Semi-Automated Quantitative Validation Tool for Medical Image Processing Algorithm Development

Viktor Jonas, Miklos Kozlovszky, Bela Molnar

Semi-Automated Quantitative Validation Tool for Medical Image Processing Algorithm Development

Viktor Zoltan Jonas[1], Miklos Kozlovszky[2,3], Bela Molnar[4]

[1]Doctoral School of Applied Informatics, Óbuda University, Budapest, Hungary
[2]Biotech Knowledge Center, Óbuda University, Budapest, Hungary
[3]MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Victor H. str. 18-22., H-1518 Hungary
[4] Second Department of Internal Medicine, Semmelweis University, Budapest, Hungary

{ viktor.jonas@3dhistech.com,
kozlovszky.miklos@nik.uni-obuda.hu,
bela.molnar@3dhistech.com }

Abstract. Cancer research and diagnostics is an important frontier to apply the power of computers. Researchers use image processing techniques for a few years now, but diagnostics only start to explore its possibilities. Pathologists specialized in this area usually diagnose by visual inspection, typically through a microscope, or more recently on a computer screen. They examine at tissue specimen or a sample consisting of a population cells extracted from it. The latter area is the area of cytometry that researchers started to support by creating image processing algorithms. The validation of an image processing approach like that is an expensive task both financially and time-wise. This paper aims to show a semi-automatized method to simplify this task, by reducing the amount of human interaction necessary.

Keywords: validation tool, automated validation, medical image processing

1 Introduction

The project [1] we are currently participating in aims to reproduce a medical research and diagnostic method called ploidy analysis (PA) through image processing means. PA is a method to measure DNA content in cell nuclei as a basic cancer marker, and is considered as a segment of pathology, more closely image cytometry (ICM). In diagnostics PA is usually done by a machine called Flow Cytometer (and the family of assessments related to it flow cytometry (FCM)), an appliance that operates with a light beam directed at a transparent capillary, where objects are traveling in sheath fluid, facilitating laminar flow. The objects are measured through their optical properties like light scattering, but the result is more one-dimensional measurement functions of time. Image cytometry takes a different approach. Digital pathology is in the process of introduction into medical diagnostics. This new approach is based on

the idea of taking traditional glass slide specimen to the computer screen through digitalization. This enables experts to use the monitor to evaluate the samples, and also use software tools to achieve the task. Expectations are that this approach enhances objectivity reproducibility and traceability of forming diagnoses. Image cytometry is the sub-field where the software tools are used to process the digitized sample, enabling its users to analyze vastly more objects than the traditional visual-manual method using microscopes for inspection and clickers for counting. The project we are working on aims to reproduce a flow cytometry analysis by image processing means [2][3]. Is the ICM approach viable as a diagnostic approach? This is a twofold question in itself: is it possible, and is it capable of sufficient throughput (equivalent of FCM)? The hypothesis of course is that it is possible, but measurements have to be taken to confirm.

A crucial part in applying (image processing) algorithms in forming medical diagnoses, is validation. There are two possible routes: validation by comparison to a validated approach (in this case FCM result), or the method used for validating FCM in the first place: clinical a study. The clinical study approach was chosen to eliminate the inherent error accumulation, and the simpler procedure regarding laboratory access and overall expenses in human work. This approach is semi-automatic in the sense of needing manual input for the quantitative validation in the form of the planar coordinates of the reference objects. The novelty in our approach is to decrease the human interaction as much as possible. To achieve this validation process was separated into two steps, only the first needing direct expert interaction: marking the objects to detect on the sample. Something similar was done traditionally: using a clicker counter, while examining the sample with a microscope. This we named the quantitative part. It is possible to add a qualitative part that relies on the result of a quantitatively sound detection as reference, for measuring object detection quality.

This paper aims to present an algorithmic tool for first approach validation of the image cytometry algorithm described, and possibly other image segmentation projects. The tool is constructed to be also useful in image processing algorithm development, by enabling continuous comparison to results of pervious variants.

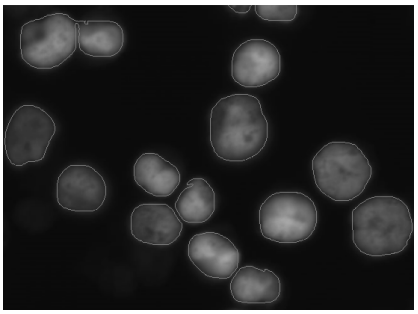

Fig. 1. Sample area with segmentation mask. On the top-left a merged pair of nuclei is visible. On the bottom left area a few (very) low-intensity objects are located.

2 Benefits from Cloud-based Engineering Systems

The image processing of medical images regarding small objects usually entail tiled image processing. This, in itself is a problem that scales well in parallel processing. Comparing or validating the results of an above mentioned algorithm can also be organized for highly parallel processing. More importantly considering the amount of samples to analyze in a pathology lab (that are usually highly centralized facilities), that should be done more quickly, than manual/visual inspection of the sample, to have relevance in diagnostics. A glass slide has the useful area of roughly 15x25mm. The optics in the hardware our project used for sample digitization enables ~0.2µm/pixel resolution. This means a bit more than 9GB of image data for each image channel. (Naturally for storage it can be compressed, but processing usually uses the uncompressed, full magnification data.) Though our project uses single channel images, but a medium sized diagnostic lab works with a few hundreds of these samples a day [4].

Recent papers discuss the role of cloud in medical image processing like [5]. Considering the amount of data accumulated in digitized glass slides the storage advantages of clouds seems also an option to explore. Creating the required storage capacity, the security issues and the problem of maintaining these may be easily solved through cloud. Taking a step back another dimension of clouds can also be valuable for this area and that is the consultation over medical cases, the possibility to easily share and inspect samples, or even cases (sets of samples corresponding to a patient) are already feasible, but with pre-cloud techniques his can be solved only in a less than ideal manner.

3 Related Literature

There is extensive literature on the evaluation of image segmentation quality and validation. We started with an earlier work in mind; that work was a case study of a clinical validation [6]. The categorizing approach of [7] was of great help to widen our field of view in this area. And other works like [8], which used the same supervised approach as we planned, were of great help to select the depth and scope of this paper.

4 Discussion

Image processing projects usually encounter at least two types of demand considering validation/comparison. One is the actual validation of the segmentation algorithm or the software solution, where segmentation result usually has to be compared to human "segmentation" results (as in a clinical study – supervised approach). The other case is during the development of the algorithm that compares the algorithm to its previous version or to a completely different approach to confirm or measure the change or difference, to be able to rank them, and improve the best one further. Accordingly the

proposed software application is constructed to compare two measurements. This enables us to consider it as an ordering relation of some sort between image processing algorithms and also as a validation tool, when comparing human validation as reference input. The analysis tries to find corresponding object pairs (or n-tuples) from both measurements (reference and the currently tested), compare them, and register their relation using result/error classes.

4.1 Data

A measurement to compare can be considered an array of records containing all measured morphometrical and colorimetrical properties of each detected object (cell nucleus in the case of the actual study). This paper focuses on reproducing the process of a manual validation project, only the morphology and location information is analyzed, though a simple check is conducted on other properties of the measured objects for enabling automatic discovery of unanticipated changes. Samples of human blood (lymphocyte nuclei) were used in this case. A 1 mm^2 area was analyzed on each of the 17 digital slides. The images processed were 0.1625µm/per pixel resolution compressed (jpeg 80). This means roughly 6150*6150 pixels on each sample.

4.2 Analysis

The simplest case is when object shapes (as polygons) match, and also all their measured (non-geometrical) parameters match. (For optimization purposes all shape comparison is preceded by the comparison of their bounding boxes, to ensure quicker analysis.) This group is labelled *Perfect match*. If the criterion for parameters fails the cluster is labelled *Parameter mismatch*. These two groups are most significant in synthetic tests, where objects with known geometry are segmented and analyzed. It can also be helpful during algorithm development to detect unwanted changes in segmentation or parameter measurement, as a step of automated testing protocol. This is why these two cases are handled separately, if at the end of this step all objects are accounted for, the analysis is over, if not, the remaining are passed to the next stage of the test.

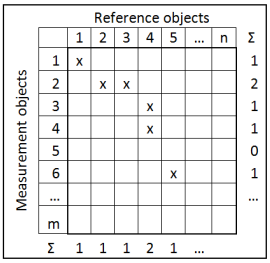

Fig. 2. Illustration of the relation matrix. Rows represent the items of the gold standard, columns contain the elements of the compared measurement. X-es designate objects where overlap is possible.

This next stage leverages the idea of neighborhood matrices and object overlap. Both measurements are assigned to a dimension of the matrix, and relations are recorded to the cell addressed by the two interacting detected objects. If their bounding boxes intersect, and also the actual shapes intersect the interaction is marked in the corresponding matrix cell as a binary flag (as visible in Fig. 2). Shape overlap is detected by simply rendering the two objects to the same image additionally, and counting overlapping pixels. Intersecting ratio is defined as the intersection area over the area of the object with greater area. When this matrix is populated a simple analysis is conducted. For simplicity name the horizontal dimension reference and the vertical the measurement. Also generate column and row sums for the matrix.

1:0 and 0:1 Object Relation

These cases can be found by collecting entries, where either the column sum or the row sum equals to zero. When there is a reference object that has no corresponding measurement object it is marked as *false negative*. The complementary case is labelled *false positive* (measurement for no reference). These objects are registered, and removed from the matrix.

1:1 Object Relation

Where the column and the row sum is also equals to one means that to the reference object only one measurement object is assigned, and also to that measurement object corresponds only (this) one reference object. These objects are further grouped into two classes. Where the centroid of the objects is on the same location the object pair is labelled as *match*, where the centroid locations differ, the label assigned is *shifted*. (In our case instead of centroid simply the center of the bounding box was used.) A shifted (or not perfect) match can be observed on Fig. 3.

Fig. 3. Reference marker and detected object overlap test. Manually placed "reference" position marker on the left, segmented object mask on the right.

This classification may later be used for a few purposes:

When a measurement is run only on a smaller sub-area of the reference area, there is difference in coordinate systems of the two samples, but otherwise comparing two identical measurements. Similarly if the tissue sample is digitized twice, similar difference may be observed. A new digitization process may assign a new coordinate system for the digital sample, or the tilling can change the coordinate system locally, because of mechanical or parametric differences of the two attempts.

Naturally these analyzed pairs of objects are also removed from the matrix.

1:n and n:1 Object Relation

In the remaining set of objects where the column sum equals one (and the row sum is greater than one, otherwise it would have been processed earlier) is the class of *Merged*, meaning more reference objects to a single measurement object. Conversely where only row sum equals one is labelled *Split*, assigned to the case where to one reference there are multiple measurement objects.

Fig. 4. A segmentation object containing two markers. Image processing "merged" the two nuclei into one segmented object.

All remaining cases are labelled as *Residuum*. (n:m relations are not desirable in this setting, found no purpose in further analyzing them.)

For the validation of the above described method the results of the previously published manual assays were used. To be able to compare these results the result classes of the two methods have to be arbitrated. The resulting classification results are visible in **Table 1**.

Table 1. Results of the manual and the automatic assay. Reference column designates the count of nuclei the expert marked. FP (false positive), FN (false negative), Match and Other columns contain the count of objects in the named cluster.

	Human rating				Algorithmic rating			
Reference	**FP**	**FN**	**Match**	**Other**	**FP**	**FN**	**Match**	**Other**
1508	34	253	1223	24	30	264	1213	17
1937	41	450	1411	66	43	453	1401	46
1301	68	243	977	71	67	238	1001	26
1766	38	348	1381	32	47	464	912	97
1674	48	347	1257	61	44	345	1238	48
2040	19	434	1535	59	20	437	1514	47
977	51	115	850	12	51	123	829	19
1586	29	305	1242	32	27	312	1235	22
1259	32	170	1075	8	31	180	1065	8
2175	28	479	1606	81	24	482	1584	55
1677	20	383	1245	42	19	382	1224	36
1524	23	423	983	111	22	425	973	64
2175	40	551	1559	53	38	548	1535	47
2110	16	528	1512	64	10	532	1503	40
1776	32	413	1275	80	22	412	1246	64
1957	11	547	1352	47	170	361	1016	11
831	61	107	706	18	61	110	703	10

Definition of the Evaluation Classes

Match: reference marker corresponds with exactly one measurement. In the proposed method the "Match" and "Shifted" classes, and also the half of the count of the "Split" class[1]

False positive: there is a measurement, but no corresponding reference object. Same in the proposed method and the other half of the "Split" class[1]

False negative: to a reference there is no measurement object. Same in the proposed method

Other: all other cases, usually where one measurement corresponds to more reference objects. Merged, Residuum classes

The result classes of the assays can be considered categorical variables, the manual results as expected and the algorithmic results as measured values. The Chi-square goodness of fit test is used to confirm that the proportions the algorithm produced do not differ significantly from the ones the manual assay states.

The test was conducted using the significance level of 95% (p = 0.05). The column **h** in Table 2 contains whether the above null hypothesis stands, column $\mathbf{p_t}$ contains the p-value needed to reject the null hypothesis at the actual significance level.

Table 2. The result of the Chi-square goodness of fit test of concordance between the human and the algorithmic rating. Sample ID designates the glass slide containing the cell nuclei; the other columns contain the results of the statistical test.

sample ID	χ^2	p_t	h	sample ID	χ^2	p_t	h
1M01	1.7038	0.7900	1	1M13	5.1632	0.2710	1
1M02	3.5843	0.4652	1	1M14	0.5417	0.9693	1
1M03	21.0897	3.0e-04	0	1M15	12.1855	0.0160	0
1M04	731.7800	0	0	1M16	0.4652	0.9768	1
1M05	1.6605	0.7979	1	1M17	6.8198	0.1457	1
1M06	1.4359	0.8379	1	1M18	3.5227	0.4744	1
1M10	2.1032	0.7168	1	1M19	227.0211	0	0
1M11	1.9999	0.7358	1	1M20	2.3150	0.6780	1
1M12	0.3481	0.9865	1				

5 Conclusion

The results of the proposed comparator algorithm mostly concur with the results of the manual assay. In case of four samples the results seem to differ significantly. Further investigation is needed to uncover the cause in those cases. This may enable faster, more objective comparison of image segmentation. Being able to compare measurements in the magnitude of a few tens of thousands of objects automatically adds the possibility for some additional testing of similar image segmentation algorithms.

A few weaknesses were discovered. Comparison processing time grows quickly with the number of measured objects. Analysis speed is inherently at least $O(n^2)$

[1] In the manual assay the rule of thumb given to the expert was to mark the best fitting segmentation result as a *Match* all others as false positives.

because of the interaction-matrix. If the processing algorithm in itself is tile processing based (as in the case of our project), comparing the result subsets tile-wise seems a viable solution. In other cases by using spatial ordering of both the reference and the measurement is possible (like the storage or indexing the measurement in a quad tree manner), thus being able to construct interaction matrices for objects that possibly overlap in a distributed (and also quicker than $O(n^2)$) manner.

Comparison of larger objects is not efficient or even not possible; rendering them on single images in memory may not be an option. Overlap calculation in their case should be implemented by an entirely different approach.

The separate step of detecting complete matches and geometric matches for testing purposes makes this twice as slow, the two "modes" should probably be used separately, mode chosen explicitly. To add a module with the possibility of supplying image masks and/or text files with a strictly set format as comparison input is also our future goal. This is necessary to be able to use the tool more generally or in other projects. Extending the comparison capabilities to the level of detail of the segmented objects' level is also a possibility.

References

1. V. Z. Jonas, M. Kozlovszky, B. Molnar; Ploidy Analysis on Digital Slides. IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI 2013), Hungary, 287--290, ISBN 978-1-4799-0194-4, DOI: 10.1109/CINTI.2013.6705207, (2013).
2. V. Z. Jonas, M. Kozlovszky, B. Molnar; Nucleus detection on propidium iodide stained digital slides. IEEE 9th International Symposium on Applied Computational Intelligence and Informatics (SACI2014), Timisoara, Romania, 139--143 DOI: 10.1109/SACI.2014.6840051 (2014).
3. V. Z. Jonas, M. Kozlovszky, B. Molnar; Separation enhanced nucleus detection on propidium iodide stained digital slides. IEEE 18th International Conference on Intelligent Engineering Systems 2014 (INES 2014), Tihany, Hungary, 157--161 DOI: 10.1109/INES.2014.6909360, (2014).
4. Nikolas Stathonikos, Mitko Veta, André Huisman, Paul J. van Diest; Going fully digital: Perspective of a Dutch academic pathology lab, Journal of Pathology Informatics, DOI: 10.4103/2153-3539.114206 (2013).
5. G. C. Kagadis, C. Kloukinas, K. Moore, J. Philbin, P. Papadimitroulas, C. Alexakos, P. G. Nagy, D. Visvikis and W. R. Hendee: Cloud computing in medical imaging, Medical Physics 40, DOI: 10.1118/1.4811272 (2013).
6. L. Krecsak, T. Micsik, G. Kiszler, T. Krenacs, D. Szabo, V. Jonas, G. Csaszar, L. Czuni, P. Gurzo, L. Ficsor and B. Molnar; Technical note on the validation of a semi-automated image analysis software application for estrogen and progesterone receptor detection in breast cancer, Diagnostic Pathology, 6:6 DOI:10.1186/1746-1596-6-6, (2011).
7. Hui Zhang, Jason E. Fritts, Sally A. Goldman; Image segmentation evaluation: A survey of unsupervised methods, Computer Vision and Image Understanding (CVIU), 110(2), 260--280, (2008).
8. Christian Ledig, Wenzhe Shi, Wenjia Bai, Daniel Rueckert; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3065—3072, (2014).