Materials Physics and Mechanics 9 (2010) 175-184 Received: March 17, 2010

DISTRIBUTED COMPUTING IN THE NATIONAL-WIDE LUNG SCREENING AND DIAGNOSIS SYSTEM: FIRST STEPS

Eduard V. Snezhko[1],* Vassili A. Kovalev[1], Aliaksandr Prus[1], Alexander Dmitruk[1], S. Kharuzhyk[2]

[1] United Institute of Informatics Problems of National Academy of Sciences of Belarus

Surganova 6, 220012, Minsk, Belarus

[2] N.N. Alexandrov National Cancer Center of Belarus, Lesnoy, 223040, Minsk District, Belarus

* E-mail: snezhko@newman.bas-net.by

Abstract. The paper is devoted to consideration of methodological, large-scale data management, medical imaging and software developing issues concerned with the distributed computing in a national-wide telemedicine system. The system substantially exploits the image data provided with the complementary pulmonary X-ray and chest computed tomography (CT) image modalities; its aim is to provide the computerized support of lung disease diagnosis. The computational environment is heterogeneous and is based on PCs and dedicated servers located in medical institutions, as well as on the supercomputer installed in the National Supercomputer Center of Belarus. For utilizing the grid infrastructure, the distributed computing architecture employs the Unicore as a middleware, as well as the Message Passing Interface. At present the project is on its early stage. However, the preliminary research and experimentation performed in framework of the project already involves X-ray lung image data of more than 100 thousands of patients and about a hundred of 3D CT scans of the lungs accumulated in the database.

1. Introduction

For several years the nation-wide program of compulsory screening of adult population as well as diagnosis and treatment of pulmonary diseases is under implementation in Belarus. The program is based on a telemedicine system which involves such image data as chest X-ray scans acquired with the help of domestic fully-digital Pulmoscan-760 scanners and 3D tomograms obtained by recent multi-spiral computed tomography (CT) scanners (Volume Zoom Siemens, Light Speed by General Electrics etc.). These pieces of imaging hardware were installed in a number of general public clinics located country-wide, in tubercular prophylactic centers playing the role of regional diagnostic centers as well as in cancer dispensaries mostly dealing with CT tomography. The supporting computer facilities include networked PCs, dedicated servers and supercomputers of the SKIF family (a joint Belorussian-Russian venture) installed in the National Supercomputer Center of Belarus. The principal structure of the underlying computer network is sketched in Fig. 1.

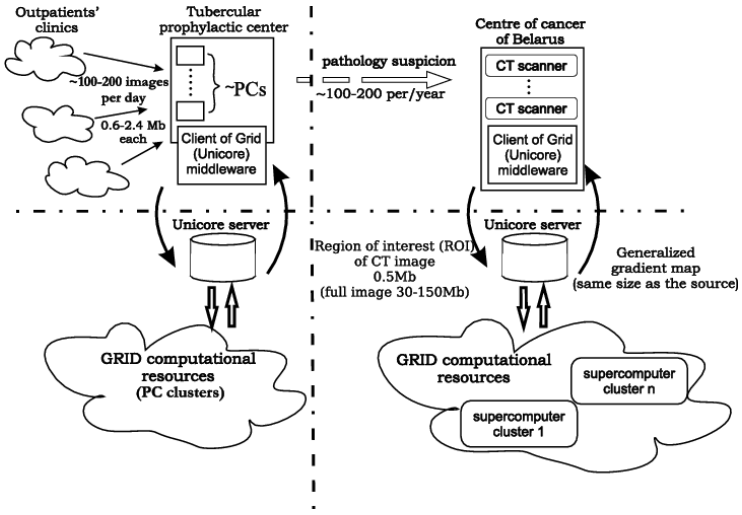

Fig. 1. The principal structure of the network.

The major application tasks being recently under development aim for computerized support of the diagnosis and treatment procedures concerned with the pulmonary diseases. They include:

1. X-ray lung image segmentation, computerized analysis and provisional diagnosis;
2. detecting the borders and highlighting the invisible internal structure of the malignant tumors on CT lung images;
3. content-based pulmonary X-ray retrieval for browsing very large image archives of a national scale and supporting the 'similar-case-in-past' diagnosis paradigm.

In this paper we are focusing on methodological, data management, medical imaging, gridification as well as cluster and grid integration issues of distributed computing in the above national-wide telemedicine system. Four national teams (specialists in computer science and medical radiology) make the contribution to the system development, each responsible for the individual part of the system (outlined by dashed lines on Fig. 1) and their integration.

2. Provisional pulmonary diagnosis based on the computerized x-ray image analysis

The analysis of X-ray images for the preliminary diagnosis is the first real-life application of the global screening of the population. Normally, about 150-200 images per day are transferred from local clinics to the single regional tubercular prophylactic center for a primary examination. For instance, at present in Minsk there are two such centers covering altogether of 24 clinics. They are interconnected to the data transfer and diagnosis network. All patients are subdivided into two groups of the healthy and pathology suspicious subjects based on the analysis results obtained at this stage.

The tool for preliminary computerized diagnosis implements two subtasks: the lung isolation and segmentation on X-ray images and computerized analysis of the segmented image regions. At the current stage of the project we implemented the algorithm for solving the segmentation task, which is able to operate with all types of images from X-ray scanners installed in the clinics of Belarus (image size vary from 0.6Mb to 2.4Mb). In general, the existing expert-based solutions of this diagnosis problem are subjective by nature and the problem of fully-automatic segmentation is not completely solved yet. The lungs segmentation experiments have been performed on a database containing 142000 of chest X-ray images and have demonstrated that in 96.31% of cases the segmentation quality was practically acceptable (see Fig. 2 for examples of segmentation performed). In about 4200 cases (2.96% of the output results) have been qualified by an automatic exit-check procedure as suspicious and were automatically directed to an additional interactive examination. The

rest 0.73% of the results were obviously wrong. Therefore we may conclude that the accuracy of the segmentation method developed varies in the range of 92-96% depending on the input image quality. Given this conclusion. we inclined to utilize it as a part of the diagnosis expert system.

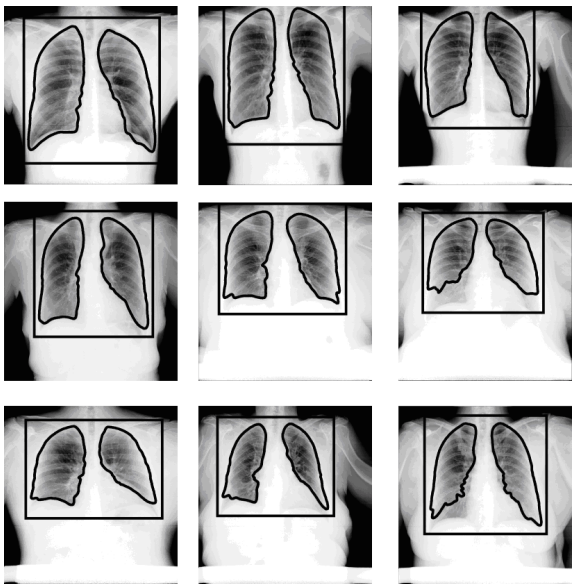

Fig. 2. Results of segmentation of lungs.

The idea of the method is to construct the outline of the lungs drawing a bunch of rays from within the lungs and estimating the length of each ray. The boundary of the lungs is a combination of different tissues and anatomical structures, which have different characteristics and X-ray absorbtion. Therefore the estimation of rays lengths requires an adaptive approach. The algorithm implemented is based on estimation of local intensity changes and it is proved to be robust for a wide variety of input images. An additional testing demonstrated that the computational costs required for the segmentation algorithm are moderate. Our experiments showed that the procedure of image scaling, histogram equalization and segmentation of the single image on the 2GHz CPU takes about 10 seconds whereas the scaling procedure occupies about 70% of time. The time expenses are close to the period of a single image transfer to/from the grid server plus the period of the task queuing if one would use the grid infrastructure for the same problem. Hence, moderate computational costs and with relatively high data transfer requirements led to the conclusion that employing grid is inexpedient for the task of the lungs segmentation.

The other aim of the project being implemented is the computer-aided automation of the most complicated manual methods of detecting pathologies such as various types of tuberculosis, pneumonia, sarcoidosis at the early stages of lung cancer. The pathologies mentioned above are well visualized on the high resolution X-ray images. However, their detection requires constant eyestrain of an expert in pulmonology. Other pathologies are not of such high interest because of their occasional occurrence and simplicity of detection. The use of grid infrastructure in this case is also not so well grounded since the data transfer and task queuing expenses per image could be higher that the period of its processing. Perhaps the optimal solution for this task could be its parallel execution using MPI implementation installed on PCs allocated within tubercular prophylactic center as it is depicted in Fig. 1. However, this restrains the use of sophisticated methods by the following reason: the expert in pulmonology analyzes a single image in several seconds; therefore developing a computerized support system that takes much more time to proceed is rather unreasonable.

Thus the segmentation and image analysis can be used for consultative purposes as a second opinion only, while a deeper analysis by means of the sophisticated classification methods and grid utilization should be done in case of a patient appointed to pass the further computed tomography examination.

3. Therapy planning based on computationally extensive non-contrasted ct image analysis

After screening in the case of pathology suspicion in lungs, a patient is referred to the CT examination. About a half of patients with centrally located lung cancer have a collapse of lung or its portion named atelectasis. Usually it is located at the outer part of lungs, i.e., closer to ribs while cancer tumors are located closer to the mediastinum.

Computed tomography (CT) is the primary modality for imaging the lung cancer patients. The introduction of a contrast agent, while CT scanning, increases the radiation stress on the patients. However, differentiating lung atelectasis and malignant tumours based on non-contrasted tomograms is hardly possible due to their very similar visual appearance: mean intensity of atelectasis area is 36.9±6.5 Hounsfield units (HU) while malignant tumors – 37.0±7.7 HU (Fig. 3). The reason may be that the human eye cannot detect the difference between image regions with equal mean value and standard deviation [1, 2], even if higher order statistics of these regions (e.g., asymmetry, kurtosis) may differ much. Yet accurate tumour segmentation is strongly necessary by the following two reasons. First, the correct tumour localization, segmentation, and precise measurement of tumour diameter play a crucial role in the therapy planning and choosing suitable surgery technique. Second, if the radiation therapy is prescribed, an exact tumour border is required for precise targeting and accurate delivery of the ionizing radiation exactly to the tumour but not to the surrounding tissue [3].

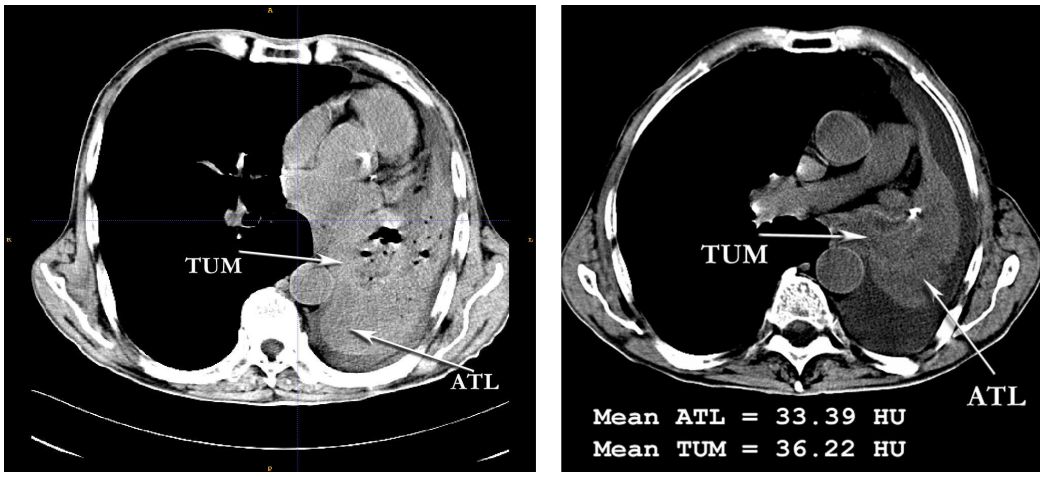

Fig. 3. Typical examples of lung CT slices with cancer tumour (TUM) and the atelectasis (ATL).

As it was mentioned above, the human eye is able to detect the difference in textures that differ only in the first- and second-order statistics, e.g. the mean value and the standard deviation. However, exactly the difference in higher order statistics between tumour and atelectasis regions on native CT lung scans was found to be noticeable using advanced the classification techniques [4].

Recently the generalized gradient method [5] was proposed to detect the borders on such images. The method consists in calculation of so-called ‘generalized’ gradient map for the original images. It performs the pass through the image with sliding window divided into two halves and calculates some statistical characteristics for each half. The difference in values

obtained for each pair of the window halves is then considered as a gradient magnitude in the window axis direction. Two kinds of the method were examined. The first one is based on the voxel sample clustering and interpreting the resultant classification accuracy as the voxel sample difference (so-called 'classification gradient') whereas in the second case the voxel sample difference is measured by the statistical significance obtained with the help of Student's t-test.

The generalized gradient method performs well on synthetic images [6] and on native CT images of lungs used for its testing [7]. Examples of resultant maps are presented in Figs. 4 and 5. Generalized gradient maps can be interpreted as follows: the darker regions on maps correspond to a more homogeneous region on the native CT images comparing to the brighter ones. In other words, the brighter voxels of gradient map are supposed to belong to the border between the relatively homogenous regions. This effect can be seen in right columns of figures in form of borders seen between regions marked by ATL (atelectasis) and TUM (tumour) on corresponding original images.

We used tomograms with the following characteristics: the voxel size was equal to 0.68 mm in the axial image plane with the slice thickness equal to the inter-slice distance of 7 mm; in-slice images resolution was 512×512. This means that their raw size is minimum 30-40 Mb and should be more if inter-slice distance is lesser than 7 mm. A visual examination of all resultant gradient maps performed by a highly experienced radiologist confirmed that the gradient maps generated definitely provide visual information useful for localization of tumour border and may be used as an additional source in the diagnosis and treatment process.

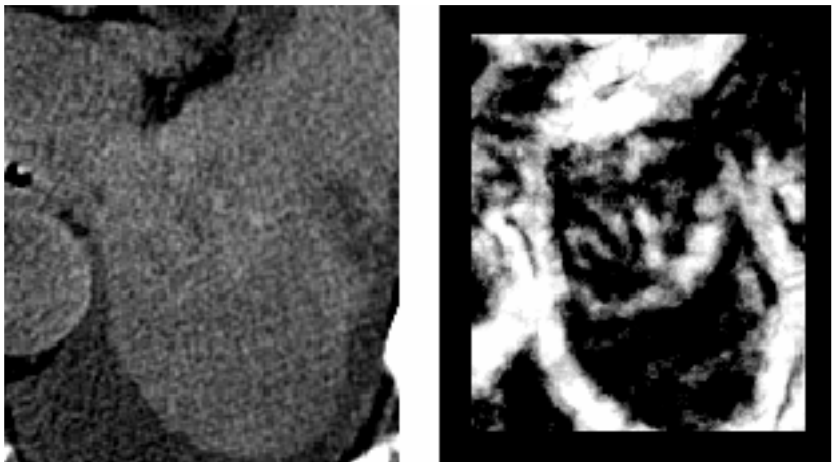

Fig. 4. Generalized gradient maps (right) of the test image (left) calculated using the triptych sliding window; R=4, d=2, n=3 and SVM classifier.

The process of calculation of generalized gradient maps is very time consuming. We have examined two different methods of generalized gradient maps calculation with different parameters. As a result, it was determined that the time expenses needed to process one chest CT image of common size are of the order of hundreds of hours on a modern 2.2GHz CPU. This is unacceptable for the use in real outpatients' settings. Yet the time needed to transfer the image of a patient being examined to the distributed computational environment is low. Moreover, this task can be naturally parallelized by data, i.e. by means of processing small data blocks independently on many computational nodes. Hence, the use of grid environment in this case is highly desirable and of natural necessity. The process of generalized gradient map generation could take several minutes depending on the computational resources available at that moment. For example, to process the image ROI of size 160 x 160 x 10 voxels on 30 computational nodes takes about 4.5 minutes using the fastest set of parameters

of the generalized gradient method, which makes it possible to provide the sophisticated diagnosis assistance on-line during the common procedure of patient examination.

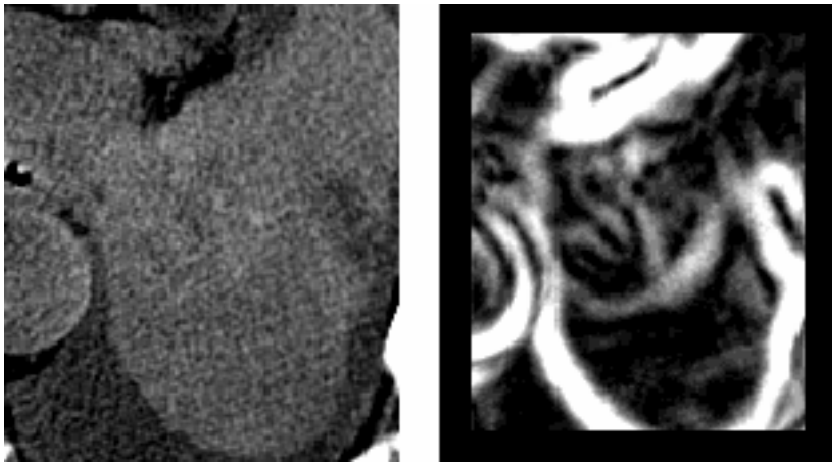

Fig. 5. Generalized gradient map (right) of the test image (left) calculated using the triptych sliding window; R=4, d=2, n=3 and the output t-test significance as a gradient measure.

4. Content-based pulmonary x-ray image retrieval

Content-based image retrieval (CBIR) is the process of retrieving images directly by visual image characteristics. In comparison to text-based image retrieval which uses textual language to describe image content, it is free from significant limitations, especially in medical domain, since image data cannot be fully described texturally. Traditionally, CBIR uses visual characteristics, such as color, texture or shape to represent image content and to retrieve images from databases which are visually similar to a query image. One of the CBIR challenges is to bridge the semantic gap between low-level features extracted automatically from images and the high-level human interpretation. Therefore CBIR for medical databases has the potential to assist clinical decision-making, research, training and easy-to-use retrieval of medical images [8]. But any image retrieval from databases of distributed organization can become a bottleneck for computer-aided diagnosis system performance. Large number of image data sets located in different institutions and clinics still present challenges for data movement. Hence, the development of efficient and robust image descriptors is the crucial point for system performance. The exchange and queries of descriptors instead of the whole images between distributed computers allow dramatically decrease the network loads. In this case only images retrieved by their descriptors are transferred to a query node.

At present the diagnosis expert support system for pulmonary deceases is under development in Belarus. One of the tasks of this system is the retrieval of visually similar images, since the comparison of some actual cases with similar examples in the past fixed in electronic archives can be a very useful diagnosis tool [9]. At the first stage of the development we decided to evaluate the quality of X-ray image retrieval by patient's age and gender based on the image descriptors proposed in this paper. This non-trivial task for evaluation was chosen on purpose since even for highly qualified physicians it is difficult to assess the patient's age and gender while observing pulmonary X-ray images.

Thus, for the experiments we used a total of 6594 images sub-sampled from image databases obtained in framework of mandatory national X-ray screening program currently running in Belarus. Of these, 3 groups of images with different patient age (22-23, 42-43, 62-63 years old respectively) were selected, equally men and women. The image size was 560×576 pixels, 16 bits of intensity resolution. Further details on the test image databases are given in Table 1.

Table 1. Image databases used in the prototype tool.

X-Ray database	
Class (age group)	**Quantity**
22-23	2198
42-43	2198
62-63	2198
Total	**6594**

One of the common features implemented in CBIR systems is the search for similar images. To accomplish this, the content should first be described in an efficient way, e.g. the so-called feature extraction should be performed. Usually the visual content of images in the database is described by feature vectors, possibly of different sort. Since original images are gray level, our tool analyzes not color but texture and its statistical features. In particular, extended co-occurrence matrices [10] for image description were used: triplets unlike pairs of pixels in classical Haralick matrices [11]. The co-occurrence matrix contains relative frequencies $P(i,j,k,d_1,d_2,d_3)$ with which pixels triplets of gray levels i, j and k respectively occur in the image at distances d_1,d_2,d_3. Every descriptor can be represented as a triangle with gray levels in its corners. Descriptors cover the entire image and represent statistical distribution of spatial relationships of gray level intensities. Consideration of the additional third pixel increases feature space and thereby enhances the sensitivity to medical texture.

In our experiments we used triangle-shaped descriptors with side sizes 1, 3 and 5 pixels. Increasing descriptor's size would lead to losing correlation between pixels and thereby local texture features. Extended co-occurrence matrices were calculated with 16 bins of gray levels. Invariance for both rotation and reflection is achieved by matrix reduction to a triangular form [12]. In other words, sequence order of intensities at sides of triangles becomes unimportant. To decrease the number of elements in matrix and simplify data, so-called feature vector was extracted from the co-occurrence matrix. Low sensitivity to image size was achieved after feature vector normalization by sum of the elements.

Image retrieval using the ‘query by example’ paradigm was performed by means of calculating the L1 distance (measure of similarity) between image feature vectors. Each of 6594 images was submitted as a query and the resultant top 30 most similar images were considered for calculating the final accuracy of image class retrieval. The results of retrieval are shown in Fig. 6. The statistical analysis of all 6594 queries showed that while searching by age the mean accuracy for 22-23 years old group was 60%, 42-43 – 40%, 62-63 – 50% and practically remained constant for the rest of retrieval results. When searching by gender, the mean accuracy approximately linearly decreases from 84% at the first result to 78% at the 30th retrieval result for male patients. Similar behavior of retrieval accuracy can be observed for female patients too.

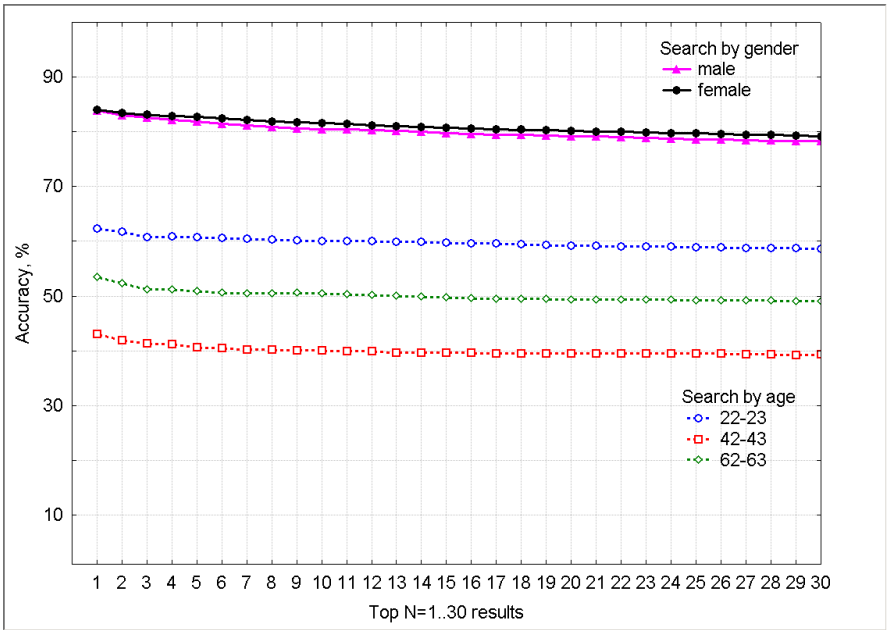

Fig. 6. Retrieval results for X-ray database.

One of the main goals of this technology is to replace original medical images by their feature vectors and small preview images (thumbnails) at the client side or at the regional server side. Considerable simplification of computational complexity and a dramatic reduction of image data needed to be transferred allow radiologists to search by image content locally, on an ordinary personal computer. Feature vector calculation performed off-line (e.g., overnight) is also not very complex: for 2084×2560 image on 2GHz Intel processor it takes about 12 seconds with the binning of gray levels lasting for 7 seconds. Therefore for this particular task, we see the applicability of grid technology in distributed data storage and management with real transferring only few resultant (retrieved) images out of hundreds of thousands.

5. Conclusion

As it was mentioned above, gridification of X-ray diagnosis and retrieval subtasks of the computerized support of the national-wide lung diagnostic system is not absolutely necessary and can be implemented based on MPI capabilities. An opposite situation is typical for large CT image analysis where the computational time may be extremely high while data transfer costs are relatively low. The team responsible for the development of sophisticated 3D image analysis methods has implemented the interconnection of the software package for the analysis of CT images with grid environment. The 3D analysis package called Voligator has been already under development for a few years in framework of the national program ‘Infotech’ and the Belarussian-Russian program ‘SKIF-Grid’. Among other, it is also based on the Insight Toolkit and Visualization Toolkit, two open-source libraries widely used for medical image processing. We also employed the Unicore middleware [13] for interconnection of Voligator with grid computing resources. Currently the Unicore server is installed in the United Institute of Informatics Problems (UIIP) of the National Academy of Sciences of Belarus and is able to grant computational memory and storage resources of the SKIF K1000-M supercomputer cluster located at the National Supercomputer Centre of Belarus in the same institution.

As soon as the user (an expert-radiologist) defines the region of interest (ROI) in the CT image, the software Voligator invokes the command line client of Unicore and transfers the ROI as well as type of analysis required for processing in grid environment. At present

configuration files for Unicore command line client should be already customized on a client machine. But it is possible and will be likely implemented within the client part of the 3D image analysis software the possibility to customize the configuration files of Unicore client concerned with the required computational resources depending on the type of analysis and the image size. Resulting generalized gradient map calculated on grid computational nodes is finally returned to the client 3D image analysis software (Fig. 1) and is superimposed on the source image as a transparent overlay shown in Fig. 7.

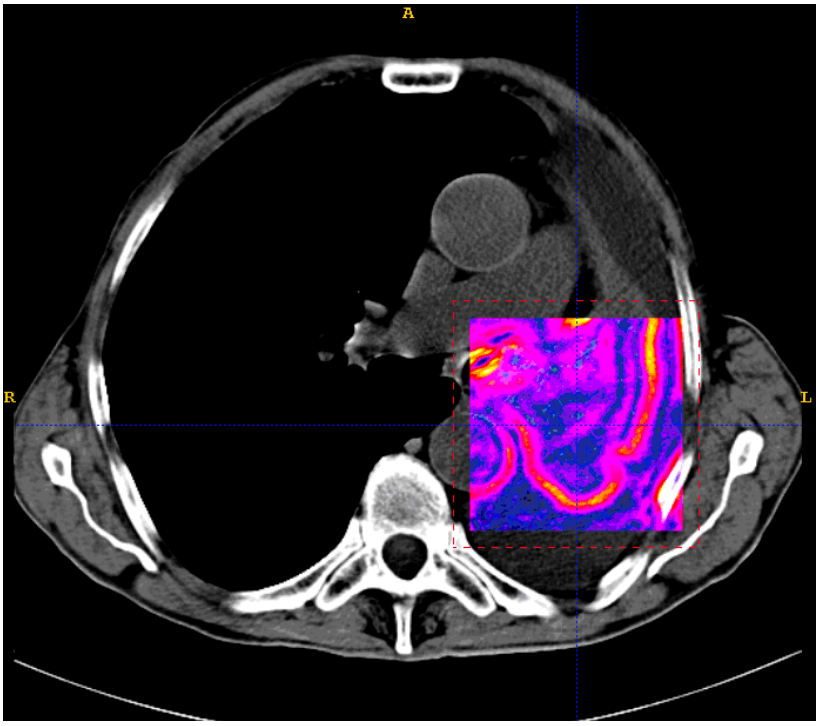

Fig. 7. Generalized gradient map visualized over the CT scan.

Acknowledgements

This work was supported by the project BalticGrid II and partly by the International Joint Project grant 2006/R1 from British Royal Society, ISTC grant B-1489 and the project No. 4G/07-225 of the 'SKIF-Grid' program.

References

[1] B. Julesz // *Nature* **290** (1981) 91.
[2] M. Petrou, V. Kovalev and J.R. Reichenbach // *IEEE Transactions on Image Processing* **15** (2006) 3020.
[3] P. Bowden, R. Fisher, M. Macmanus, A. Wirth, G. Duchesne, M. Millward, A. Mckenzie, J. Andrews and D. Ball // *International Journal of Radiation Oncology, Biology, Physics* **53** (2002) 566.
[4] V. Kovalev, M. Petrou and S. Khoruzhik // *Proc. Medical Image Understanding and Analysis 2006*, (Aberystwyth, United Kingdom, 2007) pp. 21-25.
[5] M. Petrou and V. Kovalev // *International Journal of Scientific Research* **16** (2006) 119.
[6] V.A. Kovalev and M. Petrou // *Proceedings of the 18th International Conference on Pattern Recognition*, (Washington, DC, USA, 2006) pp. 830-833.
[7] V. Kovalev and E. Snezhko // *Proceedings of the 2nd International Conference on Advanced Information and Telemedicine Technologies for Health*, (Minsk, Belarus, 2008) pp. 64-69.
[8] Z. Xue, L.R. Long, S. Antani, J. Jeronimo and G.R. Thoma // *Proceedings of SPIE* Vol.

6919 *Medical Imaging 2008: PACS and Imaging Informatics*, edited by Katherine P. Andriole and Khan M. Siddiqui, (2008) pp. 691907-691907-9.
[9] H. Abe, H. MacMahon, R. Engelmann, Q. Li, J. Shiraishi, S. Katsuragawa, M. Aoyama, T. Ishida, K. Ashizawa, C.E. Metz and K. Doi // *RadioGraphics* **23** (2003) 255.
[10] V. Kovalev and M. Petrou // *Graph. Models Image Process*. **58** (1996) 187.
[11] R.M. Haralick, K. Shanmugam and I. Dinstein // *IEEE Transactions on Systems, Man and Cybernetics*, **3**(6), 610-621 (1973).
[12] V. Kovalev, F. Kruggel, H.J. Gertz, and D. von Cramon // *IEEE Transactions on Medical Imaging,* **20**(5), 424-433 (2001).
[13] A. Streit, D. Erwin, T. Lippert, D. Mallmann, R. Menday, M. Rambadt, M. Riedel, M. Romberg, B. Schuller and P. Wieder // *Grid Computing: The New Frontiers of High Performance Processing, Advances in Parallel Computing,* edited by L. Grandinetti, **14**, 357-376 (2005).