mda2006

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

Workshop on Mass-Data Analysis of Images and Signals

Petra Perner (Ed.)

Workshop Proceedings

IBaI CD-Report ISSN 1617-2671

Published July, 2006

www.data-mining-forum.de

Organized by:

ibai

Institute of Computer Vision
and Applied Computer Sciences
Dr. Petra Perner

Preface

The Industrial Conference on Data Mining ICDM-Leipzig was the sixth event in a series of annual events which started in 2000. We are pleased to note that the topic data mining with special emphasis on real world applications has been adopted from so many researchers all over the world into their research work. We received 156 papers from 19 different countries.

The main topics are data mining in medicine and marketing, web mining, mining of images and signals, theoretical aspects of data mining, and aspects of data mining that bundles a series of different data mining applications such as intrusion detection, knowledge management, manufacturing process control, time-series mining and criminal investigations.

The program committee was working hard in order to select the best papers. The acceptance rate was 30%. All these selected papers are published in this proceeding volume as long papers up to 15 pages. Besides that we installed a forum where work in progress has been presented. These papers are collected in a special poster proceeding volume and show once more the potentials and interesting developments for data mining for different applications.

Three new workshops have been established in connection with ICDM: 1. Mass Data Analysis on Images and Signals, MDA 2006, 2. Data Mining for Life Sciences, DMLS 2006, and 3. Data Mining in Marketing, DMM 2006. These workshops are developing new topics for data mining under the aspect of the special application. We are pleased to see how many interesting developments are going on under these topics.

We would like to express our appreciation to the reviewers for their precise and highly professional works. We appreciate the help and understanding of the editorial staff at Springer and in particular Alfred Hofmann, who supported the publication of these proceedings in the LNAI series.

We wish to thank all speakers, participants, and industrial exhibitors who contributed to the success of the conference.

We are looking forward to welcoming you to ICDM 2007 (www.data-mining-forum.de) and to the new work you will present there.

July 2006 Petra Perner

Citation

Petra Perner (Ed.)
Workshop on Mass-Data Analysis of Images and Signals, MDA2006, Workshop Proceedings,
IBaI CD- Report, ISSN 1617-2671, July 2006

Example Paper:

Mauro Evangelisti, Primo Coltelli, Valtere Evangelista, Paolo Gualtieri,
An automatic real-time system for the determination of translational and rotational speeds of swimming microorganisms. In Workshop Proceedings: Petra Perner (Ed.), Workshop on Mass-Data Analysis of Images and Signals, MDA2006
IBaI CD- Report, ISSN 1617-2671, July 2006 Weixing Wang, p. 1-13.

Table of contents

Object detection in watershed partitioned gray-level images

Maria Frucci and Gabriella Sanniti di Baja

Institute of Cybernetics "E.Caianiello", CNR, Pozzuoli, Italy
(m.frucci, g.sannitidibaja)@cib.na.cnr.it

Abstract. Gray-level image segmentation is the first task for any image analysis process, and is necessary to distinguish the objects of interest from the background. Segmentation is a complex task, especially when the gray-level distribution along the image is such that sets of pixels characterized by a given gray-level are interpreted by a human observer as belonging to the foreground in certain parts of the image, and to the background in other parts, depending on the local context. It very seldom happens that the background is characterized by an almost uniform gray-level. Thus, in the majority of cases, segmentation cannot be achieved by simply thresholding the image, i.e., by assigning all pixels with gray-level lower than a given threshold to the background and all remaining pixels to the foreground. One of the most often adopted segmentation techniques is based on a preliminary partition of the input gray-level image into regions, homogeneous with respect to a given property, to successively classify the obtained regions in two classes (foreground and background). In this paper, we follow this approach and present a powerful method to discriminate regions in a partition of a gray-level image obtained by using the watershed transformation. The basic idea underlying the classification is that for a wide class of gray-level images, e.g., a number of biological images, the boundary between the foreground and the background is perceived where locally maximal changes in gray-level occur through the image. Our classification procedure works well even starting from a standard watershed partition, i.e., without resorting to seed selection and region growing. However, we will also briefly discuss new criteria to be used when applying digging and flooding techniques in the framework of watershed transformation, so as to produce a less fragmented partition of the image. By using the so obtained partition of the gray-level image, the successive classification is facilitated and the quality of the obtained results is improved. Some hints regarding the use of multi-scale image representation to reduce the computational load will also be introduced.

1. Introduction

Gray-level image segmentation is a necessary step in any image analysis process to single out the subsets of the image constituting the objects of interest (foreground) and so to distinguish them from the background.

Recent surveys of different approaches to image segmentation can be found in [1,2]. Histogram thresholding (see e.g., [3]) is characterized by low computational complexity, but is suitable mainly for images where gray-level distribution is roughly articulated in two well defined peaks, separated by a not too broad and flat valley, i.e., images perceived as naturally binary such as written documents. To overcome these limits, rule-based methods combined with learning methods such as case-based reasoning have been developed [4]. Based on a rule set the histogram is properly smoothed and the right number of peaks is selected. Case-based reasoning ensures the incremental learning of the rule set with the proper parameters. Another approach is based on feature space clustering (see e.g., [5]), which is based on the assumption that each region of the image constitutes an individual cluster in the feature space. This method is easy to implement, but the selection of the proper features is critical and, analogously to histogram-based techniques, it does not take into account spatial information. Thus, this technique fails in presence of regions that a human observer assigns to either the foreground or the background depending on the local context. Region-based approaches (see e.g., [6]) require a suitable selection of seeds from which a growing process is done to group pixels in homogeneous regions. Of course, the selection of the seeds plays a key role for the quality of the obtained results and the method works well when the region homogeneity criterion can be defined in an easy manner. A related approach is based on edge detection techniques (see e.g., [7]). This approach follows the way in which a human observer perceives objects by taking into account the difference in contrast between adjacent regions. A segmentation method exploiting both the region-based approach and edge detection is based on the watershed transformation [8]. Fuzzy approaches use a membership function to represent the degree of some properties and are generally characterized by high computational cost. Neural network techniques can also be used to perform classification of regions, but the training phase is long and the results may be biased by the initialization phase.

The segmentation procedure to be adopted depends on the specific image domain. In this paper we consider the class of images where the distinction between foreground and background is based only on the analysis of gray-level information, without involving other features, such as the shape [9] expected to characterize the foreground components. In particular, we refer to images where the foreground is either consistently locally lighter (or consistently locally darker) than the background. This class includes, for example, a number of biomedical images. In the digitized version of a histological specimen, the regions of interest are characterized by a different gray-level, either because these regions actually have different intensity in the specimen, or because they are placed at a different depth in the slide and, hence, some of them result out of focus.

For the class of images considered in this paper, a segmentation method based on the use of the watershed transformation is the most suited one. Once the gray-level image has been partitioned into homogeneous regions, we classify the regions as belonging to either the foreground or the background, depending on the analysis of the locally maximal changes in gray-level between pairs of adjacent regions. Our classification procedure can be applied to the basic partition obtained by standard watershed transformation, i.e., without taking into account suitable procedures to select the significant seeds. Better results are achieved if the classification is accomplished on a more sophisticated watershed partition, e.g., the partition obtained by using the algorithm introduced in [10], which significantly reduces the excessive fragmentation of the input image into regions. An alternative way to reduce oversegmentation is to re-sort to multi-scale image representation. When a gray-level image is observed at different resolutions, only the most significant regions are perceived at all resolutions. In turn, regions with lower significance, which can be interpreted as fine details, are perceived only at sufficiently high resolution. Thus, if the seeds for watershed segmentation are detected at lower resolution, and these seeds are used to discriminate between significant and non-significant seeds in the image at full resolution, the partition is expected to consist mainly of the most significant regions.

This paper is organized as follows. In Section 2, we briefly discuss the standard watershed transformation as well as the method [10] to partition a gray-level image into a set of regions. In Section 3, we describe the procedure to classify the obtained regions in the two classes (foreground and background). In Section 4, we give some hints regarding the use of multi-scale image representation to reduce the computational load of segmentation. Finally, concluding remarks are given in Section 5.

2. Watershed partition

The 2D gray-level input image, used in this paper as running example, has been provided by courtesy of Dr. V. Guglielmotti and includes pyramidal neurons of rabbit cerebral cortex. See Fig.1, left. Gray-levels are in the range [0, 255]. In the running example, the foreground is perceived as locally darker with respect to the background. Thus, the foreground consists of the pixels having locally lower gray-level, according to the generally followed criterion for which the highest gray-level 255 corresponds to white, while the smallest possible value 0 corresponds to black.

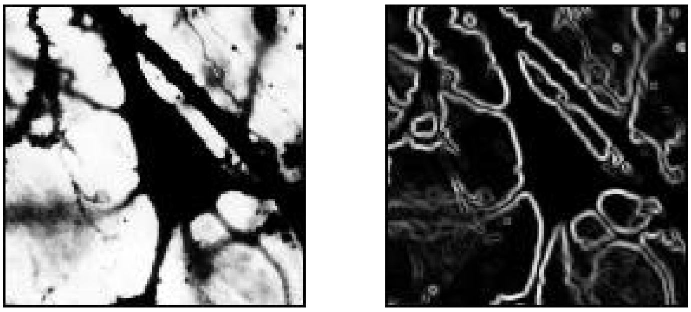


Fig. 1. The input image used as running example, left, and the relative gradient image, right.

A gray-level image can be interpreted as a 3D landscape, where for every pixel in position (x,y), its gray-level plays the role of the z-coordinate. High gray-levels are mapped into mountains of the landscape, and low gray-levels into valleys. An easy way to explain how watershed transformation produces a partition of the image is the following. Let us assume that the landscape is immersed in water, after the bottom of each valley has been pierced. As a result, the valleys are flooded. Filling of a valley begins as soon as the water level reaches the bottom of that valley. A dam is built to prevent water to spread from a catchment's basin into the neighboring ones, wherever waters from different basins are going to meet. When the whole landscape has been covered by water, the basins are interpreted as the parts into which the landscape is partitioned by means of watershed lines.

In a standard watershed transformation, the bottoms of all the valleys, i.e., the regional minima, are detected in the gradient image of the input gray-level image, see Fig. 1 right. The regional minima are used as seeds for region growing. Watershed transformation generates a partition of the (gradient) image into regions characterized by homogeneity in gray-level.

As it can be seen with reference to Fig.2 left, where the watershed lines are superimposed onto the input image, the image is fragmented in a quite large number or regions (1010 for the running example). Oversegmentation is caused by the too many detected regional minima, which are not all perceptually significant.

To reduce oversegmentation, a careful selection of the regional minima to be used for region growing is necessary. Flooding and digging techniques are generally employed to cause disappearance of those regional minima that are recognized in the gradient image as corresponding to non-significant regions. Of course, the definition of significant region is crucial to obtain a meaningful partition. In [10], a new criterion has been introduced to evaluate the significance of the regions and to merge non-significant regions only with selected adjacent regions. Merging is obtained by applying again the watershed transformation on a suitably modified gradient image, which includes a smaller number of regional minima with respect to the original landscape.

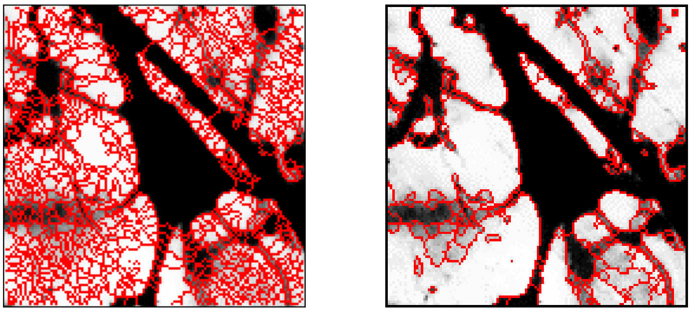


Fig. 2. Standard watershed partition, left, and watershed partition by the algorithm [10], right.

In [10], as soon as the watershed partition is available, the significance of a region X is defined by evaluating the interaction of X with every adjacent region Y. Two parameters are used to define the interaction: i) the maximal depth of X when the water reaches the *local overflow pixel*, i.e., the pixel with minimal height along the watershed line separating X from Y, and ii) the absolute value of the difference in height between the regional minima of X and Y. If non-significant regions exist in the current watershed transform, the watershed transformation needs to be applied again after the seeds corresponding to the non-significant regions have been suitably removed. Three cases are possible:

1. X is significant with respect to each adjacent region Y. Then, X is definitely meaningful and no merging is necessary.

2. X is non-significant with respect to each adjacent region Y. Then, X has to be absorbed by (some) adjacent region(s). To this aim, the regional minimum of X has

to be removed before applying again the watershed transformation. Flooding is accomplished by setting all pixels of X with gray-level lower than the lowest local overflow value *q*, to value *q*. X will result as merged to Y, when the watershed transformation is applied again.

3. X is significant in correspondence of some adjacent regions only. Then, X has to be merged with proper regions, selected among those with respect to which X is non-significant. Along the watershed line between X and any such a candidate region Y, a local overflow pixel exists, which is not necessarily the lowest local overflow pixel. Digging is performed, in this case, to open a canal connecting X with the region Y, which will absorb X when the watershed transformation is newly applied. The canal is identified as the minimal length path linking the regional minima of X and Y, and passing through the local overflow pixel common to X and Y. The gray-level of all the pixels in the path is set to the lower value between those of the regional minima of X and Y. When the watershed transformation is newly applied, the water can flow through the canal from X to Y, and the desired merging is obtained. The watershed lines of X, which were already detected as separating significant regions, are not altered.

The process is iterated until all resulting regions are significant. A remarkable reduction in the number of seeds, hence of the regions of the partition, is obtained. In turn, the computational cost is higher than that of standard watershed transformation, due both to the repeated application of the watershed transformation, and to the process aimed at computing region significance and possibly perform region merging via flooding and digging. For the running example, instead of 1010 regions characterizing the partition obtained by standard watershed transformation, only 259 regions are found. See Fig.2 right. Non-significant regions have been absorbed by adjacent significant regions. Non-significant regions have never been grouped to form a new, unexpected, significant region, or a region whose shape is altered with respect to the foreseen shape.

3. Classification of regions

The watershed transformation has partitioned the image into N regions, whose membership to either the foreground or the background has not yet been established. Since the pixels constituting a region R_i of the partition don't have all the same gray-level, we compute the average, r_i, of the gray-levels of all pixels in R_i, and use it as the representative gray-level for the whole region. Adjacent regions with the same value of r_i are interpreted as constituting a single region.

We first classify the regions whose representative gray-level is smaller (greater) than the representative gray-levels of all their adjacent regions, as belonging to the foreground (background). This initial classification is done by means of a global process, which detects, in a parallel way, all gray-level local minima and local maxima. Obviously, only the pits of the valleys and the peaks of the mountains in the landscape are classified by this process.

The still unclassified regions constitute the slopes in between peaks and pits. For these regions, our classification method is inspired by visual perception and is based on the difference in gray-level between adjacent regions. In fact, the boundary separating the foreground from the background is perceived as placed wherever strong differences in gray-level occur. Thus, for any pair of adjacent regions R_i and R_j, out of which at least one is still unclassified, we compute the difference $D_{i,j}=|r_i-r_j|$. Without losing generality, we assume that the first region, R_i, in any such a pair (R_i, R_j) is the darker one and the second region, R_j, is the lighter one, i.e., we assume $r_i<r_j$.

An iterative classification process is performed, at each iteration of which the current value $\Delta=\max\{D_{i,j}\}$, i.e., the currently maximal difference in gray-level, is used to select the pairs of regions in between which the boundary is more likely to be placed. The process is iterated until all regions are classified.

At each iteration, two cases are possible, depending on the number k of adjacent regions $R^k{}_i$ and $R^k{}_j$ with difference Δ that are found.

When k=1, we classify the darker region $R^k{}_i$ of the unique selected pair, as belonging to the foreground, and the lighter region $R^k{}_j$ as belonging to the background. Moreover, we also classify in a global way all the unclassified regions that are characterized by representative gray-level not larger than $r^k{}_i$ (and, hence, darker than $R^k{}_i$) as belonging to the foreground, and all unclassified regions with representative gray-level not smaller than $r^k{}_j$ (and, hence, lighter than $R^k{}_j$) as belonging to the background.

When k>1, classification is still done by using a global process only if the k darker regions $R^k{}_i$ have their representative gray-levels smaller than the representative gray-levels of all the lighter regions $R^k{}_j$. In other words, if the value $max_{min}= \max_k\{r^k{}_i\}$ is smaller than the value $min_{max}= \min_k\{r^k{}_j\}$, we classify all regions with representative gray-level not greater than max_{min} as belonging to the foreground, and all regions with representative gray-level not smaller than min_{max} as belonging to the background.

In turn, if at least one of the k, k>1, darker regions $R^k{}_i$ has representative gray-level not smaller than the representative gray-levels of all the lighter regions $R^k{}_j$, the same global classification would lead to conflictual assignments. For example, the region with representative gray-level min_{max} should be assigned to the background, since that region is the lighter one in the pair including it, but it should be assigned to the foreground, since it results to be darker than the region with representative gray-level max_{min}. To avoid conflicts, we classify globally only the regions with representative gray-level not larger than $min_{min}=\min_k\{r^k{}_i\}$ (not smaller than $max_{max}=\max_k\{r^k{}_j\}$) as belonging to the foreground (background). For any remaining region $R^k{}_i$ belonging to a pair of regions with difference Δ, the following local investigation is done. All ascending paths, consisting of unclassified regions with increasing representative gray-levels, are traced along the slope including $R^k{}_i$ until a classified region is met. Since along the slope, more than one pair of adjacent regions with difference Δ can be found, a decision has to be taken to select, among the encountered pairs, the pair where the separation between the foreground and the background has to be placed. We select the pair for which $r^k{}_i$ is the greatest one, so as to favor assignment of most of the slope to the foreground.

Once all regions have been classified, a final local process is accomplished, aimed at possibly changing the classification status of some regions that have been classified as belonging to the background during the iterative classification process, and are placed at the border with respect to foreground components along the slopes. This final process depends on the problem domain. If the purpose is to favor region growing without merging already detected foreground components, the change of status is done only if it does not cause a topology change. In turn, if clusters of foreground components are desired, e.g., to analyze the spatial organization of the foreground, the change of status is done only if it causes a topology change.

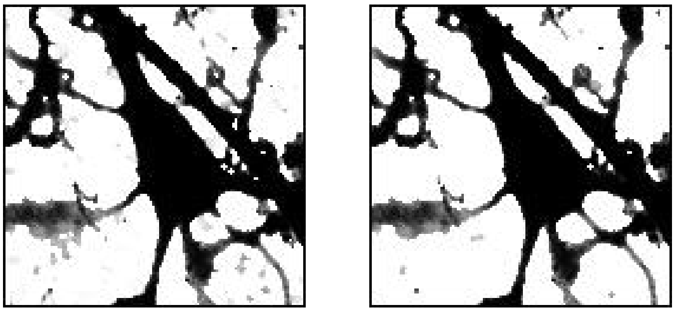


Fig. 3. Pixels classified as belonging to the foreground, starting from the standard watershed partition, left, and from the partition obtained by using the algorithm [10], right.

In Fig.3, the result of the classification process is shown for the running example, starting from the standard watershed partition, left, and from the more sophisticated watershed partition [10]. Foreground pixels are shown with their original gray-levels, while all background pixels have been set to 255. Both results can be regarded as satisfactory, even if an obviously more accurate segmentation is obtained by using the partition obtained by the method [10]. As already pointed out in Section 2, the method [10] is computationally more expensive. Thus, the choice of which partition to adopt depends on a compromise between quality of the results and cost of the process.

4. Reducing oversegmentation by multi-scale representation

We describe here an alternative way to reduce oversegmentation, based on the use of a multi-scale image representation. This method requires that the standard watershed transform be computed only twice, while it has to be computed for a larger number of times if the algorithm [10] is used. When observing a gray-level image at different scales, the most significant regions are perceived at all resolutions, while regions with lower significance, e.g., fine details, are perceived only at sufficiently high resolution. Thus, if the seeds for watershed segmentation of the gray-level image are detected in a representation of the image at a lower resolution with respect to the full resolution of the input image and are, then, used to distinguish significant and non-significant seeds in the image at full resolution, the resulting partition is expected to consist mainly of the regions that are perceived as the most significant ones.

To build a lower resolution representation of the input image I, we superimpose onto I a partition grid, each cell of which includes a fixed size block of pixels (*children*). We associate to each cell of the grid a single pixel (*parent*) in the representation of I at lower resolution, I'. The gray-level of a parent pixel is computed in terms of the gray-levels of its children. Depending on the position of the grid, the size of the cells, and the rule used to compute the gray-level of parent pixels, different lower resolution representations can be obtained. We here use the grid introduced in [11], whose cells are blocks of 2×2 pixels, so that the size of I’ is a fourth of the size of I. The rule we adopt to compute the gray-level of the parent pixels is such to produce an almost shift invariant lower resolution image representation. Moreover, the parent-child relations are preserved, so that it is easy to transfer onto the full resolution image I, the information derived by analyzing its lower resolution representation I’.

More in detail, we inspect in forward raster fashion only pixels belonging to even rows and columns of I. This means that we use the bottom right child pixel in the 2×2 block to find the coordinates of its parent pixel in I’. For each inspected pixel in position (i,j) of I, the parent pixel in I’ will be in position (i/2,j/2).

As for the gray-level of the parent pixel in position (i/2,j/2) of I’, we note that the sampling grid could be placed on I in four different ways and, hence, any pixel in the 3×3 window centered on (i,j) could be the bottom right pixel of a block of the partition grid. If we consider the nine 3×3 windows, that in I are respectively centered on (i,j) and on each of its eight neighbors, then the pixel in position (i,j) is included in all the nine windows, its edge-neighbors are included in six windows and its vertex-neighbors in four windows.

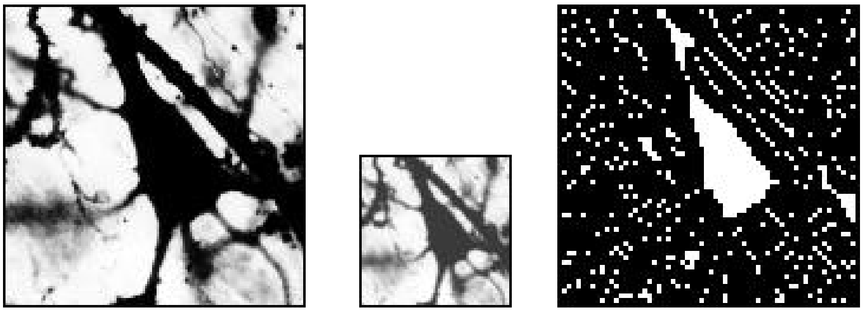


Fig. 4. Full resolution image, left, lower resolution image, middle, and markers, right.

We use the above numbers 9 for the pixel in (i,j), 6 for its edge-neighbors, and 4 for its vertex-neighbors, as the proper weights to be used in a multiplicative mask to compute the gray-level of the parent pixel (i/2,j/2) of I’. By using the mask, we take into account the gray-levels of the pixel (i,j) and of its eight neighbors in a manner independent of the position of the grid. Rescaling of the computed gray-levels is done to have them still in the range [0, 255].

In Fig. 4 middle, the representation of the running example at lower resolution is shown. The full resolution image is given to the left, for the reader’s convenience.

Since I is well represented by its lower resolution representation I’, we can use the seeds detected in the gradient image of I’, ∇', as markers to select the significant seeds in the gradient image of I, ∇. Due to the preservation of the parent-child relations, we can easily project the seeds found in ∇' onto a full resolution image. Since any parent pixel has four children, for each seed found in ∇' we identify a projected seed consisting of the union of 2×2 blocks of pixels in ∇. See Fig. 4 right.

We regard a seed detected in ∇ as significant, if the partition region associated with it in the standard watershed transform includes at least one pixel of a projected seed. Seeds originally detected in ∇, but such that the associated partition regions of the standard watershed transform do not include any pixel of projected seeds are regarded as non-significant. By means of a *flooding* process, the partition regions of the standard watershed transform corresponding to non-significant seeds are merged to adjacent regions. In practice, the gray-level of the non-significant seeds is suitably increased, so that those pixels will not be newly identified as regional minima, when the watershed transformation is applied for the second time to obtain the final partition.

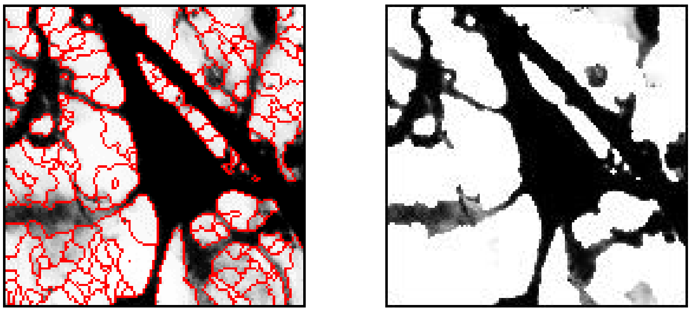


Fig. 5. Final result of the process to identify foreground components.

The results of using multi-scale representation are shown in Fig. 5. The watershed lines partitioning the running example into only 153 regions are shown to the left, and the foreground components detected by using the process described in Section 3 are shown to the right, superimposed onto a uniform background.

For completeness, we point out that the resolution of the image I' could be furthermore reduced by applying to I' the same decimation process that we have applied to I. By using an image I' with even lower resolution, the number of significant seeds detected as significant in the full resolution image is expected to diminish. However, this could produce a too rough segmentation of the input gray-level image.

5. Conclusion

We have introduced a segmentation method based on a preliminary partition of a gray-level image into regions by means of the watershed transformation. The partition regions have been classified in two classes (foreground and background) by taking into account only gray-level information. Our segmentation method has been tested on a variety of images in different domains even if, in this paper, only one running example relative to biological images has been shown. The method is suited to gray-level images, where the boundary between foreground and background is perceived in correspondence with the locally maximal changes in gray-level through the image.

We have shown the classification results obtained starting from standard watershed transform, i.e., without resorting to seed selection and region growing. When this is done, the computational burden of the whole segmentation process is rather limited.

Better results are obtained, at the expense of a higher computational cost, if the gray-level image is partitioned by using a more sophisticated watershed transformation, including digging and flooding techniques to produce a less fragmented partition of the image. This more complex procedure is necessary when a finer segmentation is indispensable. We have also suggested an alternative way to reduce oversegmentation, by using multi-scale image representation. A lower resolution representation of the input image is built and the seeds for watershed partition found in this image are used as markers to discriminate between significant and non-significant seeds in the full resolution image. Segmentation done by using this approach has a cost that is intermediate between the cost of segmentation based on the standard or a more sophisticated watershed transformation, still producing good results.

References

1. H.D.Cheng, X.H.Jiang, Y.Sun, J.Wang, "Color image segmentation: advances and prospects", *Pattern Recognition*, 34, 2259-2281, 2001.
2. D.L.Pham, C.Xu, J.L.Prince, "Current methods in medical image segmentation" *Annual Review of Biomedical Engineering*, 2, 315-337, 2000.
3. P.K.Sahoo, S.Soltani, A.K.C.Wong, Y.C.Chen, 'A survey of thresholding techniques', *Comput. Vis. Graph. Im. Proc.*, 41, 233-260, 1988.
4. P.Perner, "An architecture for a CBR image segmentation system", *Journal of Engineering Application in Artificial Intelligence, Engineering Applications of Artificial Intelligence*,12-6, 749-759, 1999.
5. J.C.Bezdek, L.A.Hall, L.P.Clarke, "Review of MR image segmentation techniques using pattern recognition", *Med. Phys.*, 20, 1033-1048, 1993.
6. R.M.Haralick, L.G.Shapiro, "Image segmentation techniques", *Comput. Vis. Graph. Im. Proc.*, 29, 100-132, 1985.
7. R.C.Gonzalez, P.Wintz, *Digital Image Processing*, Addison-Wesley, Reading, Massachusetts, 1987.
8. S. Beucher, F. Meyer, 'The morphological approach of segmentation: the watershed transformation', in Dougherty E. (Ed.) *Mathematical Morphology in Image Processing*, Marcel Dekker, New York, 433-481, 1993.
9. P.Perner, A. Bühring, "Case-based object recognition", in P.Funk and P.A.Gonzalez Calero (eds.) *Advances in case-based reasoning,* Proceedings ECCBR 2004, Madrid, Spain, Springer-Verlag, 375-388, 2004.
10. M.Frucci "Oversegmentation reduction by flooding regions and digging watershed lines", *International Journal of Pattern Recognition and Artificial Intelligence*, 20-1, 15-38, 2006.
11. M.Frucci, G.Ramella, G.Sanniti di Baja, Oversegmentation reduction via multiresolution image representation, in M.Lazo and A.Sanfeliu (eds.), *Progress in Pattern Recognition, Image Analysis and Applications,* Springer, Berlin, LNCS 3773, 989-996, 2005.

Acknowledgments

We gratefully acknowledge Dr. Vittorio Guglielmotti (Neuroanatomy Research Group, Institute of Cybernetics, CNR, Pozzuoli, Naples, Italy) for kindly providing the images that we used as input to test our procedure.

Finding cells, finding molecules, finding patterns

Carolina Wählby[1,2], Patrick Karlsson[1], Sara Henriksson[2], Chatarina Larsson[2],
Mats Nilsson[2], and Ewert Bengtsson[1]

[1] Centre for Image Analysis, Uppsala University, Sweden,
carolina@cb.uu.se, http://www.cb.uu.se

[2] Dept. of Genetics and Pathology, Uppsala University, Sweden,
http://www.genpat.uu.se

Abstract. Many modern molecular labeling techniques result in bright point signals. Signals from molecules that are detected directly inside a cell can be captured by fluorescence microscopy. Signals representing different types of molecules may be randomly distributed in the cells or show systematic patterns indicating that the corresponding molecules have specific, non-random localizations and functions in the cell. Assessing this information requires high speed robust image segmentation as well as signal detection, and finally pattern analysis. We present and discuss this type of methods and show an example of how the distribution of different variants of mitochondrial DNA can be analyzed.

1 Introduction

Data mining can be defined as the science of extracting useful information from large data sets. In this case, the input data is digital images of cells captured using fluorescence microscopy, and the information we aim to retrieve is that of spatial distribution patterns of different variants of fluorescence labeled molecular targets. New probing and staining techniques allow a large variety of molecular targets to be visualized in situ and imaged by fluorescence microscopy. Biological processes can be studied at the ultimate level of single molecules, and with sufficient precision to distinguish even closely similar variants of molecules. It is thus possible to study the inter- or subcellular context of molecules that otherwise may go undetected at the level of populations of molecules and cells. At the same time, large numbers of cells have to be analyzed to retrieve statistically significant information. Extracting information from the resulting image data will therefore require efficient and robust cell segmentation as well as signal detection and, finally, pattern analysis.

Before signals can be assigned as coming from a particular cell, each cell has to be delineated. Segmentation is the process in which an image is divided into its constituent objects, or parts, and background. Cells can be visualized in many different ways, using different kinds of probes or stains that bind to structures within a cell. It is therefore difficult to define a single algorithm that will always find the individual cells in an image, independent of method for visualization and cell morphology. Instead, cell image segmentation can be seen as a modeling problem where different approaches more or less explicitly are based on models of the cells. For example, thresholding methods can be seen as being based on a model stating that cells have an intensity that is different from the surroundings. More robust segmentation can be obtained if a combination of features, such as intensity, edge gradients, and cellular shape, is used.

In many applications in cell biology, where fluorescence marked probes are applied, the resulting images are composed signals seen as spots of different shapes and intensities. The localization of these regions can yield important biological information. In multiple labeling experiments in particular, measurements of relative positions of regions labeled with different marker molecules can provide insight in the functional relationship between organelles and/or processes. Visual inspection is, apart from being tedious, beset with various sources of error. The positions of signals in an image should be determined automatically to derive objective information and allow further extraction of image information, such as signal intensity distribution, relative positioning and pattern analysis. The human mind is exceptional at finding patterns, it will even find patterns in data that is completely random. It is therefore valuable to have computerized methods that can search for patterns in a more objective way.

We present an image based data mining example where the distribution of different variants of the genetic information contained in mitochondria (i.e., mtDNA) has been examined. MtDNA is present in multiple copies in the mitochondrion of the cell. It is inherited together with the cytoplasm during cell replication. Genetic diseases are often caused by mutations where one single nucleotide has been substituted by another, a so-called point mutation. To be able to study and diagnose such disease with limited material from patients, there is a need for methods to detect point mutations in situ. Padlock probes and rolling circle amplification (RCA) combines highly specific target sequence recognition with a high signal-to-noise ratio. Padlock probes have been successfully used for detecting point mutations in mitochondrial DNA [10]. We combine cell segmentation, padlock probes, signal detection and pattern analysis to examine the distribution of mtDNAs. This type of methods could also be used in applications ranging from detection of infectious organisms to studies of tumors.

2 Methods

The methods section is divided into three parts, describing methods for segmentation of cells, detection of signals, and evaluation of patterns in the detected signal distribution. A specific example is thereafter brought up in the Experiments and results section.

2.1 Cell segmentation: finding cells

The difficulty of the segmentation problem is highly dependent of the type of specimen that is to be analyzed, and the result of the segmentation usually determines eventual success of the final analysis. If we are dealing with cytological specimens where the cells are lying singly on a clean background with well stained nuclei, and if the analysis task is limited to nuclear properties, then a simple automatic thresholding method may be sufficient. Thresholding is often based on histogram characteristics of the pixel intensities of the image [19]. In order to get a satisfactory segmentation result by thresholding, a sufficiently uniform background is required. The transition between object and background may be diffuse, making an optimal threshold level difficult to find also after background correction. At the same time, a small change in the threshold level may have a great impact on the further analysis; feature measures such as area and volume are directly dependent on the threshold. Adaptive thresholding, i.e., local automatic thresholding, can be used to circumvent the problem of varying background or as a refinement to a coarse global threshold [17]. The problems of segmenting clustered objects and choosing a suitable threshold level for objects with unsharp edges will, however, remain.

If we model the objects as consisting of connected regions of similar pixels we obtain region growing methods. A popular region growing method which has proved to be very useful in many areas of image segmentation is the so called watershed algorithm. The method was originally suggested by Digabel and Lantuéjoul, and extended to a more general framework by Lantuéjoul and Beucher [3]. Watershed segmentation has then been refined and used in very many situations, see [14, 20] for an overview. If the intensity of the image is interpreted as elevation in a landscape, the watershed algorithm will split the image into regions similar to the drainage regions of this landscape. To avoid over-segmentation, i.e., splitting of the image into too many regions, water can be allowed to rise only from places marked as seeds [2, 9, 11, 14, 20]. Seeds may be found manually or by automated methods. Over-segmentation can also be reduced by rule-based merging [15].

Cell nuclei are usually convex and fairly round or elliptic and the shape can therefore be used as part of the object model. Touching nuclei that are not separated by an intensity threshold can be separated by distance transforming [4] the binary image and applying watershed segmentation [12, 17, 21].

None of the above described methods will alone produce a satisfactory result on the more difficult types of cell and tissue images. We may for instance have problems if (1) the cells are clustered, (2) the image background is varying, and (3) there are intensity variations within the cells. By combining the methods, more powerful models can be created, and more complex segmentation problems be solved. Our experience is that the seeded watershed approach is a useful core component in such segmentation models.

2.2 Signal detection: finding molecules

The most common method for finding structures such as proteins and organelles in situ is using antibodies labeled with fluorescing molecules. Fluorescence labeled secondary antibodies can be used to amplify the signal and increase signal to noise ratios. The genetic information contained in the DNA in a cell can be stained as a whole using non-specific chemical dyes, or in a more specific way using oligonucleotide probes that search for a particular DNA sequence. Fluorescence in situ hybridization (FISH) is such a method, and can detect larger mutations such as duplications, translocations and deletions, but it is not sensitive enough to distinguish between single nucleotide sequence variations. Primed in situ labeling (PRINS) [7] reaction uses a specific primer that will initiate synthesis of DNA from fluorescent labeled nucleotides at the site of sequence detection. The method does however not give signals from single-copy genes that are distinguishable from noise caused by insertions of fluorescing nucleotides in other places in the genome. In the oligonucleotide ligation assay (OLA) [8] oligonucleotides are hybridized juxtaposed with the junction at the point mutation. If there is a perfect match, the two probes can be enzymatically hybridized and detected. There is however a risk of wrong probes being ligated, especially when trying to find many different sequence variants in the same sample. This can be avoided by, instead of using two separate probes, using a single linear probe, a so called padlock probe. The padlock probe has ends that are designed to hybridize juxtaposed at the point mutation and if correctly base paired at the point mutation, the two ends can be enzymatically ligated, forming a circular DNA molecule [16]. The specifically reacted circular DNA can thereafter be amplified using rolling circle amplification (RCA) [1] generating molecules that are bound by hundreds of fluorescing probes. These signals can be detected by fluorescence microscopy as bright spots at or below the resolution of the microscope, the image resolution limited by the point spread function of the microscope.

An image that contains multiple, and sometimes clustered spots with different maximum intensities can be segmented in many different ways. Regions found by procedures such as intensity thresholding often contain more than one local maximum of intensity, indicating that the region consists of more than one spot. Top-hat transforms [5] in combination with threshold procedures fail to divide the image into separate domains each containing one local maximum of intensity as the top-hat transform is unable to distinguish a local maximum from a saddle-point. If each spot contains a single local maximum, watershed segmentation, as described above, in combination with a background threshold, may be used to delineate individual signals. Another approach is the largest contour segmentation [13], where the domain of each signal is defined by a local maximum and an iterative region-growing. If two or more signals are clustered into a spatially large signal, where the individual signals do not contain individual intensity maxima (due to tight clustering or signal saturation), the shape of the signal can provide clues as to how the signals should be detected. In [6], the curvature of the edge of each signal cluster is examined, and signals are positioned within the cluster starting from the position where the greatest curvature is found.

2.3 Hypotheses testing: finding patterns

Patterns in image data can be evaluated by interpreting the signal distribution as image texture, and using texture measurements. Some of the most commonly used texture measures are derived from the Grey Level Co-occurrence Matrix (GLCM). The GLCM is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in a pixel pair in an image [5]. Different kinds of distance measures can also be used to evaluate spatial relationships between signals once they have been detected. In the case where we want to know if red and green signals are randomly distributed in the cytoplasm or not, we can simply count how often a red signal has a green signal as its closest neighbor, and how often a red signal has a red signal as its closest neighbor (and the other way around). To evaluate the outcome, we have to know what distributions can be expected. By creating a virtual cell, where possible positions and number of signals of different types is given as input, different hypotheses can be tested. We can then compare the spatial relationships between signals in real cells with those in a virtual cell with the same input parameters. Signals in the virtual cell can be positioned based on a hypotheses, i.e., either randomly, or according to a pattern. Thousands of randomized virtual cells can thereafter be created, and the probability of the real cell having the hypothesized signal distribution pattern can be examined. Factors such as staining efficiency and noise may also be added to the virtual cell for comparison.

3 Experiments and results

To illustrate the concepts discussed we will here describe a project where model based cell segmentation is combined with padlock probing for molecule detection, model based signal detection, and pattern analysis to examine the spatial distribution of mtDNAs.

3.1 Finding cells

In the presented experiment, no general stain defining the cytoplasm is available. We do however have a general stain defining the nucleus of each cell. Combining this information with the fact that the over all signal variance is higher within the cytoplasm than in the image background, a model defining cytoplasms is created. The three markers (i.e. nuclear stain, padlock probe 1, and padlock probe 2) are shown as three images, see Fig. 1 A, B, and C, each captured with a different filter set in the fluorescence microscope. Cells segmentation is initiated by intensity thresholding of the image showing the nuclear stain. A suitable threshold is found using Otsu's method, which searches for the threshold level that minimizes the intra-class variance of foreground as well as background [18]. The resulting binary image is shown in Fig. 1 D. Intensity thresholding is not enough to separate nuclei that are very close to one another. Thanks to their round shape, touching nuclei can be separated by applying watershed segmentation to a distance map of the binary image. Over-segmentation due to multiple local maxima is avoided by smoothing of the distance map. The distance map is shown in Fig. 1 E, and the result after watershed segmentation is shown in Fig. 1 F. The region surrounding each nucleus, not belonging to the image background, is the cytoplasm. The image background has less variance than the parts of the image containing cells, and can thus be found by variance filtering. The variance map of the sum of B and C (Fig. 1 G) is thresholded using Otsu's method. Small, disconnected regions are removed by morphological opening, and larger disconnected regions are re-connected by dilation. In both cases, a disc-shaped structure element of radius 10 pixels was used. The choice of structure element depends on size and density of signals.

Each pixel of the resulting binary image (Fig. 1 H) is assigned to the closest nucleus by seeded watershed segmentation, using the segmentation result from the segmentation of the nuclei as seeds. Non-seeded regions are discarded as background. Fig. 1 I shows the final segmentation result on top of a projection of the three input images.

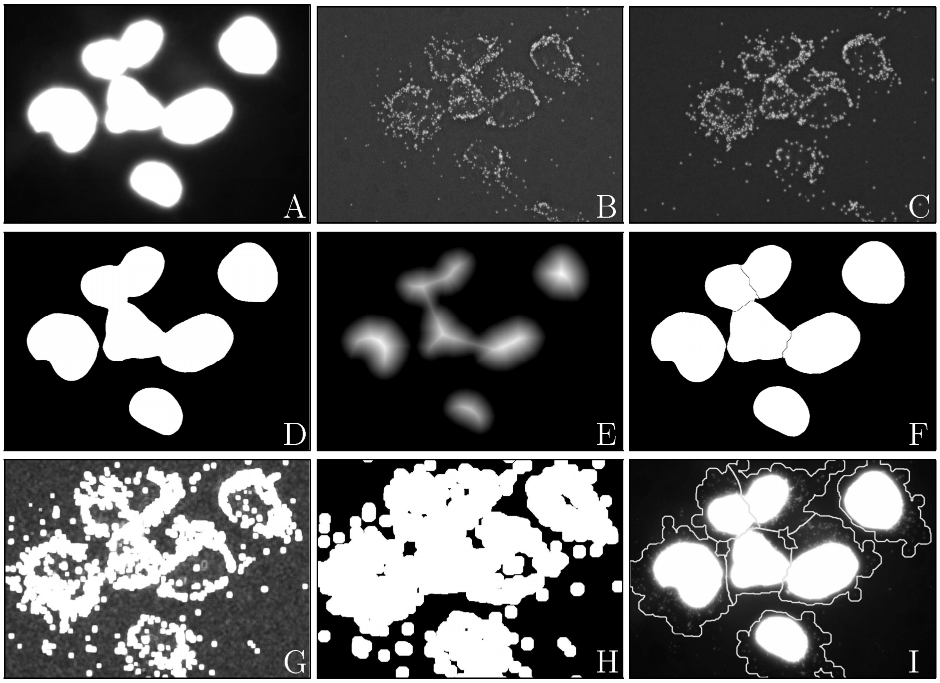


Fig. 1. A: Part of an image of DAPI stained cell nuclei. **B**: Image of the same cells showing logarithm of signals from padlock probe 1 (stained with Cy3) and **C**:, logarithm of signal from padlock probe 2 (stained with FITC). **D**: Binary image after thresholding of **A**. **E** shows the same image after distance transformation, and **F** is the result after watershed segmentation on the distance map, i.e., the final segmentation of the nuclei. **G**: The cytoplasm is found by combining the two images showing padlock probes (**B** and **C**) and applying a variance filter. **H**: Potential cytoplasm after thresholding of variance map and morphological opening to remove noise. **I**: Final segmentation result based on shape of nuclei and variance of cytoplasm.

3.2 Finding molecules

In the presented experiment, a model system with padlock probes was used. It consists of different detection sequences that represent real point mutations. Four different padlock probes were used for testing efficiency of staining and evaluation of signal distribution patterns. Two of the padlock probes hybridize to different sites on the same mtDNA fragment, i.e., they are non-competing. One is detected using Cy3 (red), and one using FITC (green). The other two probes bind to the same site, and are therefore competing.

Signal detection was initiated by first reducing the background variation present in the images. As the cells are cultured on a glass surface, they are comparably flat. Despite this, it is necessary to image them in more than one focal plane to make sure that all signals are detected. In the presented study, the slides were studied in a fluorescence microscope (Axioplan II Zeiss) using a 63x objective. Images were collected with Axiovision 4.3 software as a 16 layers z-stack with 0.5um between consecutive layers. The nuclear stain DAPI emission was collected at 360nm excitation wavelength for 200ms, green padlock detection fluorochrome FITC at 470nm for 200ms and red padlock detection fluorochrome Cy3 at 546nm for 450ms. Background was reduced in each z-image separately by morphological tophat filtering, using a disc of radius 10 pixels. Tophat filtering removes intensity variations that have a spatial extent greater than that of the disc. The 16 layers were thereafter combined using maximum intensity projection. Projection of the 3D information to a single image will result in loss of spatial information in the z-direction. As the extension of cultured cells in z-direction is only a fraction of their extension i x-, and y- direction, the 3D information was considered less important. This would however not be the case if cells in tissue were observed.

The result after pre-processing and maximum intensity projection of a small fraction of an image is shown in Fig. 2 A. Simple intensity thresholding will separate the signals from the image background, but signals that are clustered will not be separated from each other. To separate clustered signals, Watershed segmentation, starting from all local maxima, is applied to the image, and the watershed regions are allowed to extend until they reach a predefined background threshold. The resulting signal centers after watershed segmentation of Fig. 2 A are shown in Fig. 2 B.

3.3 Finding patterns

The patterns of red and green signal distributions were examined by searching for aggregations of signals, i.e., the existence of groups of signals with the same color. The affinity of red and green signals was measured as the number of red signals with a green nearest neighbor, the number of red signals with a red nearest neighbor, the number of green signals with a green nearest neighbor, and finally the number of green signals with a red nearest neighbor. To normalize the observed result and evaluate the probability of non-random pattern, virtual cells with truly random patterns were created. Virtual cells with random signal

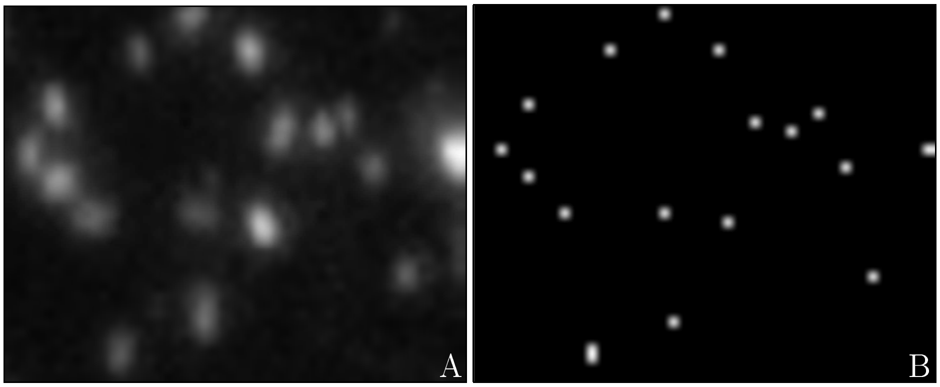


Fig. 2. **A**: Enlarged image showing signals from padlock probes after image preprocessing. **B**: Positions of detected signals using watershed segmentation.

distributions were created by keeping the number of red and green signals the same as in the real cell. Red and green signals were then randomized within an area corresponding to the cytoplasm region with the 10% greatest variance. Virtual cells were re-created 1000 times, neighborhood relations were examined, and the resulting distributions were compared with observed distributions.

A number of restrictions have to be taken into consideration when creating the virtual cells. First of all there is a limit in closeness between signals in the real data due to the point-spread function. Two signals that are of the same color will not be separated if they are closer than the width of a single signal. This has to be compensated for in the randomized data, or else it will affect the outcome of the analysis of the neighbor relations. Randomized signals that appeared closer to one another than the two closest signals in the real data were simply removed, and a new pair of random signals was created and tested for closeness with existing randomized signals. Fig. 3 A, top, shows the true signal distribution within the cytoplasm of a cell with competing padlock probes, red and green signals as + and o respectively. Fig. 3 A, bottom, shows one of the 1000 virtual cells with randomized signals. Fig. 3 B, top, shows the true signal distribution within the cytoplasm of a cell with non-competing padlock probes, red and green signals as + and o respectively. Fig. 3 B, bottom, shows one of the 1000 virtual cells with randomized signals in the non-competing case. As can be seen, it is not trivial to pick out the cells showing random distributions compared to those showing a non-random affinity between red and green signals. Comparing the randomized data with the true signal distributions shows that the pattern falls within the randomized distribution in the case with competing probes, while the non-competing probes show a red-green affinity three standard deviations greater than that of the randomized distribution. This agrees with what one would expect as the non-competing probes can bind to the same mtDNA fragment.

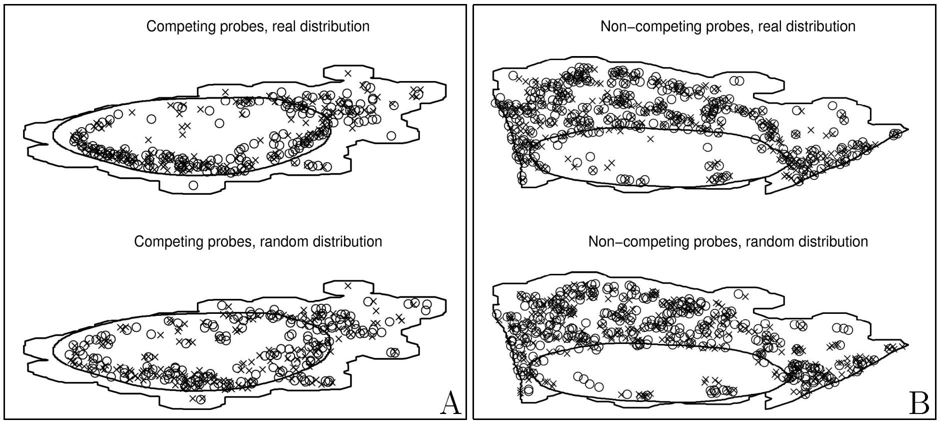


Fig. 3. Cells treated with two competing (**A**) or non-competing (**B**) padlock probes in a 50-50 concentration. Top: true signal distribution, bottom: randomized signal distribution. Red and green signals as + and o respectively. Nucleus and cytoplasm outlined.

4 Conclusions

Patterns and spatial relationships between molecules in cells are of great interest in many types of analysis. One way of examining patterns in cells is by visualizing the molecules of interest using highly specific detection probes, and comparing observed signal distributions with randomized distributions in a virtual cell. Before patterns can be examined, signals from detection probes have to be found, and clustered signals separated. More than one cell is often observed simultaneously, and automated identification of each individual cell in an image provides efficient analysis of large data sets with little impact from observer bias. In order to obtain a successful cell segmentation method it is important to use as much a priori information as possible about the appearance of the objects that are to be segmented, without resorting to models that are too complex or too difficult to train or apply.

References

1. J. Banér, M. Nilsson, M. Mendel-Hartvig, and U. Landegren. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acid Research*, 26:5073–5078, 1998.
2. S. Beucher. The watershed transformation applied to image segmentation. *Scanning Microscopy*, 6:299–314, 1992.
3. S. Beucher and C. Lantuéjoul. Use of watersheds in contour detection. In *International Workshop on Image Processing: Real-time and Motion Detection/Estimation*, Rennes, France, Sept. 1979.

4. G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics and Image Processing*, 34:344–371, 1986.
5. R. M. Haralick and L. Shapiro. *Computer and Robot Vision*, volume 1 and 2. Addison-Wesley, REading, Mass, 1992.
6. P. Karlsson and J. Lindblad. Segmentation and separation of point like fluorescent markers in digital images. In *Proceedings of 2004 IEEE International Symposium on Biomedical Imaging*, Washington D.C., USA, 2004.
7. J. Koch, S. Kolvraa, K. Petersen, N. Gregersen, and L. Bolund. Oligonucleotide-priming methods for the chromosome-specific labeling of alpha satellite dna in situ. *Chromosoma*, 98:259–265, 1988.
8. U. Landegren, R. Kaiser, J. Sanders, and L. Hood. A ligase-mediated gene detection technique. *Science*, 241:1077–1080, 1988.
9. G. Landini and I. E. Othman. Estimation of tissue layer level by sequential morphological reconstruction. *Journal of Microscopy*, 209(2):118–125, 2003.
10. C. Larsson, J. Koch, A. Nygren, G. Janssen, A. K. Raap, U. Landegren, and M. Nilsson. In situ genotyping individual dna molecules by target- primed rolling-circle amplification of padlock probes. *Nature Methods*, 1:227–232, 2004.
11. S. J. Lockett, D. Sudar, C. T. Thompson, D. Pinkel, and J. W. Gray. Efficient, interactive, and three-dimensional segmentation of cell nuclei in thick tissue sections. *Cytometry*, 31:275–286, 1998.
12. N. Malpica, C. Ortiz de Solorzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M. Garcia-Sagredo, and F. del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28(4):289–297, 1997.
13. E. M. M. Manders, R. Hoebe, J. Strackee, A. M. Vossepoel, and J. A. Aten. Largest contour segmentation: a tool for the localization of spots in confocal images. *Cytometry*, 23:15–21, 1996.
14. F. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image Representation*, 1(1):21–46, 1990.
15. L. Najman and M. Schmitt. Geodesic saliency of watershed contours and hierarchial segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(12):1163–1173, 1996.
16. M. Nilsson, H. Malmgren, M. Samiotaki, M. Kwiatkowski, B. P. Chowdhary, and U. Landegren. Padlock probes: Circularizing oligonucleotides for localized dna detection. *Science*, 265:2085–2088, 1994.
17. C. Ortiz de Solorzano, E. Garcia Rodriguez, A. Jones, D. Pinkel, J. Gray, D. Sudar, and S. Lockett. Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *Journal of Microscopy*, 193:212–226, 1999.
18. N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. on System Man and Cybernetics*, 9(1):62–69, 1979.
19. P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen. A survey of thresholding techniques. *Computer Vision, Graphics and Image Processing*, 41:233–260, 1988.
20. L. Vincent. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE Trans. on Image Processing*, 2(2):176–201, 1993.
21. C. Wählby, I.-M. Sintorn, F. Erlandsson, G. Borgefors, and E. Bengtsson. Combining intensity, edge, and shape information for 2D and 3D segmentation of cell nuclei on tissue sections. *Journal of Microscopy*, 215(1):67–76, 2004.

An automatic real-time system for the determination of translational and rotational speeds of swimming microorganisms

Mauro Evangelisti+, Primo Coltelli*, Valtere Evangelista+, Paolo Gualtieri +°

* CNR ISTI, via Moruzzi 1, 56124 Pisa Italy
+ CNR Istituto di Biofisica, via Moruzzi 1, 56124, Italy
° to whom correspondence should be addressed

Abstract. This paper describes a digital system designed for the automatic detection and measurement of the velocity of moving objects in images acquired by means of a common TV-camera mounted onto a microscope. The main characteristics of this system are the following: 1) it can perform a real-time gray level difference between two successive frames in order to detect moving objects and to suppress stationary objects (subtraction procedure); usually the delay between two successive frames varies linearly from 40 msec to 1920 msec; 2) it reduces the size of images resulting from the subtraction procedure (difference images) and stores them in the frame memory; the result of these operation, all performed in real-time, is a film of time sequences; 3) it performs an automatic labelization in order to recognize the moving microorganisms and to calculate their area in each difference image; 4) it calculates and plots the variation of the average area of the cells moving in the microscope field; 5) it completes the analysis in few seconds.

1 Introduction

The exact determination of the speed parameters of swimming microorganisms can be a very useful tool for the study of both behavioral and physiological aspects of motility, that is an essential. Speed parameters can be obtained by means of photomicrographic [1] and cinematographic techniques [2]. These methods, however, turn out to be time consuming. Statistical counting techniques can be utilized as well [3], but they are tedious, and prone to human errors. Other methods utilize more sophisticated techniques such as spectral analysis of the light scattered by the microorganisms [4], or analogic elaboration of the video signal of a TV-camera mounted onto a microscope [5]. In these cases, however, the instrumentation necessary for speed parameter determination has the drawback to be not portable. A further alternative is represented either by the digital tracking microscope, which can determine speed parameters by reconstructing the entire movements of swimming microorganisms [6], [7], [8] or by the simple method we will describe. This method automatically and in real time determines the speed of swimming microorganisms. The biflagellate algae *Dunaliella salina* has been used as experimental subject. Substantially, this method utilizes the subtraction operation, which has already been used by other authors for the detection of motion [9]; however our procedure performs automatically and in real-time both the detection of the moving microorganisms and the determination of their speed parameters. Our results are consistent with previous published speed data of *Dunaliella salina* obtained with the other methods [10].

2 Materials and Methods

A Pulnix TM860 (Pulnix, USA) CCD video camera was mounted onto a Zeiss Axioplan microscope (Zeiss, Germany) equipped with 16x and 60x objectives and 100W halogen lamp as light source. Cells were placed in a small chamber obtained by fixing a PVC ring onto a microscope slide. The chamber was closed by means of a cover slip so as to avoid sample drying-out. The microorganisms can freely swim within a narrow layer of growth medium placed between a slide and a cover slip.

The signal of the camera was the input of a FG100 AT Frame Grabber (Image Technology, USA) plugged into a Pentium V personal computer 750MHz clock. For the translational speed determination experiment, a sequence of images taken at known intervals of time was acquired, stored, and processed using the automatic procedure of Gualtieri and Coltelli [11]. For the rotational speed experiment, the light reflected by the cell eyespot was measured. The experimental set-up was the same used previously, with the addition of a custom-made slide. This custom-made slide allowed the lateral illumination of the cell sample by means of an optical fiber delivering the light coming from a Schott KL1500 fiber optic illuminator (Schott, Germany).

Photographs were recorded with an Olympus Camedia C-30303 digital camera (Olympus, Japan) mounted on the Zeiss Axioplan microscope (Zeiss, Germany).

3 Operation Procedures

Real time detection of microorganisms under the microscope is performed by differencing continuously each frame of the video image from a previous frame, with a variable delay, during the acquisition process. This operation is made possible by programming the 12-bit input Look-up-table (LUT) of the board. This LUT, which is located between the digitization circuit and the frame memory, transforms the image before it is stored into the frame memory. Thanks to a feed-back circuit between the frame memory and the LUT, operations are made on combinations of stored and newly-acquired data. We program the LUT in order to move the six most significant bits from the A/D converter (the newly acquired data) to the six most significant bits in the frame memory; then the LUT subtracts the same six most significant bits of

A/D data from those previously stored in the six most significant bits of the frame memory (the previous frame); the resulting six bits are then stored in the least significant bits of the frame memory. The resulting difference image is always available in the lower six 1-bit planes of the frame memory; while the upper six 1-bit planes contain the most recent data of the A/D converter, which are used as input for the next frame subtraction. In the case that the images of a moving cell in two successive frames are partly overlapping, the subtraction operation gives a zero value for the overlapping region of the cell and for the background, a negative value for that part of the cell image which is present only in the previously acquired frame and a positive value for that part of the cell image which is present only in the newly acquired frame. In order to follow the increasing of the cell image, which will increase up to the whole cell size during the delay progression, we program the LUT to clip to zero the negative value, (Fig. 1).

In order to store several difference images in real time in the frame memory, we could reduce the spatial resolution of the image being acquired by means of the hardware Zoom. In order to store the (reduced) image into its proper position of the frame memory, the X and Y coordinates of its origin are shifted in real time by means of Pan and Scroll operations. In this way we can store images, by moving the coordinates of their origin toward the right and downward. At the end of this procedure, the frame memory is displayed as a patchwork of (reduced) images. In order to identify moving cells and to extract its features such as baricenter coordinates, contours, axis and areas, a segmentation and labelization procedure is applied to each difference image [11]. This procedure lasts 2 seconds for the whole memory.

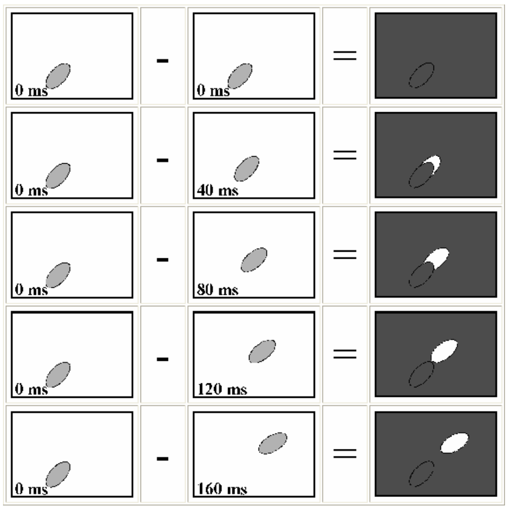


Fig. 1. Subtraction operation.

4 Results and Discussion

Fig. 2 represents 10-images time sequence (400 msec). Each difference image is represented as a framed image. The number visible in the first column represents the delay between the two frames on which the system has performed the difference. The system acquires a couple of frames utilizing for each frame six bits of the frame memory and performs the subtraction operation as previously explained. Usually we choose a delay that varies linearly, but the delay can progress in a different way as well. In our case, because of the CCIR standard, the delay between two successive frames is 40 msec, or a multiple of 40 msec. The first image (40 msec) represents the real time difference between the first acquired frame and the second acquired frame; the second image (80 msec) is the real time difference between the third acquired frame and the fifth acquired frame. The position of the frames in the successive couples can be easily extrapolated by the delay number. After subtraction every difference image is placed in its proper position of the frame memory in real time by means of pan and scroll operations. For the determination of the translational speed value of the cell we have to measure the distance covered by the cell and the time lapse; if no stimuli are applied to the environment, the swimming speed of the cells can be considered constant. Therefore, the time a cell takes to cover a distance equal to its long axis can be used for the determination of its speed. As the difference procedure presented in Fig. 1 suggests, until the cell doesn't cover a distance equal to its size, the area value will be lower than the real one. Greater the delay between the frames, less two successive images of the cell are superimposed; there is a delay for which the subtraction operation gives two separate images of the same cell, i.e. the area value of the cell is the real value. A higher delay between the two frames still separates the two images of the cell, but the area value of the cell will remain constant. In Fig. 2, the cells, which are represented by the whitish areas, can be hardly recognized in the first frames because the difference between two successive images of the same cell consists of a small agglomerate of pixels. In the last frame of the same figure, the cells are instead easily recognizable, because in this case the difference between the two images of the same cell is the whole cell area. In order to determine quantitatively the cell area variation every reduced image is segmented and labelized, (last column of Fig. 2). Because of the reduced thickness of the medium, the swimming path of the cells is planar, i.e. the cells are always focused. The average area of the cells moving in the field is calculated and the detected cells are contoured. For each reduced image a pre-established standard deviation value determines the selection of an area value range. Therefore, touching cells are automatically rejected because their area is too big; similarly, small agglomerate of pixels, produced by the subtraction operation in the case of moving cells which intrudes onto the area formerly occupied by a different cell, or when a cell enters the field of view between the two frame on which the subtraction operation is performed, are rejected. The variation of cell number in the microscope field during the acquisition is not critical for the analysis, because we calculate the average area of the labeled cells present in each difference images. Twenty sequences are measured and the calculated areas averaged.

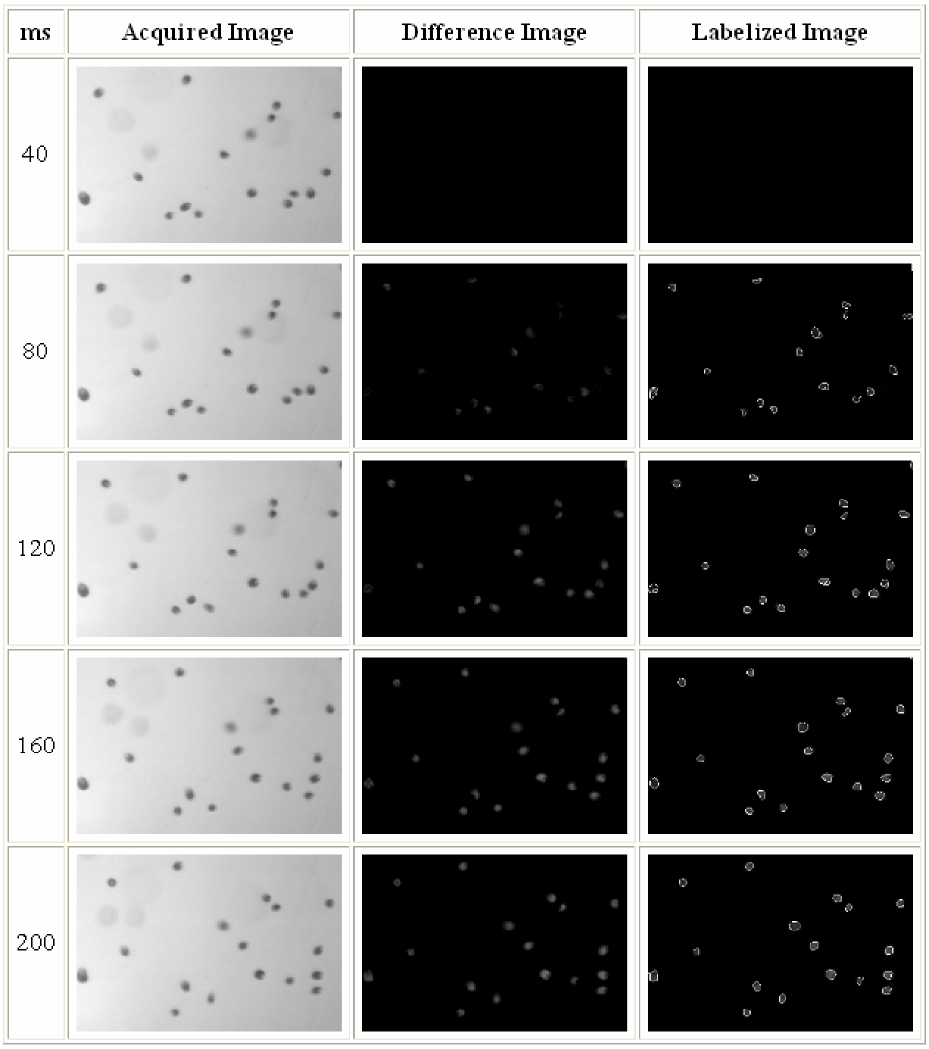


Fig. 2a. Procedure for translational speed determination (first five frames).

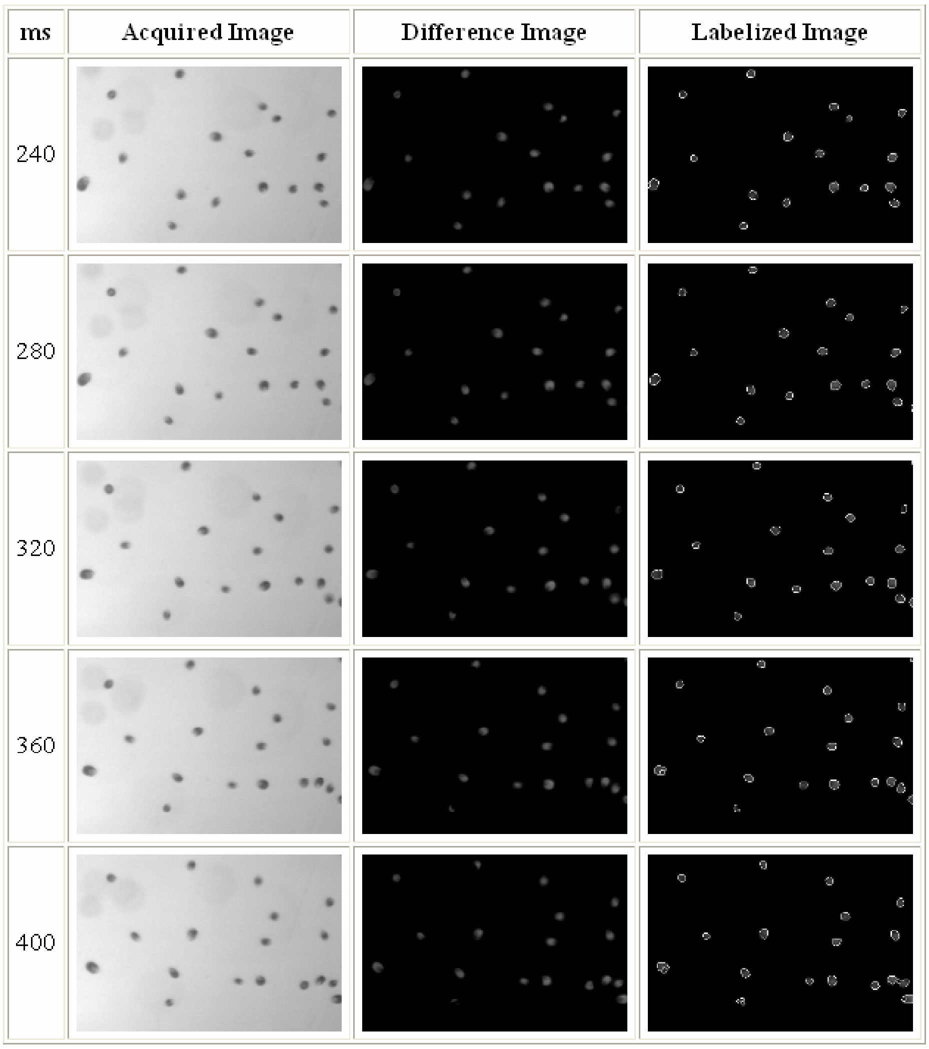


Fig. 2b. Procedure for translational speed determination (last five frames).

Fig. 3 shows the plot of the average cells area versus the delay progression. By interpolating the data of this plot, we obtain two intersecting straight lines. The first line shows that the average area value increases with the increasing of the delay because the overlapping of two successive images of the same cell decreases. The second lines shows that the average area value becomes steady because there is no more overlapping between the two successive images of the same cell. The intersection of these two lines identifies the time delay which has to be used for the determination of the exact swimming speed of the microorganisms. In our case about 150 msec is the time *Dunaliella* cells need to cover a distance equal to its long axis. A quantitative determination of the long axis of *Dunaliella*, by means of our labelization procedure, gives an average value of about15 μ. Previous studies reported a velocity of 100 μ/sec for *Dunaliella* cells, [12], therefore we can state that our system gives a correct evaluation of the swimming speed of this microorganism.

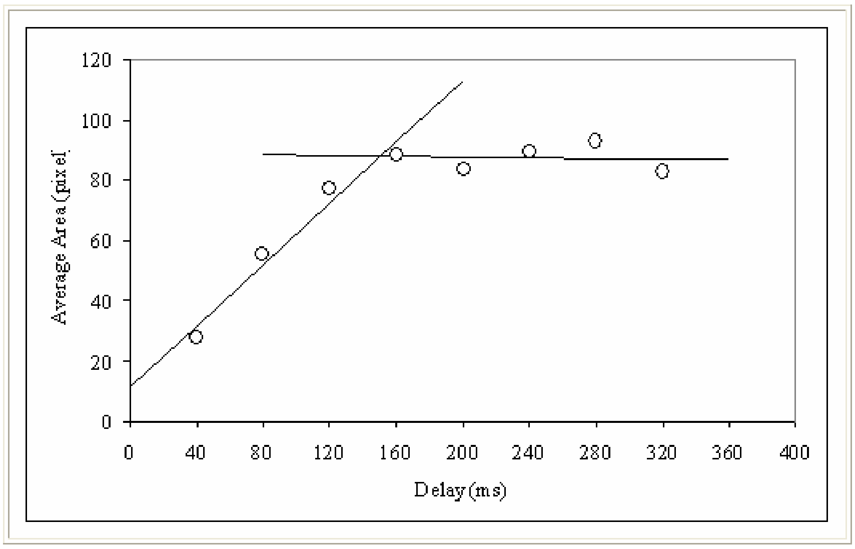


Fig. 3. The plot of the average area value of the cells vs. the delay between frames.

To investigate the rotational speed we store in the computer memory the frames acquired under lateral illumination as described in the Material and Methods section. The eyespot of Chlorophyta such as *Dunaliella* is a quarter-wavelength multi layered organization of osmophilic granules, which reflects very efficiently the light that strikes upon it. As the cell moves, we can detect this brilliant spot and verify if the cell rotates or not. For the wild type of *Dunaliella* frames were acquired every 40 ms, were thresholded and labelized so the eyespot is recognized as present in the image, (Fig. 4).

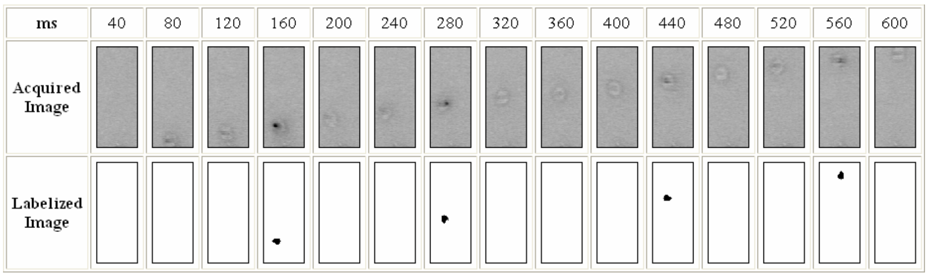


Fig. 4. Procedure for rotational speed determination.

The resulting duty cycle from a 600 ms recording shows that these cells rotate with a frequency of 8 Hz. (Fig. 5).

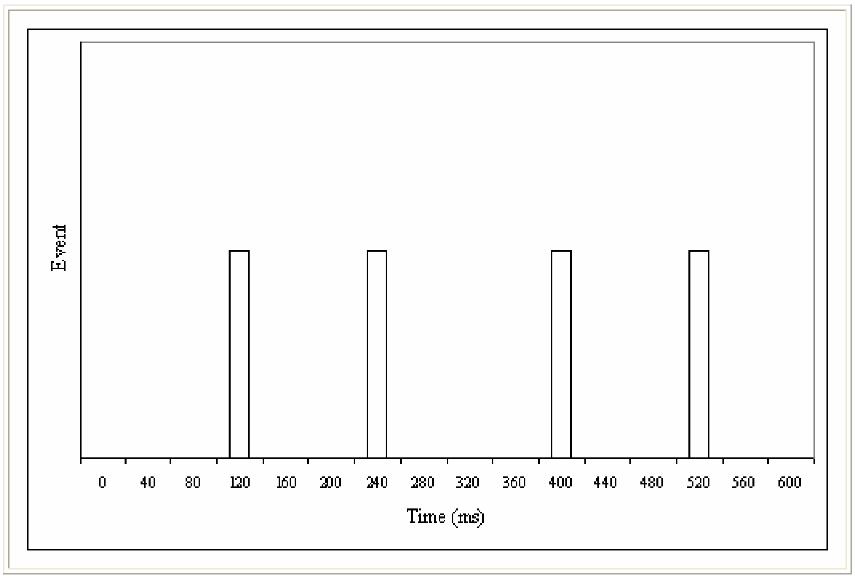


Fig. 5. The plot of the event of the eyespot detection vs. the elapsed time 600.

The time resolution of our system, which is 40 msec, can be considered sufficient to determine speed parameters of moving microorganism, as the study of physiological aspects of motility is usually based on the microscope observation of these phenomena. Due to its integration time, the human visual system has a time resolution of 250 msec, which is 6-time greater than that of our system [13]. Because the problem to be solved is the quantitative determination of visual phenomena, our system can be considered quite adequate for this purpose.

References

1. Gibbons, B.H., and Gibbons, I.R.: Flagellar movement and adenosin triphosphatase activity in sea urchin sperm extracted with triton X-100, J. Cell. Biol. 54 (1972) 75-82
2. Phillips, D.M.: Comparative analysis of mammalian sperm motility, J. Cell. Biol. 53 (1972) 561-565
3. Ojakian, G.K., and Katz, D.F.: A simple technique for the measurement of swimming speed of Chlamydomanas, Exptl. Cell. Res. 81 (1973) 487-491
4. Ascoli, C., Barbi, M., Frediani, C., and Mure, A.: Measurements of Euglena gracilis motion parameters by laser light scattering, Biophys. J. 24 (1978) 585-599
5. Biedert, S.W., Barry, W.H., and Smith, T.W.: Inotropic effects and changes in sodium and calcium contents associated with inhibition of monovalent cations active transport by ouabain in cultured myocardial cells, J. Gen. Physiol. 74 (1979) 479-494

6. Yachida, M., Asada, M., and Tsuji, S.: Automatic analysis of moving images, IEEE Patter. Anal. Match. Intel. 3 (1981) 12-20
7. Kondo, T., Kubota, M., Aono, Y., and Watanabe, M.: A computerized video system to automatically analyze movements of individual cells and its application to the study of circadian rhythms in phototaxis and motility in *Chamidomanas reinhardtii*, Protoplasma Suppl. 1 (1988) 185-192
8. Evangelista, V., Barsanti, L., Passarelli, V, Gualtieri, P. (eds.): From Cell to Proteins: Imaging Nature Across Dimensions, Springer, Dordlecht, (2005)
9. Allen, R.: New direction and refinements in video-enhanced microscopy applied to problems in cell motility, in Cowden, R.R., and Harrison, F.W. (eds), Advances in Microscopy, vol 139, Alan Riss New York (1983) 3-11
10. Vismara, R., Verni, F., Barsanti, L., Evangelista, V., Gualtieri, P.: A short flagella mutant of *Dunaliella salina* (Chlorophyta, Chlorophyceae), Micron 35 (2004) 337–344
11. Coltelli, P., and Gualtieri, P.: A procedure for the extraction of object features in microscope images, J. Biomed. Comput. 25 (1990) 169-176
12. Barsanti, L., Gualtieri, P.: Algae, Anatomy, Biochemistry, and Biotechnology. Taylor and Francis, Boca Raton (2006)
13. Rose, A.: Vision: Human and Electronic, Plenum Press, New York (1977)

Automatic Fuzzy-Neural based Segmentation of Microscopic Cell Images

Sara Colantonio[1], Igor B. Gurevich[2], Ovidio Salvetti[1]

[1] Istituto di Scienza e Tecnologie dell'Informazione, ISTI-CNR
Via G. Moruzzi 1,
56124 Pisa, Italy
{Sara.Colantonio,Ovidio.Salvetti}@isti.cnr.it
http://www.isti.cnr.it/ResearchUnits/Labs/si-lab/
[2] Dorodnicyn Computing Centre of the Russian Academy of Sciences
40, Vavilov str., Moscow GSP-1 119991 The Russian Federation
{igourevi}@ccas.ru

Abstract. In this paper, we propose a novel, completely automated method for the segmentation of lymphatic cell nuclei represented in microscopic specimen images. Actually, segmenting cell nuclei is the first, necessary step for developing an automated application for the early diagnostics of lymphatic system tumors. The proposed method follows a two-step approach to, firstly, find the nuclei and, then, to refine the segmentation by means of a neural model, able to localize the borders of each nucleus. Experimental results have shown the feasibility of the method.

1 Introduction

A great deal of research has concerned, in the last years, the development of automated systems for the early diagnosis of lymphatic tumors based on the morphological analysis of blood cells in microscopic specimen images. Actually, pathologists usually make diagnosis by analyzing the morphology of specimen cells [1, 2].

The first and necessary step for automating cell analysis is an accurate segmentation of the cells themselves, which is, then, followed by the extraction of significant morphological parameters. Unfortunately, cell segmentation is usually an ill-posed problem: due to poor dye quality, cell boundary could be not well distinguishable and parts of the same tissue could be not equally stained; two or more cells could be very close to each other or even overlapping; the chromatin distribution inside the cells could generate strong computed edges which mislead the segmentation.

In past years, many segmentation methods have been presented [3, 4]. They include watersheds [5, 6], region-based [7] and threshold-based methods [8]. The problem with these methods is that they do not employ any shape information of the cell, which can be useful in presence of noise.

Recently, the application of Active Contours has been widely investigated for cell segmentation [9, 10]. However, such methods require an initialization of the *snake*, making the segmentation not completely automated. Moreover, having to select which cell the snake should be apply to, much information regarding all the cells represented in the images is lost.

Other contour-based methods include Active Shape Models (ASM) [11], Active Appearance Models (AAM) [12] and variational deformable models (*Strings*) [13]. In the first two cases, a boundary model and its allowed variations are learned from a set of example boundaries and represented by a set of labeled points, encoding only shape information in ASM, also image features in AAM. The Strings method differs from the previous ones in adopting a continuous instead of discrete boundary representation, together with a multiple features description, giving place to a multivariate curve representation in functional space (instead of a point representation in vector space). All these methods require initialization and allow modeling only the variation seen in the training set of boundary examples.

The method we propose in this paper has the main characteristic to be completely automated. Moreover, it is suitable to segment all the cells contained in the images, allowing to extract information not only from the malignant ones.

Following a two-step approach, images are first clustered, in order to perform a rough segmentation and localize the cells. In a second processing step, an Artificial Neural Network (ANN) is applied to the image portions containing the localized cell for individuating cell borders.

Such an approach assures a high level of robustness, because the ANN performs a classification of the image and then it can distinguish among different kinds of structures, e.g. cell nucleus, cytoplasm, background, artifacts and so forth.

2 The Fuzzy-Neural Segmentation

Microscopic cell images are acquired as footprints of lymphoid tissue stained according to the Romanovsky-Giemsa technique and digitized as color images.

Each image *I* contains a number, say *n*, of cells which are constituted by the internal body – the nucleus –, which is the structure of interest to be segmented, and the cytoplasm. Due to the staining procedure, artifacts can be present in the images, as well as not perfectly stained cells that can be then considered as added *noise*.

The proposed method is suitable to detect nuclei borders and consists in applying to each image *I* a two-stage procedure as follows:

1. *Cell dislocation detection*: a cluster analysis, based on the *fuzzy c-means* algorithm, is applied to identify and label homogeneous regions in the image. The clustered regions are then used to divide the entire image in disjoint sub-parts for further processing (image partition).
2. *Cells contours extraction*: from each image partition relevant features are extracted and a dedicated ANN is used to complete the segmentation by identifying the contours of each cell.

A sketch of the method is shown in Fig. 1.

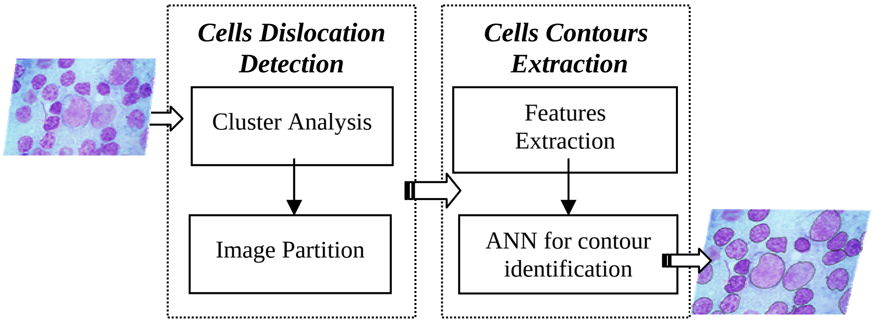

Cells Dislocation Detection
Cluster Analysis
Image Partition
Cells Contours Extraction
Features Extraction
ANN for contour identification

Fig. 1. The two-step method for cell segmentation

In the following, each step is described in more details.

2.1 Cells Dislocation Detection

In order to individuate how cells are dislocated in the microscopic images, a fuzzy cluster analysis is performed and each image is partitioned in disjoint parts for next step elaboration.

Cluster Analysis. Homogeneous image regions are labelled using an unsupervised clustering method, based on the *fuzzy c-means* algorithm (FCM) [14]. This algorithm groups a set of data in a predefined number of classes so as to iteratively minimize a criterion function, namely the sum-of-squared-distance from region centroids, weighted by a cluster membership function. A membership grade $p \in [0,1]$ is associated to each element of the data set, describing its probability to belong to a particular cluster.

For each cell image I, a features vector

$$(I_0(x), I_1(x), I_2(x), \dots, I_q(x))$$

is computed for any pixel x, considering $I(x)$ as a vector of the three color component $I(x)=(r,g,b)$. Then $I_0(x) = I$, and for $k = 1,\dots,q$, $I_k(x) = I * \Gamma_k(x)$, where Γ_k is a Gaussian filter with $\sigma = k$. In this way, we obtain a data set $D = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ where each $\mathbf{v}_h$, $h=1,\dots,m$ is a vector in $\Re^p$ representing image elements at different resolutions.

Let U_{cm} be a set of real $c \times m$ matrices, with c being an integer, $2 \le c < m$; the fuzzy c-partition space for D is, then, the set:

$$\Omega = \{U \in U_{cm} : u_{ih} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1, 0 < \sum_{h=1}^{m} u_{ik} < m\} . \qquad \textbf{(1)}$$

where u_{ih} is the membership value of $\mathbf{v}_h$ in cluster i ($i = 1,\dots,c$).

By applying FCM, an optimal fuzzy c-partition and corresponding prototypes are found minimizing the objective function:

$$J_\eta(U,\Lambda;D)=\sum_{h=1}^{m}\sum_{i=1}^{c}(u_{ih})^\eta\left\|\mathbf{v}_h-\boldsymbol{\lambda}_i\right\|^2 . \qquad \textbf{(2)}$$

where $\Lambda = (\lambda_1, \lambda_2,\ldots, \lambda_c)$ is a matrix of unknown cluster centers $\lambda_i \in \Re^p$, $\|\cdot\|$ is any norm, e.g. the Euclidean norm, expressing the similarity between each data vector $\mathbf{v}_h$ and the center λ_i, and the weighting exponent $\eta \in [0,\infty)$ is a constant that influences the membership values.

Fuzzy partition is carried out through an iterative minimization of (2), calculating the cluster centers at each iteration t = 1,2,… as:

$$\boldsymbol{\lambda}_i^{(t)}=\frac{\sum_{h=1}^{m}(u_{ik}^{(t)})^\eta\mathbf{v}_h}{\sum_{h=1}^{m}(u_{ik}^{(t)})^\eta} . \qquad \textbf{(3)}$$

and updating the membership values as:

$$u_{ik}^{(t)}=\left[\sum_{j=1}^{c}\left(\frac{\left\|\mathbf{v}_h-\boldsymbol{\lambda}_i^{(t)}\right\|^2}{\left\|\mathbf{v}_h-\boldsymbol{\lambda}_j^{(t)}\right\|^2}\right)^{\frac{2}{\eta-1}}\right]^{-1} . \qquad \textbf{(4)}$$

The iterative process stops when $|U^{(t+1)}-U^{(t)}|$ follows under a certain threshold or the maximum number of iterations is reached.

Applying the FCM on the cell images induces a partition of each *slide* into a set $P=\{R_1,R_2,\ldots\}$ of disjoint connected regions R, where the indices 1,2,… are region labels. In other words, by clustering, we obtain a rough segmentation which can be refined reducing the computation by the following step of image partitioning.

Image Partitioning. Once clustered the image, the convex hull of each connected region is calculated in order to delimitate the largest image portion (*convex image*) containing the corresponding connected region.

Starting from the convex hull, an image partition is extracted slightly enlarged in both directions the *convex image*. Such partition contains what the FCM has classified as a unique cell. However, the contour of the clustered region can be inaccurate, including, for instance, the cytoplasm; moreover, it can happen that two very closed or touching cells are clustered as a unique region. For these reasons, it is necessary to refine the clusterization in a further step.

2.2 Cells Contour Extraction

In order to detect the exact cell contour, from each image partition, a set of features is extracted and classified by a dedicated ANN.

Features Extraction. Analyzing the properties of cell images and of the similar cells, the following vector of features $\Im(x)$ is computed for characterizing each pixel x of the segmented image partition:

- *Color values*: $I(x) = (r,g,b)$;
- *Mean color value*: $M(x) = (M_r, M_g, M_b)$ computed applying an average filter $F(x)$, i.e. $M(x) = I(x) * F(x)$;
- *Gradient norm*: $\|\nabla I(x)\|$ and its mean, computed along the three color components;
- *Radial gradient*: $G_{rt}(x)$, defined as the gradient component in the radial direction $\hat{r}$ from the center of the connected region;
- *Membership value to the clustered region*: $u_i(x)$, where i is the cluster index considered as a cell in the image partition.

ANN for contours identification. The vectors of the extracted features $\Im(x)$ are processed by a dedicated ANN. It consists in a *Multilayer Perceptron*, trained according to the *Error Back-Propagation* (EBP) algorithm [15] to recognize five different classes. At present, to resolve ambiguity in case of touching cells and let the network learn and generalize better, five pixel classes are selected:

1. *Cell border*
2. *Cell internal body*
3. *Cytoplasm*
4. *Background*
5. *Artifact*

Let $o_j(\Im(x))$ be the answer of the output units of the network when the features vector $\Im(x)$ is being processed; then, the pixel membership to one of the above mentioned classes can be computed as

$$\Phi(x) = \text{argmax}_{j=1,\ldots,5}(o_j(\Im(x))) \ . \qquad \textbf{(5)}$$

A set of pre-classified images has been used to train the network, using the *Resilient Back-Propagation* [16] version of the EBP algorithm. Once defined the desired ψ_p output for each input vector of the training set TS = $\{\Im_p(x)\}$, the cost function

$$E = \frac{1}{2}\sum_{p=1}^{|TS|}(\psi_p - \mathbf{o}_p)^2 \ . \qquad \textbf{(6)}$$

where $\mathbf{o}_p = (o_1, o_2, \ldots, o_j)$ is the output vector of the network, is minimized iteratively computing the weight update at each iteration step t as follows:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)} & \text{if } \dfrac{\partial E}{\partial w_{ij}}(t) > 0 \\ -\Delta_{ij}^{(t)} & \text{if } \dfrac{\partial E}{\partial w_{ij}}(t) > 0 \ . \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{(7)}$$

where w_{ij} is the weight between the network units i and j, and Δ_{ij} is the amount of weight change which, starting from a chosen value Δ_0, varies at each step t according to the following equation:

$$\Delta_{ij}^{(t)} = \begin{cases} \varepsilon^{+}\Delta_{ij}^{(t-1)} & \text{if } \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{\mathrm{ij}}}(t-1)\cdot\frac{\partial \mathrm{E}}{\partial \mathrm{w}_{\mathrm{ij}}}(t) > 0 \\ \varepsilon^{-}\Delta_{ij}^{(t-1)} & \text{if } \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{\mathrm{ij}}}(t-1)\cdot\frac{\partial \mathrm{E}}{\partial \mathrm{w}_{\mathrm{ij}}}(t) < 0 \\ \Delta_{ij}^{(t-1)} & \text{otherwise} \end{cases} . \qquad (7)$$

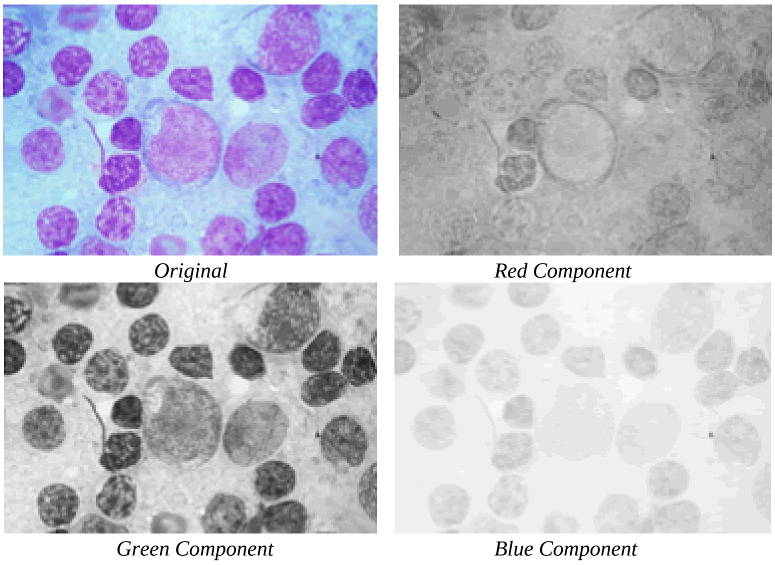
where $0 < \varepsilon^{-} < 1 < \varepsilon^{+}$ are parameters used to regulate weight modifications.

The final result of this step is discussed in the following section.

3 Results

Footprints of lymphoid tissues were Romanovsky-Giemsa stained and digitized with digital camera mounted on Leica DMRB microscope using PlanApo 100/1.3 objective. The equivalent size of a pixel was 0,0036 μ^2; 24-bit color images were stored in TIFF format of dimensions 1200 × 1792. A total number of 800 microscopic images were considered, with an average number of 20 cells for each. An example of a microscopic cell image and its three color components is reported in Fig. 2.

The cluster analysis was designed to be performed on the features vectors ($I_0(x)$, $I_1(x)$, $I_2(x)$,…, $I_q(x)$) with q = 5, but, among such components, only $I_3(x)$ and $I_5(x)$ were considered relevant. The input vectors represented in the form of color images are shown in Fig. 3.



Original *Red Component*

Green Component *Blue Component*

Fig. 2. An example of microscopic cell image: the original image and the three color components.

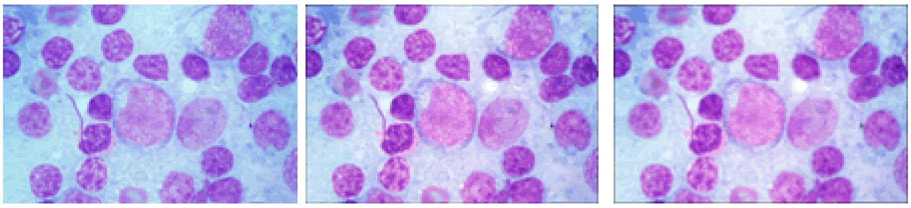
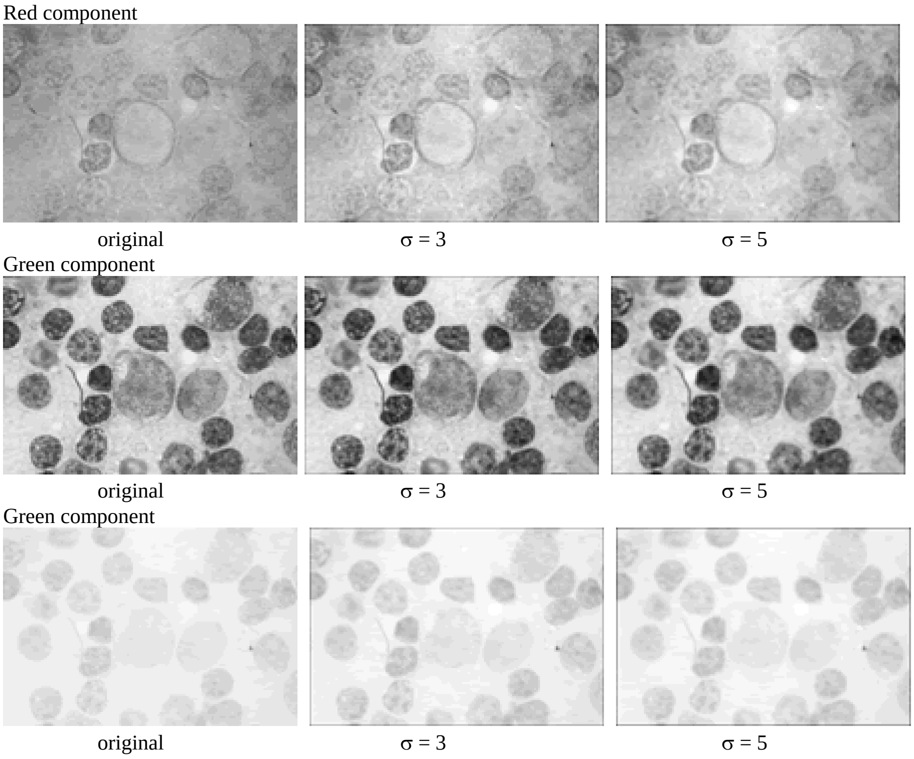


Fig. 3. An example of the three-component feature vector used for clustering: from left to right, original, σ = 3 and σ = 5.

The same feature vector for each of the color components of the image is reported in Fig. 4.

The FCM algorithm is applied to divide image pixels into two clusters corresponding to cell and background. A filling operation is performed to eliminate little holes, while clustered regions of negligible area are deleted. An example of the clustering results is reported in Fig. 5.



Red component

original σ = 3 σ = 5

Green component

original σ = 3 σ = 5

Green component

original σ = 3 σ = 5

Fig. 4. An example of the feature vector with the original values $I_0(x)$ and $I_3(x)$ and $I_5(x)$ for each of the three color components.

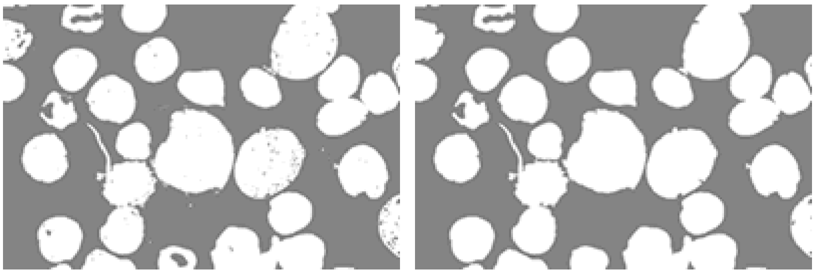


Fig. 5. Example of the clustering results: rough clustered image (left), clustered image after a filling operation and after deletion of regions of negligible area (right).

Examples of image partitions extracted for detecting the exact borders of a cell are shown in Fig. 6.

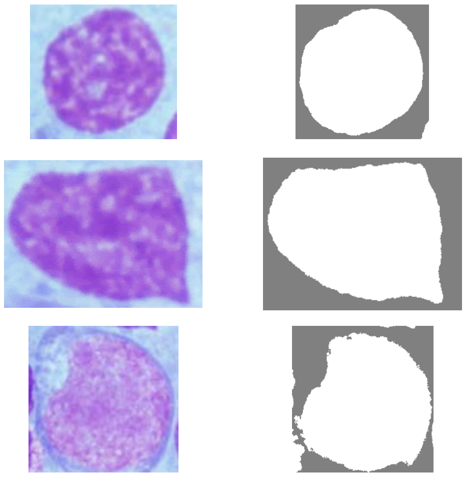


Fig. 6. Image partitions containing the cells to be segmented.

From each partition, the set of the mentioned features is extracted. To illustrate the significance of such set, Fig. 7 shows an example of the gradient regarding the green component.

The set of 800 images was partitioned in (i) a sub-set of 300 images, used for training, and (ii) a sub-set of the remaining 500 images used for the testing phase. A semi-automatic segmentation was performed for the training set, consisting in a classification of images according to the different classes of pixels.

Different architectures were tested, varying the number of the hidden units: the best performance was achieved with only one hidden layer of 20 units. An example of the segmentation results is illustrated in Fig. 8, where the entire classification results are reported too.

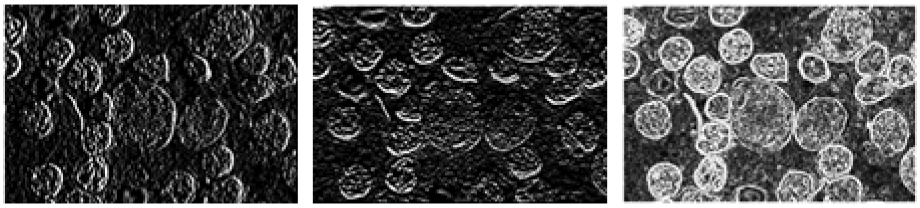


Fig. 7. Example of the computation of the green component gradient along the horizontal axis (*left*), along the vertical axis (*middle*) and the norm of the same gradient (*right*).

4 Discussion and Conclusions

A two-step method for segmenting microscopic cell images has been presented.

The first step consists of a fuzzy clustering of images performed to obtain a rough segmentation and to detect cell dislocation. In the second step, a dedicated ANN is applied to refine the segmentation by discriminating image components, i.e. cell borders, cell internal body, cytoplasm, background, and artifacts.

The main features of the proposed method are

- complete automation of segmentation
- possibility of extracting all the cells represented in the images
- robustness due to the ANN application which allows resolving ambiguity of closed or touching cells.

An example of the last characteristic is shown in Fig. 9, where it can be seen how two cells that are clustered as a unique region by the FCM are well separated by the ANN thanks to the individuation of cytoplasm.

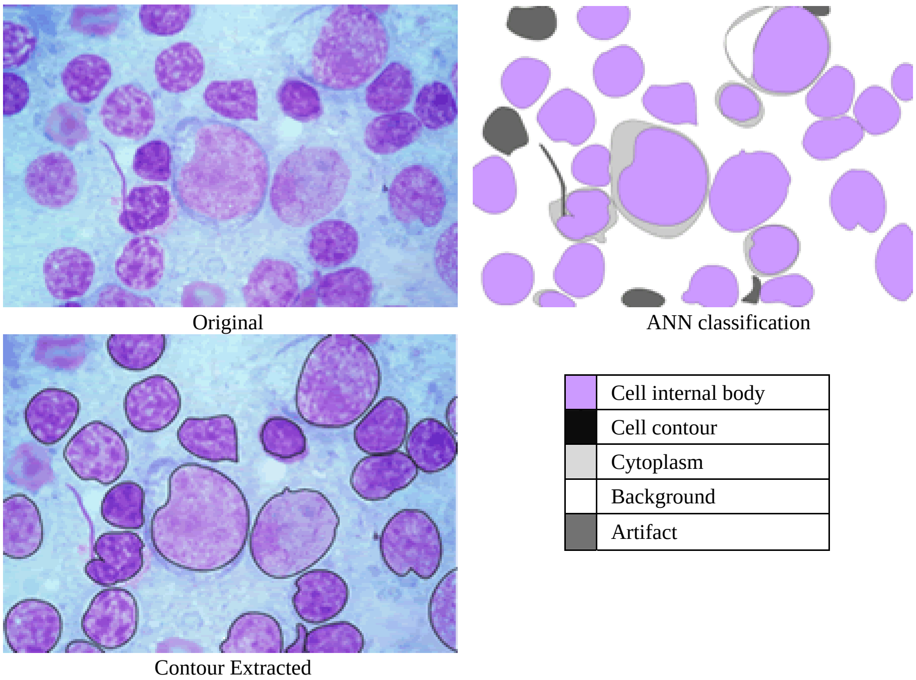


Fig. 8. Example of segmentation. *upper left*: original cell image; *upper right*: results of the ANN classification (five classes with different colors); *lower left*: identified contours of each cell; *lower right*, legenda.

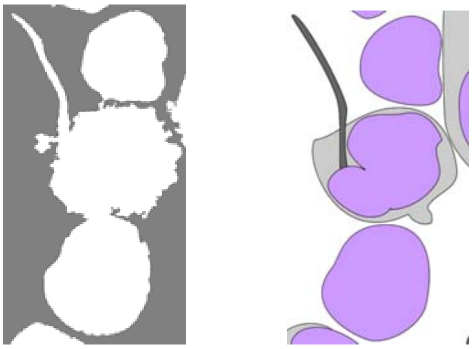


Fig. 9. Example showing the robustness of the proposed method: (*left*) rough segmentation obtained by FCM that individuates a unique region corresponding to three different cells; (*right*) result of the ANN algorithm where the cells are correctly separated by classifying pixels in cell body, cytoplasm and artifact (see Fig. 8 for explanation of colors).

Acknowledgment

This work was partially supported by INTAS Grant N 04-77-7067, by the Cooperative grant “Image Analysis and Synthesis: Theoretical Foundations and Prototypical Applications in Medial Imaging” within bilateral agreement between Italian National Research Council and Russian Academy of Sciences, by European Project Network of Excellence MUSCLE – FP6-507752 (Multimedia Understanding through Semantics, Computation and Learning), and by the Russian Foundation for Basic Research Grant N 05-07-08000.

References

1. Zh.V. Churakova, I.B. Gurevich, I.A. Jernova, et al. Selection of Diagnostically Valuable Features for Morphological Analysis of Blood Cells, Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Vol. 13. (2003) 2 381-383
2. Jaffe E. S , Harris N. L., Stein H., et al. Pathology and Genetics of Tumors of Haematopoietic and Lymphoid Tissues. Lyon: IARC Press (2001)
3. Di Rubeto, C., Dempster, A., Khan S., Jarra, B.: Segmentation of Blood Image using Morphological Operators. Proc. 15th Int. Conference on Pattern Recognition (2000), 3 397-400
4. Anoraganingrum, D.: Cell Segmentation with Median Filter and Mathematical Morphology Operation. Proceeding International Conference on Image Analysis and Processing (1999), 1043-1046
5. Lin G., Adiga U., Olson K., Guzowski J.F., Barnes CA, Roysam B.: A hybrid 3D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. Cytometry A. (2003), 56, 1, 23-36

6. Umesh Adiga P.S. and Chaudhuri B.B., “An efficient method based on watershed and rulebased merging for segmentation of 3-D histopathological images. Pattern Recognition (2001), 34, 7, 1449-1458
7. Mouroutis, T., Roberts, SJ., Bharath. AA.: Robust cell nuclei segmentation using statistical modelling. BioImaging (1998), 6, 79-91
8. Wu, HS., Barba, J., Gil, J.: Iterative thresholding for segmentation of cells from noisy images. J. Microsc. (2000), 197,296-304
9. Karlosson, A., Strahlen, K., Heyden, A.: Segmentation of histological section using snakes. In J. Begun and T. Gustavsson, eds. LNCS 2749, Proc. of 13 Scandinavian Conference, SCIA 2003, Halmstad, Sweeden (2003), 595-602
10. Murashov, D.: Two-level method for segmentation of cytological images using active contour model. Proc. 7th Int. Conference on Pattern Recognition and Image Analysis, PRIA -7 (2004), III, 814-817
11. Cootes T.F. and Taylor C.J.: Active Shape Models – ‘Smart Snakes’, Proc. British Mach. Vision Conf., Springer-Verlag (1992), 266-275
12. Cootes T. F., Beeston C., Edwards G. J., Taylor C. J.: A unified framework for atlas matching using active appearance models, in Information Processing in Medical Imaging, A. Kuba and M. Samal, Eds. Berlin, Germany: Springer-Verlag, Lecture Notes in Computer Science (1999), 322–333.
13. Ghebreab S. and Smeulders A.W.M.: Strings: Variational Deformable Models of Multivariate Continuous Boundary Features, IEEE Transactions on Pattern Analysis And Machine Intelligence (2003), 25, 11, 1399-1410
14. Bezdek, L.C.: Pattern Recognition with Fuzzy Objective Function Algorithm, New York: Plenum Press, 1981
15. Rumelhart, DE., Hinton, GE., Williams, RJ.: Learning internal representations by error propagation” Parallel Distribuited Processing (1986) MIT Press, Cambridge, MA, 318-362.
16. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The RPROP algorithm. Proc. of the IEEE International Conference on Neural Networks – ICNN, (1993), 586-591

Semi-Automated Mapping of Cell Nuclei in 3D-Stacks from Optical-Sectioning Microscopy

Martin Heß

Biozentrum der LMU München, Großhadernerstr. 2, 82152 Planegg, Germany
hess@zi.biologie.uni-muenchen.de

Abstract. 3D-stacks of optical sections through the vertebrate retina with fluorescent stained cell nuclei were measured with a laser scanning microscope. The evaluation of the data volumes with dedicated digital imaging algorithms gives access to complex morphometric tissue-characters that are discussed in terms of functional morphology. The thickness of nuclear layers and the 3D-coordinates of cell nuclei are detected automatically to measure cell densities, cell ratios and to create character-distribution-maps of the entire retina.

1 Introduction

Confocal laser scanning microscopy combined with any fluorescence staining technique is a powerful and elegant method to get three-dimensional structural data from biological tissues. Usually the result of a single xyz-scan-measurement is a stack of evenly spaced and perfectly aligned greyscale images (“optical sections”) with a considerable data volume and information content. Frequently these stacks are used to generate attractive displays of the stained structures (e.g. brightest point projections, colour channel overlays, surface renderings), but rarely for thorough evaluation of the stack’s information content by means of three-dimensional morphometric analysis. Many complex tissue characters are hardly revealed to an intuitive understanding by mere visual contemplation of 3D-data stacks or evade from a precise manual evaluation in an acceptable period of time. Digital imaging algorithms, however, allow the extraction of both simple and complex characters from huge data stacks in a time-saving way even on a standard PC. Still usually they have to be programmed and tailored to the specific object and scientific question by the scientist himself. Since biologists normally recoil from this challenge, most morphometric studies usually do not reach the third dimension to this day. In this study an example is presented for a computer-aided investigation and description of three-dimensional patterns of cell nuclei in the vertebrate retina. On the one hand retinal tissue is particularly suitable for optical-sectioning microscopy due to its transparency and low thickness, on the other hand the layered structure and high degree of geometrical order of this brain-derivative carries valuable information for the functional morphologist. For this study the retina of the European anchovy *Engraulis encrasicolus* (Teleostei, Engraulididae) was chosen to make a contribution to the morphometric description of the vertebrate retina in general and to approach to a more profound understanding of an uncommon retina in special, that is specialized for polarization contrast vision [1,2].

2 Material and Methods

2.1 Tissue preparations

Adult European anchovies (*Engraulis encrasicolus*) where obtained from local fisherman just returning from their nocturnal catches in the Mediterranean (Adriatic sea, Rovinj). For the time of death dated back less than 1.5 hours, the retinal tissue of cooled animals could be regarded as *in-vivo*. Eyes where enucleated, the eyeballs perforated by razorblade-cuts through the cornea and fixed with 4% formaldehyde in 0.1M phosphate buffer at pH 7.4 plus 3% sucrose for several hours. The cornea, lens and vitreous body were removed in cold buffer, thereafter the entire retina of a right eye (diameter 8mm) was cut into 48 pieces whose original positions were documented. The retinal fragments were rinsed in buffer and embedded in 4% agarose at 45°C in separate dishes of two 24-well culture plates. From the centre of each fragment radial slices (thickness 50 µm) were made with a Leica VT1000S vibratome for subsequent radial optical sectioning. The slices were submersed in a 1µM-solution of TO-PRO-3 (Invitrogene, λ_{max}(Excitation) = 642nm, λ_{max}(Emission) = 660nm) in buffer for 10 to 60 minutes at ambient temperature for fluorescent staining of the cell nuclei. After the staining each slice was placed in a drop of anti-fading mounting medium (Vectashield®) between a glass slide and a cover glass and sealed with nail varnish. To avoid deformation of the slices by squeezing, the cover glass was braced by two pieces of cover glass as spacers (thickness 150µm) directly glued to the slide. A second preparation was accomplished to obtain 24 retinal fragments directly placed between 300µm spacers, pigment epithelium oriented downward for sub-sequent tangential optical sectioning.

2.2 Microscopy

The tissue preparations were imaged with a confocal laser scanning microscope (Leica TSC SP2 on an inverse Leica DM IRBE). For excitation of TO-PRO-3 the 633nm HeNe-line was used and attenuated to 10% to restrict bleaching. The beamsplitter was a triple dichroic (488, 568, 633nm) by default, the spectral detection window of the photomultiplier was set to 650-740nm. For the radial optical sections a Leica UV 25x PL Fluotar NA 0.75 oil objective was used (working distance 180µm, nominal resolution xy: 260nm, z: 1108nm), the voxel size was adjusted to 405nm in xy-plane by 810nm in z-direction (voxel-geometry: integral multiple of a cube). This allows both the display of radial slices through the thickest part of the retina in the “visual field” of the photomultiplier (207.4µm x 207.4µm allocated to 512 x 512 pixels) and a comfortable digital slice spacing without interpolation. Gain and offset of the photo-multiplier were optimized to exploit the 8bit-dynamic of the sensor with reference to the available signal. For the tangential optical sections a Leica UV 63x HCX PL Apo NA 1.32 oil objective was used, voxel size adjusted to 310nm in xy-Plane (158.7µm x 158.7µm allocated to 512 x 512 pixels) by 936nm in z-direction. The xyz-scans started near the cover glass towards the glass slide (against gravity), in each plane four optical sections were averaged to improve signal-to-noise ratio.

Depending on the retina thickness and the tilt angle of nuclear layers in the tissue slices the number of optical slices was varied between 27 and 118. The resulting stacks of greyscale images had a data volume between 7.1 and 30.4 MB, altogether 0.9 TB of raw data were generated.

2.3 Digital Image Analysis

For further processing the image data stacks generated by the acquisition software of the confocal laser scanning microscope were imported in IDL (interactive data language, Research Systems Inc.) on a standard PC (2.7GHz, 1MB RAM) and subjected to several home-made IDL-algorithms. The line of actions - i.e. pre-processing, semi-automated detection of cell nuclei, mapping of measurements etc. - is subject of the results chapter.

3 Results

3.1 Data import

3D-measurements at the CLSM usually deliver sequences of tiff-images as export-files. Every greyscale image can be regarded as a table of measurements sorted by columns and rows with entries between zero and 255 (8 bit). To get access to the entire data set of a 3D-measurement the respective image sequence was imported (via IDL software) into a single array-variable with three dimensions according to the x-, y- and z-axis of the measured tissue volume. The x- and y-index of the array mirrors the pixel-position in the original 2D-image, the z-index stands for the image number or its z-Position of the volume respectively. This allows to directly interrogate the measured value of the fluorescence signal of any point in the volume (voxel) specified by three index values. To get the correct proportions, every xy-plane was doubled (radial mechanical slices; z-spacing of optical sections 2x the pixel size) or trebled (tangentially oriented retina fragments; z-spacing 3x the pixel size). The last step can be omitted to save memory and to speed up calculations - for a correct display in perspectives (at any angle of view deviating from the z-axis) and for spatial measurements, however, the elongate voxel-shape has to be taken into consideration.

3.2 Display of raw data

In almost every case displays of the raw data show the nuclear layers of the retina oriented obliquely in the kartesian coordinate-system (Fig. 1). It is true that the vibratome sections were cut as close to the radial plane as possible under visual control, but a precisely radial orientation is not obtainable in practice, not least because of the hollow-sphere shape of the whole retina. Likewise in tangential view (whole mounts) the retina fragments always showed orientations tilted against the xy-plane. Nevertheless, as a simplification, a small retina fragment with 200 µm edges cut out of an eye with 8 mm diameter is regarded as not-curved in this study.

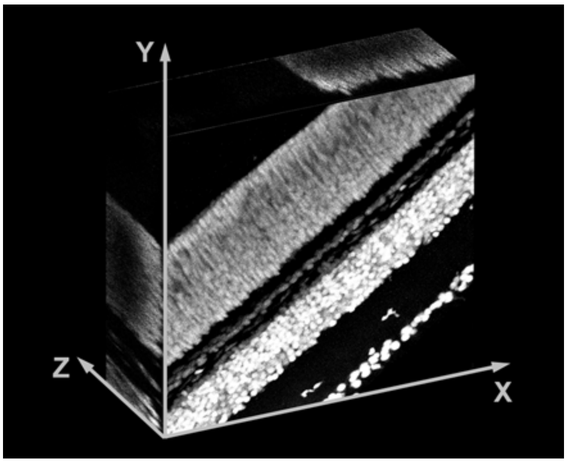


Fig. 1. Cubic display of a 3D-stack of optical slices through a small fragment of the anchovy-retina (radial vibratome slice) with coordinate system (*X, Y, Z*). The XY-view (207µm x 207µm) is a brightest-point-projection of the entire stack. The XZ- an YZ-views are single planes. *Note* nuclear layers lying obliquely in the data volume

3.3 Cell layer alignement

To simplify the following measuring methods and to get the common depiction of the retina with horizontally aligned histological layers, the fluorescent stained nuclear horizons of the scanned retinal volumes are to be oriented as parallel to the xz-planes (radial sections) or the xy-planes (whole mounts) of the data volume as possible. This happened in two orthogonal directions either interactively (with auxiliary lines parallel to the x-, y- or z-axis) or automatically. An appropriate digital imaging algorithm is demonstrated on a radial 3D-scan exemplarily (Fig. 2, 3): a brightest point projection (BPP) of the data volume in the xy-plane (i.e. z-axis shortened to zero) serves to determine the inclination angle of the nuclear layers to the x-axis. To do this, the BPP is rotated around the z-axis in 1° increments and convolved with a bright horizontal bar shifted vertically over the image in every angle-position. The result of this double-loop operation is a 2D-data set with a maximum indicating the rotation-angle that leads to horizontal alignement of the nuclear layers (Fig. 2). After rotation of the raw data stack around the z-axis by the determined angle (extension of the data volume on all sides helps to avoid clipping artefacts but increases memory demand and calculation time) the procedure is repeated with the yz-BPP and subsequent rotation around the x-axis. As a rule a second iteration of these two steps leads to a very good alignement of the nuclear layers for radial scans parallel to the xz-planes of the kartesian coordinate system (Fig. 3).

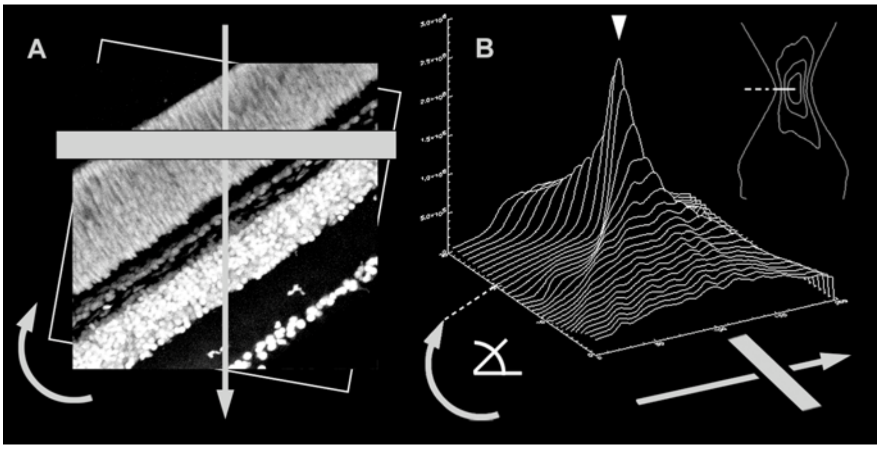


Fig. 2. Automated detection of rotation angle for the horizontal alignement of the nuclear layers. A) The XY-view of the data volume is convolved with a *horizontal bar* in vertical direction and then incrementally rotated. B) The resulting profiles build up a *2.5D-landscape* with a peak (*arrowhead*) that indicates the wanted rotation angle. *Inlay:* contour plot of the "mountain" with "summit"-position

3.4 Simple measurements

Based on BPPs of the aligned data volume the thickness of retinal layers, e.g. outer nuclear layer (ONL), inner nuclear layer (INL), inner plexiform layer (IPL) and ganglion cell layer (GCL), can be determined easily manually or automatically (Fig. 3). For the semi-automated morphometric analysis of the nuclear layers the definition of "regions of interest" (ROIs) containing unclipped fluorescence signals is required. The ROIs are defined on BPPs of the three orthogonal main-planes (XY, YZ, XZ), this way enclosing a "volume of interest" (VOI) completely filled with 3D-images of cell nuclei.

3.5 Detecting nuclear positions

To get the number and reliable centre-of-gravity positions of the cell nuclei quickly, every manually defined VOI was convolved with an idealized image (kernel) of the wanted structure (cell nucleus). The cell nuclei of the anchovy retina have diameters of 5-6µm depending on the cell type, resulting in circular profiles of 12 to 15 pixels maximally using the image acquisition settings indicated above for radial optical sections. Due to the almost spherical shape of cell nuclei in the retina the convolution can be executed with 2D-kernels plane by plane instead of 3D-kernels (spheres) in space. To do this, a kernel-array (2D-variable equivalent to the image of a white circle

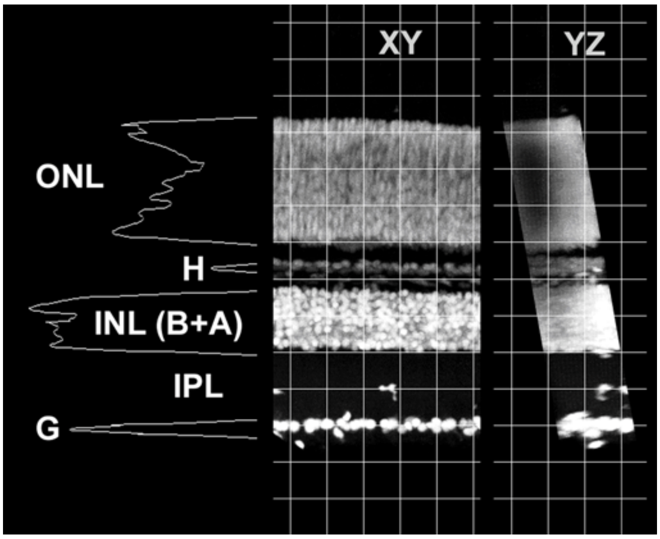


Fig. 3. Nuclear layers of the retina aligned parallel to the XZ-planes. Vertical intensity-profiles (*left*) help to measure the thickness of the outer nuclear layer (*ONL*), vitreal part of the inner nuclear layer (*INL* with bipolar (*B*) and amacrine (*A*) cells) and inner plexiform layer (*IPL*). Horizontal cells (*H*) and ganglion cells (*G*) form separate layers. *Note* restricted infiltration depth in the *ONL* (YZ-view)

with the approx. nuclear diameter on a black background) is centred over each XY-pixel of the data volume subsequently. The overlapping pixel-values of the kernel and the image are multiplied and the result is stored at the corresponding centre-position in a new 3D-variable. The result of this convolution procedure is a data set containing “blurred light-clouds” with local maxima at the centre-positions of the wanted nuclei (Fig. 4). Starting with the brightest maximum of the entire volume, the 3D-coordinates of the corresponding nucleus was written into a table. Then the nucleus around the local maximum was deleted in the VOI by multiplication with a black sphere of the approx. nuclear diameter and the convolution was repeated with the modified VOI and so on. Stop-criterion for this procedure was an estimated and pre-defined number of iterations combined with a test for erroneous measurements. Starting from the first coordinate the distances to all other detected points in the VOI were calculated. If a value lower than twice the expected nuclear radius occurred, the relevant coordinate was deleted. Such “misdetections” accumulated at iteration numbers equal or larger than the actual number of cell nuclei in the VOI.

3.6 Evaluation of position data

The corrected list of centre-coordinates allowed computation of the cell density of the VOI (converted to cells per $10^4 \mu m^2$ of retinal area), compilation of a neighbour-distances histogram for pattern-description (Fig. 5) and finally the correlation of measurements between the three neuron layers of a single retina fragment. After having analyzed several tissue fragments scattered over an entire retina 2D- or 3D-mapping of simple or complex measured characters can be demonstrated (e.g. density map of one cell type, ratio map of two parameters, Fig. 5). Every calculation step described above was performed on a standard PC between a few seconds and several minutes.

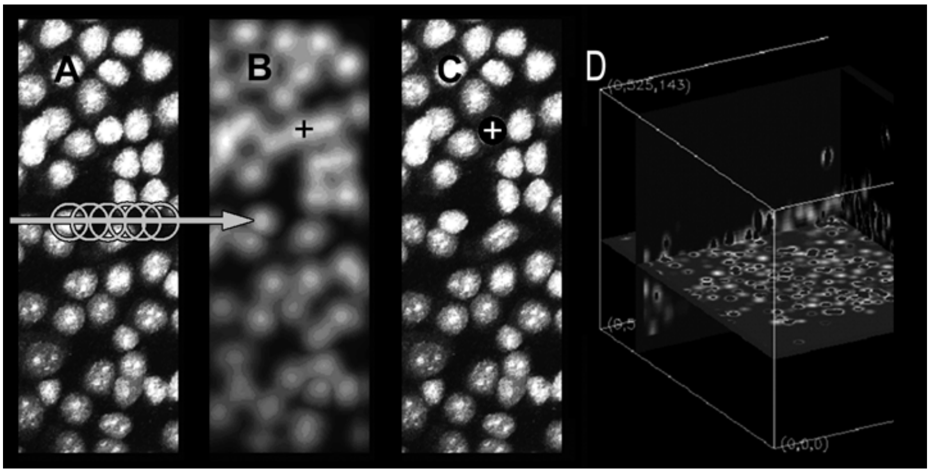
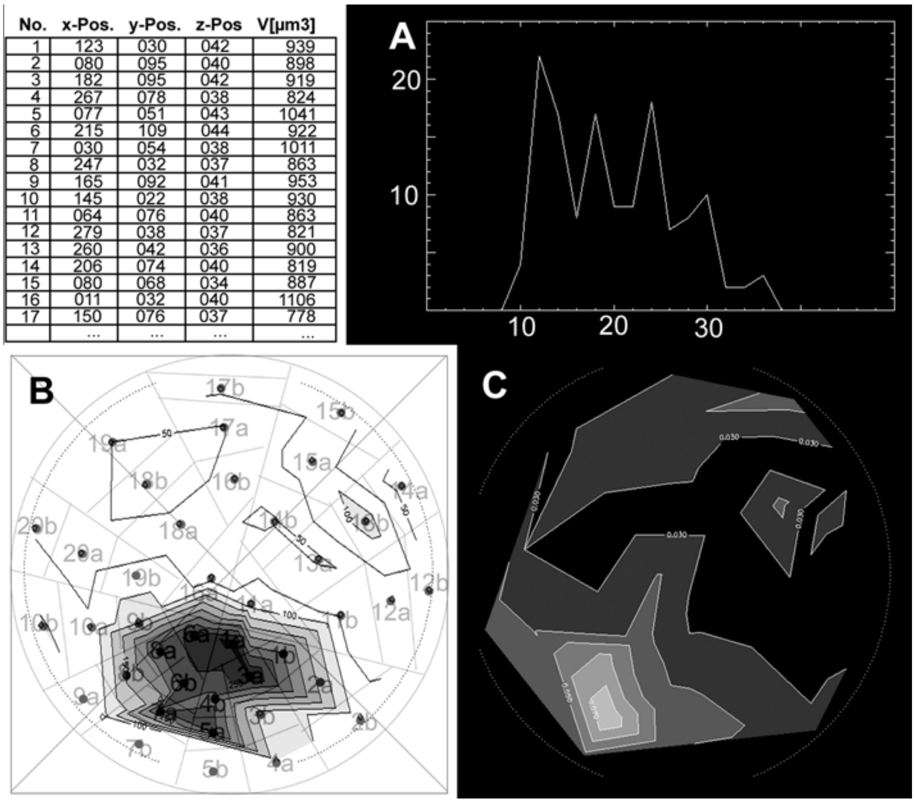


Fig. 4. Convolution of measured signals from the ganglion cell layer (*A*) with a discoidal kernel leads to a blurred picture with local maxima at the centre-positions of the cell nuclei (+ in *B*, 2D aspect of a 3D-operation). The detected nucleus is deleted (+ in *C*) prior to iteration of the convolution. *D*) Data stack with deleted nuclear centres

4 Discussion

This study outlines a method that gives access to complex morphological tissue-characters arising from the spatial arrangement of cell nuclei (in the vertebrate retina as example) by the use of fluorescence staining, optical sectioning microscopy and digital image analysis algorithms tailored for special purposes. Additionally it is intended to demonstrate the usefulness of mass-data analysis in the field of histology and functional morphology and to encourage the ambitious life scientist to design his own application software. The study provides the following lessons and impulses:



No.	x-Pos.	y-Pos.	z-Pos	V[µm3]
1	123	030	042	939
2	080	095	040	898
3	182	095	042	919
4	267	078	038	824
5	077	051	043	1041
6	215	109	044	922
7	030	054	038	1011
8	247	032	037	863
9	165	092	041	953
10	145	022	038	930
11	064	076	040	863
12	279	038	037	821
13	260	042	036	900
14	206	074	040	819
15	080	068	034	887
16	011	032	040	1106
17	150	076	037	778
	...	...	...	...

Fig. 5. Evaluation of position data. *Top left*: Table of nuclei detected in the INL with 3D-coordinates and measured nuclear volume. *Top right*: Histogram of neighbour-distances in a small VOI of H-cells with peaks indicating a square pattern (x-axis in µm). *Bottom left*: density distribution of G-cells with a ventral maximum around 250 cells/10^{4}µm^2. *Bottom right*: Ratio-map of INL-thickness / ONL-thickness indicating an area of high computing potential (light-grey) in the ventral retina

4.1 Optical sectioning depth

Prior to any programming the image acquisition parameters have to be adapted or rather optimized to get data sets that are suitable for evaluation. To be able to excite and collect fluorescence light from a sufficient tissue volume a penetration depth of at least 50 µm is desirable. Despite a relative high optical transparency, formalin-fixed retina tissue considerably scatters the visible light inversely proportional to its wavelength. To get deep optical sections with a satisfactory signal-to-noise-ratio a fluorescent stain with excitation- and emission maxima in the “red part” of the electromagnetic spectrum should be favoured (e.g. TO-PRO-3). The limiting factor in terms of penetration depth turned out to be the restricted infiltration of the ONL by different dyes even with infiltration times of more than 1h at 30° (see YZ-view in Fig. 3). As not much more than 25 µm of the tissue can be stained in z-direction the thickness of mechanical radial sections should not exceed 50 µm in this case.

4.2 Field-of-view and resolution

The use of a red fluorescent dye happens somewhat at the expense of spatial resolution but doesn’t influence the conspicuousness of cell nuclei. In fact the field-of-view of the microscope’s sensory device has to be adjusted to gather fluorescent light from all three nuclear layers of the retina at the same time (radial slices) to be able to correlate cell counts of radial staggered VOIs and to minimize the total data volume. In the examined material the distance between the vitreal border of the GCL and the scleral border of the ONL peaked at about 230 µm, fitting diagonally in the chosen field-of-view. The resulting nuclear diameters of around 15 pixels turn out to be an acceptable trade-off. Tangential optical sections of mechanically not-sectioned retina fragments were made to record VOIs of the GCL (radial scans with a z-size of ≤ 50 µm were less suitable to image this nuclear monolayer). In this orientation the resolution was increased by the factor of 1.3, the field-of-view restricted respectively.

4.3 3D-arrays

The import of image stacks into a single 3D-array in general and the use of IDL in particular allows comfortable access to every single voxel-value and to easily apply a series of powerful imaging-routines and other logical operations. Programming in a compiler language like IDL opens up the possibility to compute large data sets and frequent iterations even on standard PCs and notebooks relatively fast – ImageJ for example does the same job in a comparatively unacceptably long period of time. Of course similar approaches were developed and conducted also by other investigators for their special research problems independently [3].

4.4 Functional morphology

From the zoological perspective the semi-automated analysis of nuclear patterns opens the door to the description and interpretation of tissue-characters (not only in the retina) that hitherto has been too time-consuming or even impossible with paper and pencil. This is especially true for an accurate counting of objects (e.g. cell nuclei) not only on single microscope-slides [4], but also in high-content data stacks, also for generating neighbour-distance histograms in 3D (to describe cellular sphere-packing patterns or developmental processes at the retinal margin) and for the display of standardised density- and ratio-maps. Some examples are given to illustrate the scope of the functional morphological discussion of complex retinal characters: The density distribution of photoreceptors in the retina gives an indication of the visual acuity in different sectors of the visual field. This is, however, only reliable if the density distribution of ganglion cells mirrors the first mentioned pattern. The ratio of photoreceptors to ganglion cells in a small area reveals the degree of radial signal convergence and thus an relative indicator of light sensitivity (cone- and rod-pathways have to be analysed separately), the ratio of photoreceptors to secondary neurons, on the other hand, gives indications about the potential computing power or computing complexity of the examined retina fragment etc [5].

4.5 Outlook

To continue with this subject it is planned to expand the image analysis applications to the automated recording of shape-parameters and the high-resolution distribution of the fluorescence signal within single nuclei for cell-classification (in combination with neuroanatomical techniques), to the mapping of nuclei in hemispheric coordinate systems (e.g. small eyes) and ultimately to the full-automated adaptive acquisition and analysis of fluorescence signals from entire retinae with motorized microscopes. High-content applications would be also the comparison of developmental stages or related species.

References

1. Fineran, B.A., Nicol, J.A.C.: Studies on the photoreceptors of *Anchoa mitchilli* and *A. hepsetus* (Engraulidae). Phil. Trans. R. Soc. Lond. B **283** (1978) 25-60
2. Heß, M., Melzer, R.R., Smola, U.: The Pattern of cone pedicles and horizontal cells in the retina of the European anchovy *Engraulis encrasicolus* L. (Engraulididae, Clupeiformes). J. Submicroscopic Cytol. Pathol. 34(4) (2002) 355-365
3. Bredno, J., Metzler, V., Nacimiento, W., Lehmann, T.M., Spiter, K.: Detektion und Quatifizierung der Membranstrukturen von Nervenzellen. In: Lehmann, T.M., Metzler, V., Spitzer, K., Tolxdorff, T. (eds.): Bildverarbeitung für die Medizin. Springer Verlag, Berlin (1998) 407-411
4. O'Connell, C.P.: The structure of the eye of Sardinops caerulea, Engraulis mordax and four other pelagic marine teleosts. J. Morph. 113 (1963) 287-329
5. Archer, S.N., Djamgoz, M.B.A., Loew, E.R., Partridge, J.C., Vallerga, S. (eds.): Adaptive mechanisms in the ecology of vision. Kluwer Academic Publishers, Dordrecht, Boston, London (1999) 668pp

A General Approach to Shape Characterization for Biomedical Problems

Davide Moroni[1], Petra Perner[2], and Ovidio Salvetti[1]

[1] Istituto di Scienza e Tecnologie dell'informazione, ISTI-CNR, Pisa, Italy
davide.moroni@isti.cnr.it, ovidio.salvetti@isti.cnr.it
[2] Institute of Computer Vision and Applied Computer Sciences, Leipzig, Germany
pperner@ibai-institut.de

Abstract. In this paper, we present a general approach to shape characterization and deformation analysis of 2D/3D deformable visual objects. In particular, we define a reference dynamic model, encoding morphological and functional properties of an objects class, capable to analyze different scenarios in heart left ventricle analysis.

The proposed approach is suitable for generalization to the analysis of periodically deforming anatomical structures, where it could provide useful support in medical diagnosis. Preliminary results in heart left ventricle analysis are discussed.

1 Introduction

Deformable structures arise frequently in human anatomy and, in many cases, their deformation modes are of key importance in understanding the functional properties of the related organs and assessing their health-state. The main example is given by cardiac dynamic analysis, since many heart pathologies are correlated to the deformation pattern of the organ. In cardiac analysis, well-established imaging techniques are of great support in medical diagnosis, since they allow to acquire video sequences of the heart, from which its dynamical behavior can be inferred. However, the interpretation of the acquired data (temporal sequences of 2D/3D images, possibly from different imaging modalities) is difficult or, at least, time consuming; in daily practice, sometimes, physicians extract the most salient frames from the video sequence (end diastole and systole) and perform direct comparison among images in the selected subset. It is likely that, considering the full video sequence, more precise and rich information about the state of the heart can be discovered.

Motivated by these problems and extending the works [1, 2], we believe that it is fruitful to define, in some generality, the concept of periodically deforming visual objects (see section 2 for a precise definition) and to propose a methodological approach to their study.

Besides providing modules for structures reconstruction and characterization, that have their own importance in biomedical applications as automatic tools to speed up diagnosis, the main idea is to define a reference dynamic model of an objects class: this model can be understood as an encoding of morphological and functional properties of a periodically deforming object during its full deformation cycle. In particular, shape changes and evolution of local object properties are depicted in a coincise form in the reference dynamic model, thus allowing for deformation analysis and deformation pattern classification.

The paper is organized as follows. In section 2 we define the class of objects we are interested in, making explicit the necessary assumptions. Then in section 3, the proposed approach is outlined and its basic modules leading to the reference dynamic model are described in detail. More precisely three modules are considered: *object reconstruction* (sec. 3.1), in which every object is reconstructed in Euclidean space as a collection of manifolds, *object characterization* (sec. 3.2), in which local shape descriptors and functional features are coded into property functions and, finally, *deformation pattern assessment* (sec. 3.3) where the reference dynamic model is actually built. Preliminary results in heart dynamic analysis are then presented in section 4, whereas conclusions and directions for further work are briefly discussed in section 5.

2 Periodically deforming visual object

A visual object O embedded in the background space $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a collection

$$O = \{(V^\alpha, P^\alpha)\}_{\alpha=1,2,\ldots,k}$$

where each V^α is a smooth manifold (possibly with boundary) embedded in Ω and $P^\alpha : V^\alpha \to \mathbb{R}^{d(\alpha)}$ is a smooth *properties function* assuming its values in a suitable properties space.
The smoothness assumption is a quite common hypothesis in computational anatomy (see e.g. [3]) and it is satisfied in practice to a large extent; it implies for example that differential geometric properties (like normals, curvatures,...) can be computed everywhere. We use, moreover, collection of manifolds -instead of a single one- to be able to describe object subparts (possibly of different dimensionality) by attaching them specific salient attributes via a dedicated properties function. For example, in heart left ventricle modelling, the object of interest is the myocardium, that can be modelled as a 3D manifold, whose boundaries are two surfaces: the epicardium and the endocardium. It is convenient to attach to the boundary surfaces a different (actually richer) set of attributes than those used for internal points.

A deforming visual object $\mathcal{O} = (O_t)_{t=1,2,\ldots}$ is a temporal sequence of visual objects satisfying some smoothness constraint. Each $O_t = \{(V^\alpha, P^\alpha)\}_{1\leq\alpha\leq k}$ should be regarded as the *snapshot* of the deforming object at time t.
We require that each manifold V_t^α appearing in the snapshot at time t can be smoothly deformed into V_{t+1}^α in the subsequent snapshot. Tears or crack of any object subpart are, therefore, ruled out; moreover, in such a way, we avoid dealing with changes in topology, that would require to model shape transitions, a task far beyond our present scopes.

Finally, a periodically deforming visual object is a deforming object for which there exists an integer T such that $\forall t : O_t = O_{t+T}$. In other words, the deforming object depicts a periodic motion; thus, a periodically deforming object is characterized by a finite list of snapshots $(O_1, O_2, \ldots, O_T)$, which will be referred to as its deformation cycle.

We make a final assumption about the data available to describe a periodically deforming visual object. It is assumed that a sufficiently rich set of synchronous signals and images, possibly from different modalities, has been acquired so as to represent faithfully a physical body or phenomenon of interest. In particular, the data set should include at least one 2D/3D image sequence $(I_t)_{1 \leq t \leq T}$, from which morphology and regional properties of the object can be inferred.

3 Methodology definition

With the previous assumptions, a reference dynamic model of an object of interest is constructed by coding the dynamics of the object in a rich representation of its shape and functional properties.

The approach consists in three modules, each one performing specific tasks. Essentially, the first two modules are dedicated to extract a suitable periodically deforming visual object from image data. Then the periodically deforming visual object is analyzed and used to construct the reference dynamic model. A more precise outline of the modules used to obtain the aforementioned model is as follows:

Object reconstruction: For each phase t, the collection of manifolds $\{V_t^\alpha\}$ is identified and reconstructed in 2D/3D space by applying neural algorithms to the image sequence $(I_t)_{1 \leq t \leq T}$;

Object characterization: Morphological features and dynamic descriptors are extracted and coded in a property function P_t^α that for each point x of the manifold V_t^α returns the property vector $(P_1^\alpha(x), \ldots, P_m^\alpha(x))$, where each P_i^α represents one of the selected features;

Deformation pattern assessment: Suitable and significant shape descriptors are extracted and spatial distribution of the property functions are evaluated in order to obtain a description of the object dynamics.

In the following sections, these steps are described in more details.

3.1 Object reconstruction

The 3D reconstruction of the visual object $\mathcal{O}$ is achieved via voxelwise classification, that is by labeling each voxel in the image domain with semantic classes which describe voxel membership to the collection of manifolds $\{V_t^\alpha\}$.

The classification is performed applying an advanced neural architecture to a set of extracted features. The involved features can be divided into two classes.

First, low-level features are considered: they are context-independent and do not require any knowledge and/or pre-processing. Some examples are voxel position, gray level value, gradients and other differentials, texture, and so forth. Middle-level features are also selected, since voxel classification can benefit from more accurate clues, specific of the problem at hand. In particular, if an intrinsic reference system can be individuated to describe the object shape, it can be used to define a relative voxel position. If, in addition, a priori information about the object shape is available, a reliable clue for detecting edges in the images is given by the gradient along the normal direction to the expected edge orientation.

Moreover, a multiscale approach is adopted: the features are computed on blurred images, supplying information about the behavior of the voxel neighborhood, which results in a more robust classification.

The set of selected features are processed to accomplish the voxel classification by means of a Multilevel Artificial Neural Network (MANN), which assures various computational advantages [4]. For each voxel x, its computed features vector is splitted into vectors $\mathcal{F}_k(x)$, each one containing features of the same typology and/or correlated. Then each $\mathcal{F}_k(x)$ is processed by a dedicated classifier based on an unsupervised Self Organizing Maps (SOM) architecture. The set of parallel SOM modules constitutes the first level of the MANN which aims at clustering each portion of the feature vector into crisp classes, thus reducing the computational complexity. The output of this first level is then passed to a second and final level, consisting in a single Error Back-Propagation (EBP) module, which supplies voxel classification.
Its output describes voxel membership to the various manifolds V_t^α in the collection $\{V_t^\alpha\}_{1\leq\alpha\leq k}$.

3.2 Object characterization

The reconstructed object is further characterized by assigning a significant properties function $P_t^\alpha : V_t^\alpha \to \mathbb{R}^{d(\alpha)}$ to each manifold V_t^α.

Three types of properties are considered:

- intensity based properties;
- local shape descriptors;
- local dynamic behavior descriptors.

Examples of properties of the first type are gray level value, gradients, textures and so on. They are extracted form the image sequence I_t –the one which leads us to object reconstruction. If data collected from other imaging modalities are available, after performing registration, we can fuse this information to further annotate the object (for example, in the case of the heart, information regarding perfusion and metabolism, obtained e.g. by means of PET imaging, can be referred to the reconstructed myocardium). Geometric based properties, belonging to the second type, are extracted directly from the collection of manifolds $\{V_t^\alpha\}$, and are essential to describe locally the shape of the object. Again, we may distinguish between context independent features (automatically computable for every manifold of a given dimensionality, such as Gaussian and mean curvature for surfaces) and problem-specific properties.

Finally, the local dynamic behavior may be described by properties borrowed from continuous mechanics (such as velocity field and strain tensor); they, however, require, at least, local motion estimation, that we haven't pursued yet.

3.3 Deformation pattern assessment

The periodically deforming visual object obtained in the previous steps can be used to assess the dynamic behavior of the object and identify its deformation pattern. However, the voxelwise characterization of the reconstructed objects is not suited for state assessment. Indeed, the given description of the whole objects (collection of manifolds described by functions) has a dimensionality far too high to make the problem computationally feasible. Moreover, it would be essential to be able to compare anatomical structures belonging to different patients and, at the moment, the idea is to use a deformable model (given for example by mass-spring models [5]) and to normalize every instance of anatomical structure to that model: in this way anatomical structures (belonging to the same family) are uniformly described and can be then compared.

Combining these two issues, we should look for a new set of 'more intrinsic' features $\mathcal{F}_t$ that should be enough simple and, at the same time, capturing essential information about the objects.

To obtain these new kinds of features, global information about the objects can be extracted from the properties function, without introducing any model. For example, one may consider the 'property spectrum', by which we mean the probability density functions (PDF) of a given component of the property function $P_t^\alpha(\cdot)$. This consists in a function capturing how the property is globally distributed; thus, comparison of different property spectra is directly feasible; to reduce dimensionality, moreover, it is effective to compute the momenta of the PDF (mean, variance,. . .).

However, properties spectrum does not convey any information at all about regional distribution of the property. In practical situation, this is a drawback which cannot be ignored: for example, a small 'highly abnormal' region may not affect appreciably the PDF, but its clinical relevance is, usually, not negligible. Hence, spatial distribution of properties has to be analyzed; in some cases, approaches which do not need a refined model of the object (e.g., Gaussian image, spherical harmonics or Gabor spherical wavelets) may be suitable. However, in general one should define a model of the objects (whose primitives -elementary bricks- are regions, patches or landmarks) and then propagate it to the set of instances to be analyzed by using matching techniques. Then, we may consider the average of a property on regions or patches (or the value in a landmark) as a good feature, since comparisons between averages on homologous regions can be immediately performed.

Following this recipe, a vector of features $\mathcal{F}_t$ with the desired properties is obtained for each phase of the cycle. The deforming object is then described by the dynamics of the temporal sequence of feature vectors obtained at different phases of the deformation cycle.

A further fruitful feature transformation may be performed exploiting our assumptions on deformable visual objects. Indeed, the smoothness of deformations implies that a visual object has mainly low frequency excited deformation modes. We extend this slightly assuming that this holds true also for the features lists $(\mathcal{F}_t)_{1\leq t\leq T}$. We assume that the fundamental frequency of the motion is also the main component of each feature tracked on time. With these assumptions, an obvious choice is given by the Fourier transform, followed by a low pass filter, which supplies a new features vector Θ.

The evaluation of the above mentioned parameters $\mathcal{F}_t$, at each phase t, implicitly codifies information regarding object dynamics. Actually, we avoid defining a complex model of the object kinematics and exploit its periodic characteristic by constructing a rich representation of each phase of the deformation cycle.

4 Results

An elective case study for the presented methodology is cardiac analysis, whose clinical relevance can be hardly overestimated. We restrict our analysis to the left ventricle (LV) that, pumping oxygenated blood around the body, is the part of the heart for which contraction abnormalities are more clinically significant.

The proposed methodology is, of course, not universal, in the sense that there are some intrinsic limitations that prevent it to be potentially applied in any scenario. Indeed, our analysis is limited to a single deformation cycle and so only pathologies that affect every deformation cycle can be considered. Moreover, we require that physiological and (selected) pathological states induce different feature dynamics. This requirement is not too restrictive; actually, it is well known that many pathologies are correlated to abnormal shape patterns at end systole.

The LV structure is modelled as a 3D manifold (the myocardium) with boundary. The boundary has two connected components which are the surfaces corresponding to epicardium and endocardium.

We describe henceforth how the steps of the methodology are applied. First, the deformable visual object structure is extracted from the available data, consisting in a sequence of short axis gradient echo MR images, acquired with the FIESTA, GENESIS SIGNA MRI device (GE medical system), 1.5 Tesla, TR $= 4.9$ ms, TE $= 2.1$ ms, flip angle 45° and resolution $(1.48 \times 1.48 \times 8)$ mm. Sets of $n = 30$ 3D scans, consisting of $k = 11$ 2D slices, were acquired at the rate of 30 ms for cardiac cycles [diastole-systole-diastole]. Various clinical cases were considered, for a total of 360 scans, corresponding to 12 cardiac cycles.

To perform reconstruction, we first used a pre-processing step devoted to the automatic localization of the left ventricle cavity (LVC) [6].

The located LVC is then exploited to define an Intrinsic Reference System (IRS), given by a hybrid spherical/cylindrical coordinates system. This choice is dictated by the fact that LV approximately resembles a bullet-shaped structure;

moreover, in the IRS, image partial derivatives w.r.t. radial coordinate are an efficient clue for heart surfaces detection.

The IRS is used to extract the following features for voxel classification:

- Position w.r.t. IRS
- Intensity and Mean intensity (computed applying Gaussian filters)
- Gradient norm $||\nabla I_t||$
- Partial derivative in the radial direction $\frac{\partial I_t}{\partial r}$.

Using the 2-level ANN, voxels are classified on the basis of their features vector as belonging or not to epi- and endocardial surfaces. More in detail, the set of extracted features is divided into two vectors $\mathcal{F}_1$, $\mathcal{F}_2$ containing respectively position, intensity and mean intensity, and position, gradient norm and partial derivative $\frac{\partial I_t}{\partial r}$. The position w.r.t. IRS is replicated in both vectors because it reveals salient for clustering both features subsets. Then, the first level of the MANN consists of two SOM modules, which have been defined as 2D lattice of neurons and dimensioned experimentally, controlling the asymptotic behaviour of the number of excited neurons versus the non-excited ones, when increasing the number of total neurons [7].

A 8 × 8 lattice SOM was then trained, according to Kohonen's training algorithm[8], for clustering the features vector $\mathcal{F}_1$, while $\mathcal{F}_2$ was processed by a 10 × 10 lattice SOM.

A single EBP module has been trained to combine the results of the first level and supply the final response of the MANN. The output layer of this final module consists in two nodes, which are used separately for reconstructing the epicardium and the endocardium. Since each cardiac surface divides the space into two connected regions (one of which is bounded), each output node can be trained using the signed distance function with respect to the relative cardiac surface. In this way, points inside the surface are given negative values, whereas positive values are given to points in the outside. Henceforth the surface of interest correspond to the zero-level set of the output function.

Different architectures have been tested, finding the best performance for a network with only one hidden layer of 15 units, trained according to the Resilient Back-Propagation algorithm [9].

The voxel classification, supplied by the MANN, may be directly used for visualization purposes by using an isosurface extraction method, as shown in figure 1.

Characterization of the reconstructed structure is obtained annotating every voxel with intensity, Gaussian and mean curvature, wall thickness and IRS properties. In particular, Gaussian and mean curvature have been included as shape descriptors whereas wall thickness, which is a classical cardiac parameter, is one example of problem-specific property: it is defined as the thickness of the myocardium along a coordinate ray and it is expected to increase during contraction (since myocardium, being almost water, is, with good approximation, incompressible).

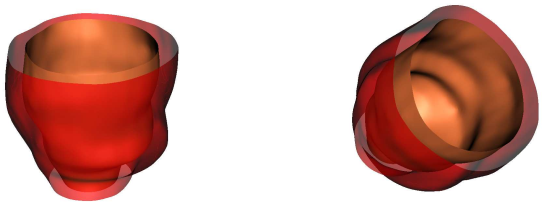


Fig. 1. Different views of the rendered left ventricle at end diastole. The surfaces are obtained applying marching cubes on the two output functions of the network. To eliminate satellites, a standard island removing procedure is applied.

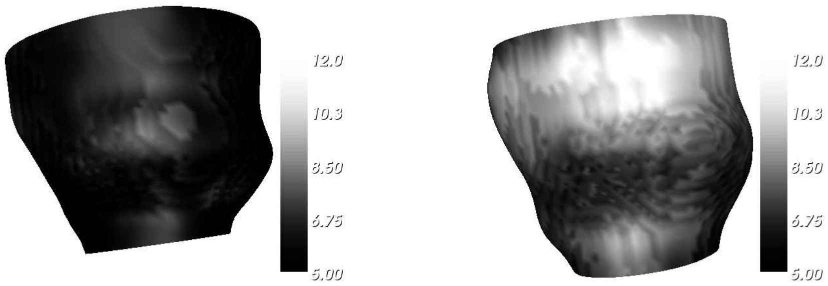


Fig. 2. Wall thickness at end diastole and systole, shown as an attribute of epicardial surface. Estimation is performed according to the centerline method and values are expressed in millimeters.

This characterization is translated in a more amenable form by computing properties spectrum and regional features. In computing spectrum, coordinates w.r.t. IRS have been disregarded, with the exception of radial coordinate; intensity has also been excluded. For any property only mean and variance have been considered. For computing regional features, so far, we used a popular model of the LV (see [10] for a review of 3D-cardiac modelling). In 2D, as shown in Figure 3, it is defined by the intersections of cardiac surfaces with a pencil of equally spaced rays. The 3D version is obtained by stacking the 2D construction along the axis of the LV.

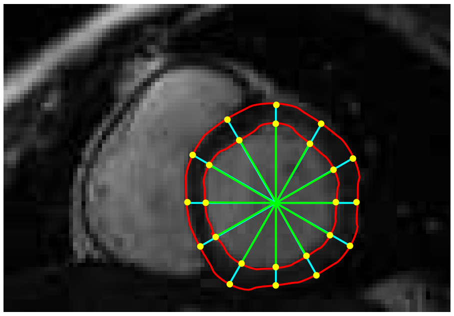


Fig. 3. The pencil of equally spaced rays used to computed local features.

5 Conclusions and further work

In this paper, we define a reference dynamic model, encoding morphological and functional properties of an objects class, capable to analyze different scenarios in heart left ventricle analysis. In particular, a framework for the shape characterization and deformation analysis has been introduced for the study of periodically deforming objects.

This framework consists of several modules performing a) object reconstruction, b) object characterization, c) pattern deformation assessment. Solutions to specific tasks proposed in each module are, to a large extent, independent and may be combined with other methods, thus broadening the potential application field of the framework. In particular, an approach based on multi-level artificial neural network has been selected as a general purposes strategy for object reconstruction, motivated by the promising results presented in [4]. A quantitative evaluation of segmentation performance, based on comparison between images automatically segmented and images annotated by a committee of expert observers, however, is still in progress.

The elective case studies are represented by the analysis of heart deformable anatomical structures. Actually, for demonstrating the effectiveness of the proposed framework, we have shown the preliminary results in the study of the heart left ventricle dynamics. The next step will be to employ the obtained results for defining a general method to classify the state of the deformable object, and, in particular, the physio-pathological states of the left ventricle.

6 Acknowledgments

This work was partially supported by European Project Network of Excellence MUSCLE - FP6-507752 (Multimedia Understanding through Semantics, Computation and Learning) and by European Project HEARTFAID "A knowledge based platform of services for supporting medical-clinical management of the heart failure within the elderly population"(IST-2005-027107).

References

1. Colantonio, S., Moroni, D., Salvetti, O.: A methodological approach to the study of periodically deforming anatomical structures. In: Proc. International Conference on Advanced Information and Telemedicine Technologies, AITTH 2005. (2005) 32–36
2. Colantonio, S., Moroni, D., Salvetti, O.: Shape comparison and deformation analysis in biomedical applications. In: Eurographics Italian Chapter Conference. (2006) 37–43
3. Grenander, U., Miller, M.I.: Computational anatomy: an emerging discipline. Q. Appl. Math. **LVI**(4) (1998) 617–694
4. Di Bona, S., Niemann, H., Pieri, G., Salvetti, O.: Brain volumes characterisation using hierarchical neural networks. Artificial Intelligence in Medicine **28**(3) (2003) 307–322
5. Di Bona, S., Lutzemberger, L., Salvetti, O.: A simulation model for analyzing brain structures deformations. Physics in Medicine and Biology **48** (2003) 4001–4022
6. Colantonio, S., Moroni, D., Salvetti, O.: MRI left ventricle segmentation and reconstruction for the study of the heart dynamics. In: IEEE ISSPIT, Athens, Greece (2005) 213–218
7. Di Bono, M., Pieri, G., Salvetti, O.: A tool for system monitoring based on artificial neural networks. WSEAS Transactions on Systems **3**(2) (2004) 746–751
8. Kohonen, T.: Self-Organizing Maps. 2 edn. Volume 30 of Springer Series in Information Sciences. (1997)
9. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: Proc. of the IEEE Intl. Conf. on Neural Networks, San Francisco, CA (1993) 586–591
10. Frangi, A.F., Niessen, W.J., Viergever, M.A.: Three-dimensional modeling for functional analysis of cardiac images: A review. IEEE Trans. Med. Imaging **20**(1) (2001) 2–5

Statistical Analysis of Myocyte Orientations of the Left Ventricular Myocardium

Kai Rothaus and Xiaoyi Jiang

Department of Computer Science, University of Münster
Einsteinstrasse 62, D-48149 Münster, Germany
{rothaus,xjiang}@math.uni-muenster.de

Abstract. The commonly used model of the heart for medical applications suffers from some incompleteness when explaining different kinds of measured forces in vivo studies by medical experts. In this paper, we make a statistical analysis of the so-called angle of intrusion automatically. The basis of the proposed method is a set of histological preparations showing heart fibre tissue. We adapt a multi-scale midline extraction process to extract the myocyte strings out of these images and measure the angles of intrusion. Furthermore, a statistical model is derived and validated by the result of a novel parameter estimation technique.

1 Introduction

In this work, we present an approach to analysing the orientation of myocyte strings automatically. For this, digitised images showing heart tissue (see Figure 1 for an example) are processed. The dark elongated structures are the strings of myocyte cells, which cause the contraction of the heart muscle. The top border of the image is oriented parallel to the epicard (the outer border of the heart). Substantially, the myocyte cells form strings, which are situated parallel to the epicard with slight variations. For medical purpose, the distribution of the myocyte orientations, also denoted as angle of intrusion, has a high impact. Using former models of the heart, where the myocyte structures are essentially ignored, the forces observed by the physicians in vivo studies [5] cannot be simulated or at least explained suitably. Lunkenheimer et al. [6] try to enhance the existing model of the heart by investigating two different kinds of observed forces. Their assumption is that there must be not only tangential directed myocyte strings, as assumed so far. Furthermore, they expect a larger portion of transversal myocyte strings. The first step to document this assumption is an appropriate analysis of the angles of intrusion, which we present in this paper.

Some former work has been done to perform quantitative assessments of myocytes by Karlon et al. [2]. They compare manual measurements with two automatic approaches. The first method is based on a Hough transform technique. Firstly, edges are computed using four different gradient masks. The responses of these filters are thresholded and a connected component analysis is performed. Afterwards, the image is divided into smaller regions and some constraints are checked to filter out false regions. On all remaining regions, a Hough transform is performed to compute one mean orientation of the structures in each. The mean orientations of all regions are collected and constitute the observation set.

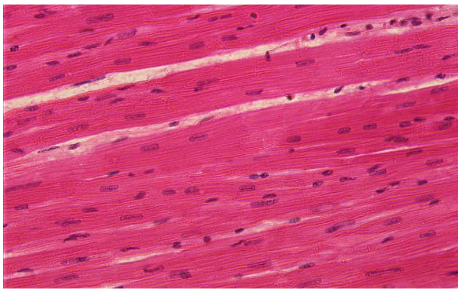


Fig. 1. Sample image of heart tissue: Myocyte strings are visible as dark elongated structures.

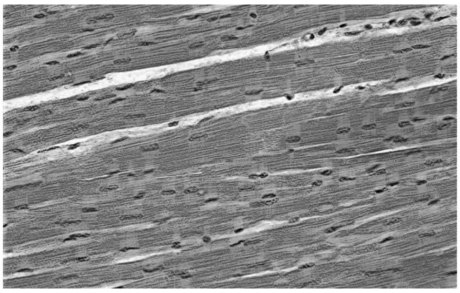


Fig. 2. Grey-level transform and adaptive contrast enhancement of the slice in Figure 2.

The second method described by Karlon et al. [2] is based on the intensity image gradient directly. Once again, the image is divided into regions and a statistical analysis is performed in each. As statistical model, the class of von Mises distributions with parameter kappa (for statistical background see [1]) is used. Since the result of the two methods are justified by a manual analysis, an automatic analysis of the angle of intrusion is well founded [2].

In contrast to these two methods, our approach does not divide the images into regions. Since we are not only interested in the mean angle of intrusion, but also want to analyse the underlying distribution, we try to take as much information as possible into consideration. For this, we locate the midlines of the myocyte strings in the images and take the tangential vectors at sample points of the extracted midlines as observations.

The remainder of this paper is organised as follows. At first, we describe the image analysis part (Section 2). This process results in the observation set consisting of local measured angles of intrusion. In Section 3 we give the statistical analysis of the observation set. We will present two different distribution models, based on classes of von Mises and Gaussian distributions, respectively. Afterwards, the results of our approach are shown (Section 4). Finally, we end up by drawing some conclusions (Section 5).

2 Analysis of heart fibre images

The histological preparations are cuts of pig hearts, which are dissected using pairs of cylindrical knives with different diameters (see [6]). After pinning the slices flat, they are fixed in formaldehyde, embedded in paraffin and sectioned. A treatment with several substances is done to achieve a swelling of the preparations, so that the myocyte strings are clearly visible. The colour images show these preparations 100 times magnified. All slices are adjusted with the upper border parallel to the epicard and recorded on digital camara.

The image analysis part of our method consists of three steps: enhancement of the myocyte strings (Section 2.1), midline extraction (Section 2.2) and measurements of the tangential orientations of the midlines at equidistant sample points (Section 2.3).

2.1 Image Acquisition and enhancement of the myocyte strings

At first, we transform the colour images to intensity images $I(x,y)$ by a linear combination of the three RGB colour channels.

$$I(x,y) = 0.2626 \cdot R(x,y) + 0.4116 \cdot G(x,y) + 0.3258 \cdot B(x,y) \tag{1}$$

This linear combination scheme is computed once by analysing a subset of images by a principal component analysis, to keep as much contrast in the images as possible.

Afterwards, we enhance the contrast of the image $I(x,y)$ by an adaptation of the method proposed by Yu et al. [9]. This method works by propagating the minimum ($lmin$), average ($lavg$) and maximum ($lmax$) value towards different scan directions by a conditional propagation scheme. The initialisation of the three arrays $lmin$, $lavg$ and $lmax$ is the image, which should be processed. In contrast to the original approach of Yu et al. [9] we use different conductivity factors for the minimum ($C_{\text{min}} = 0.95$), average ($C_{\text{avg}} = 0.75$) and maximum ($C_{\text{max}} = 0.55$) values, to steer the propagation behaviour of the enhancement algorithm. Thus, we use the conditional propagation scheme at the actual scanned image position:

$$lavg = (1 - C_{\text{avg}}) \cdot lavg \; + \; C_{\text{avg}} \cdot \overline{lavg} \tag{2}$$

$$lmin = (1 - C_{\text{min}}) \cdot lmin \; + \; C_{\text{min}} \cdot \overline{lmin} \qquad \text{iff } lmin \; > \; \overline{lmin} \tag{3}$$

$$lmax = (1 - C_{\text{max}}) \cdot lmax \; + \; C_{\text{max}} \cdot \overline{lmax} \qquad \text{iff } lmax \; < \; \overline{lmax} \tag{4}$$

where the bar denotes the value at the previous scanned image position. Furthermore, we choose two scan directions, namely from top to bottom and vice versa. This adaptation of the approach is motivated by our goal to keep the dark structures, but lighten the bright structures to enhance the contrast. For each pixel p the three resulting values $lmin$, $lavg$, $lmax \in [0,1]$ reflect the minimum, average and the maximum intensity in a neighbourhood of p. The original intensity old at p can now be emphasised against its neighbourhood. Therefore, we define the local intensity range as $\delta = lmax - lmin$ and the local enhancement factor as $\omega = \sqrt{\delta \cdot (2 - \delta)}$. The new intensity value at p is then computed as (adaption of the transformation proposed by Yu et al. [9])

$$\frac{lmax + lmin - \omega}{2} + \frac{2\,\omega\,(old - lmin)\,(\delta + \omega\,(lavg - old)\,(old - lmax))}{\delta^3}. \tag{5}$$

The result of this grey level transform and enhancement step applied on the preparation of Figure 1 is shown in Figure 2.

2.2 Midline extraction

After the pre-processing procedure (Section 2.1), now the extraction of the midlines will be explained. For this task we have developed a multi-scale extension [7] of López's Level-Set-Extrinsic-Curvature approach (LSEC) [3, 4]. In the following, we give a brief summary of this extension. The pixel array I of the enhanced intensity image (Figure 2) is taken as input image, in which we want to localise the midlines of dark elongated structures (myocyte strings).

(1) Computation of the intensity gradient vector field:
After a slight smoothing of I with a Gaussian kernel ($\sigma = 0.75$) we apply the Sobel operator, which results in the partial deviation I_x and I_y. These deviations are used to

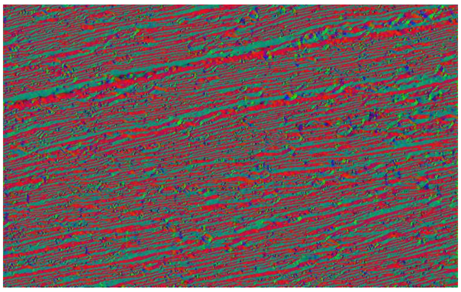


Fig. 3. Gradient image after multi-scale smoothing of the gradient vector field

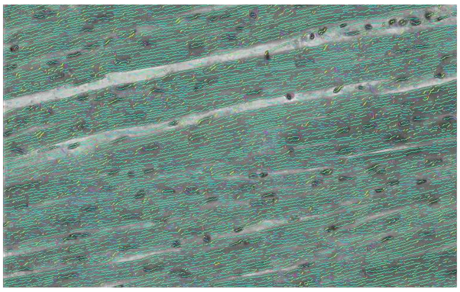


Fig. 4. Extracted midlines are laid over the intensity image.

compute the edge magnitude array S and a local edge orientation array Θ. Since at each pixel, the corresponding deviations I_x and I_y could be recomputed on S and Θ, in the following we use the notation (I_x, I_y) or (S, Θ) for the gradient vector field, equivalently.

(2) Enhancement of edge magnitude:
Since the LSEC approach of López et al. [3, 4] is valid on normalised gradient images only, we boost the edge magnitudes S pixel-wise using the function

$$b_t(S) = 1 - \exp\left(-\frac{S^2}{2 \cdot t^2}\right) \tag{6}$$

with threshold $t = 0.075$. This optimistically choice is made to preserve even weak edges, since they would be filtered out in the further process, if there are no equally orientated edges in the neighbourhood.

(3) Iterative smoothing process:
The goal of this step is to smooth the gradient vector field in the sense that the gradient vectors are propagated towards the interior of dark elongate structures. Firstly, the structured tensor is computed for each pixel

$$ST(x,y) = \begin{pmatrix} I_x(x,y)^2 & I_x(x,y) \cdot I_y(x,y) \\ I_x(x,y) \cdot I_y(x,y) & I_y(x,y)^2 \end{pmatrix}, \tag{7}$$

where (x, y) are the image coordinates. This tensor field is smoothed element-wise with Gaussian kernels G_k of different scales $\sigma_k = k \cdot \sigma_0$ ($\sigma_0 = 0.75$). For each pixel (x, y) and each scale we compute the edge magnitude $S_k(x, y)$ and the new edge orientation $\Theta_k(x, y)$ based on the largest eigenvalue and corresponding eigenvector of the smoothed structured tensor (for details see [7]).

Finally, for the pixel (x, y) the vector with the highest edge magnitude value $S_k(x, y)$ of all considered scales k is chosen as the result vector of the iterative smoothing process. In Figure 3 the result of the sample image (Figure 1) is shown. The magnitude S of the vectors are visualised by the intensity (V) and the direction Θ by the colour (H) in the HSV colour space.

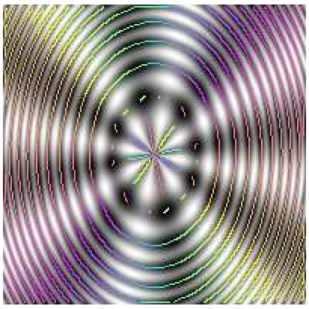


Fig. 5. Example image of López et al.[3].

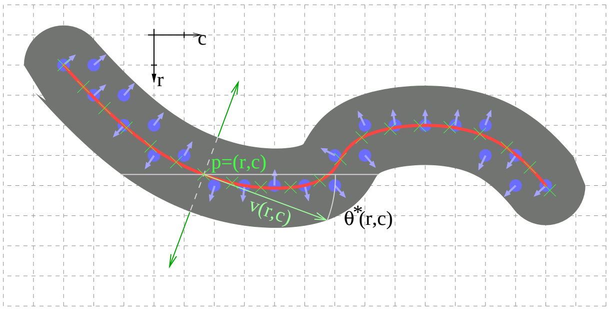


Fig. 6. Equidistant resampling of the midlines and computing of the angle of intrusion

(4) Computation of local creaseness in the intensity image: After a single additional smoothing, we apply the divergence operator on the smoothed gradient vector field (I_x, I_y). The level set extrinsic curvature is a creaseness measure of an image function I. The midlines we want to detect consist of pixels with maximum creaseness. Therefore, the level set extrinsic curvature is computed as the negative divergence κ of the smoothed gradient at each pixel. López et al. [3, 4] have proved, that this is equivalent to the direct computation

$$\kappa(x, y) = \frac{2I_x(x, y)I_y(x, y)I_{xy}(x, y) - I_y^2(x, y)I_{xx}(x, y) - I_x^2(x, y)I_{yy}(x, y)}{\left(I_x^2(x, y) + I_y^2(x, y)\right)^{3/2}} s \quad (8)$$

of the LSEC in continuous domains under certain preconditions. In fact, the divergence of gradients gives even better results in discrete domains than the direct computation. This advantage is depicted in the example image of Figure 5. The divergence approach leads to continuous segments, where as the direct computation leads to gaps [4].

(5) Grouping points of maximal creaseness to line segments: Pixels, which hold a local maximum creaseness, value in direction of the local smoothed gradient vector are taken as candidates for midline pixels. We link two neighboured candidates together if the gradients and the creaseness at the corresponding pixel are similar. This can be done by simple threshold rules. After this grouping process and an additional filtering step (discard segments of less than three pixels), we get the midline segments, which represent the myocyte strings. Result of this step are presented in the Figures 4 and 5, where the extracted midline segments are overlaid on the grey-scale images.

2.3 Measurement of the tangential orientations

The midlines are represented as strings of neighboured midline pixels (see Figure 6). With each midline pixel (blue dots) the smoothed gradient vector is stored (connected arrows). The drawback of this representation is that diagonal midlines are represented by pixels, which form a stairway, but a straight pixel line represents horizontal midlines. Obviously, this representation does not regard the true length of the midline.

Since we want to make a statistical analysis of the myocyte orientation, we have to resample the midlines at equidistant points. We decide to choose the width of one pixel as distance, so that a midline of n pixel length should be represented by $n + 1$ midline pixels. This resampling can be done by scanning over the midline and interpolating the pixel coordinates as well as the assigned gradient vectors using the nearest two pixels on the midline. Thereby, the tangential of the midline defines the scanning direction, which is orthogonal on the smoothed gradient vector at the last considered midline point. In Figure 6 the resulting pixel coordinates (green crosses) are shown. In this figure the computation of the tangential vector v at pixel p with coordinates (r, c) is drafted.

Naturally, the situation in Figure 6 is idealised, but in fact we are mainly interested in the analysis of the tangential vectors. For the purpose of midline visualisation, we keep the representation at pixel grid points, whereas for the statistical analysis we choose the equidistant representation with sub-pixel accuracy.

3 Statistical data analysis

The tangential vectors at the equidistantly distributed sample points of the myocyte string midline are taken as observation set. Obviously, these vectors are represented in a cyclic domain with period of 180^{o} or π, respectively. For this reason, we have to derive a model for cyclic data spaces. Fisher [1] gives a general introduction in the statistical analysis of such data spaces. In the following, we treat each observed angle as a point on a circle. Since the period of the domain is 180^{o}, an angle does not correspond to the normal angle in the Euclidian manner. For the purpose of geometrical interpretations of the observation set in the cyclic domain, all angles have to be multiplied by 2.0.

One can observe three common characteristics of the observed angles of intrusions:

1. Presence of equally distributed noise, introduced by falsely detected structures.
2. The angles are unimodal distributed (see Figure 7) with only a slight variation.
3. The distribution seems to be symmetrical.

Based on these observations, we have derived a suitable model for the underlying distributions. Due to the first characteristic, the noise is regarded as an additive constant term α/π, where α is the portion of noise. To model the signal (i.e. the angle of intrusion) a symmetrical density function with one local maximum should be used. In Section 3.1 we present two models (Gaussian and von Mises, respectively). The parameter estimation procedure is the same for both models. Since the equally distributed noise has no impact to the calculation of the mean orientation, at first the mean orientation $\hat{\mu}$ is estimated. We have tested several estimators [8] and figured out that a least median error approach works best (Section 3.2). Subsequently, the portion of noise is estimated by inspecting the neighbourhood of the antipole of $\hat{\mu}$ (Section 3.3). At last we compute the estimation of the shape parameter (Section 3.4), which are $\hat{\kappa}$ (von Mises) and $\hat{\sigma}$ (Gaussian), respectively.

3.1 Distribution models

Both distribution models, which we have taken into considerations, are explained by three parameters: the portion of noise $\alpha \in [0, 1]$, the mean angle $\mu \in [-\pi/2, \pi/2)$, and one shape parameter. In detail, the PDFs are given as:

$$p^{\text{Gauss}}_{\alpha,\mu,\sigma}(\phi) = \frac{\alpha}{\pi} + \frac{(1-\alpha)\cdot\exp(-\frac{(2\phi-2\mu)^2}{2\sigma^2})}{\pi\,\sigma} \tag{9}$$

$$p^{\text{von Mises}}_{\alpha,\mu,\kappa}(\phi) = \frac{\alpha}{\pi} + \frac{(1-\alpha)\cdot\exp(\kappa\,\cos(2\phi-2\mu))}{\pi\,I_0(\kappa)} \tag{10}$$

where I_0 is the modified Bessel function of order zero. The first summand of these density functions represents the equally distributed portion of noise with fraction α and the second summand (with fraction $1 - \alpha$) is an adaptation of the signal distribution to the cyclic domain of $\phi \in [-\pi/2, \pi/2)$.

The Gaussian model cannot be applied directly to the observation set, due to the cyclicity of the data space. For this, we have to cut the data space at the antipole of the mean angle. Afterwards, the whole observation is mapped to $\mathbb{R}$ linearly, such that the mean value is mapped to 0 and finally, a standard analysis on the line is performed to estimate the standard deviation σ of the underlying Gaussian.

The use of a von Mises model offers the evidence that the shape parameter κ is directly deducible on the cyclic domain. Note that a reduction of κ leads to an enlargement of the variance. The von Mises distribution is a kind of standard distribution on cyclic domains (see [1]). Its importance is comparable with the role of the normal distribution on a line. Unfortunately, things becomes more difficult on cyclic domains, so that not all properties of the normal distribution are adaptable to von Mises distributions. The close relationship becomes clear, by inspecting the density function of the von Mises model. If the cosine in Equation (10) is approximated by its first ordered Taylor polynomial the model is (beside normation and $\kappa := \sigma^{-2}$) the same as the Gaussian (see Eq. (9)). However, in all conducted experiments the von Mises distribution seems to model the observation more precisely (Table 1).

For validating the goodness of our models and to compare different estimators, we have used the following quality measure, denoted as match score (MS). The data space is divided into finite intervals of equal length (we use 180 bins $b_1, \dots b_{180}$ of 1^{o} width as default). For this discretisation we construct the relative histogram h of the observed angles of intrusion and compare for each bin b_i the measured frequency $h(i)$ with the expected frequency $f(i)$. Assuming a distribution with parameter vector $\boldsymbol{p}$ the match score is computed as:

$$\text{MS}(\boldsymbol{p}) \; = \; 1 - 0.5 \cdot \sum_{i=1}^{180} \, |f(i) - h(i)|. \tag{11}$$

3.2 LME-estimation of the mean orientation

For a fixed μ the median error to the observation set $S = \{\phi_1, \dots, \phi_n\}$ is computed as

$$\text{mederr}(\mu, S) = \text{median}_{i \in \{1,\dots,n\}} \; \{ \, d_{\text{arc}} \, (\mu, \phi_i) \, \} \tag{12}$$

where d_{arc} is the arc-length distance in our cyclic domain. Now, we minimise this error function err(μ) to estimate μ

$$\text{err}(\mu) = \min_{\mu \in [-\pi/2, +\pi/2]} \; \{ \, \text{mederr}(\mu, S) \, \} \, , \tag{13}$$

$$\hat{\mu} = \arg\min_{\mu \in [-\pi/2, +\pi/2]} \; \{ \, \text{mederr}(\mu, S) \, \} \, . \tag{14}$$

The statistical interpretation of this method is minimising the width of the 50%-quantile interval centred at μ.

The computation of $\hat{\mu}$ can be done by a single scan through the sorted observation set $\phi_{(1)}, \dots, \phi_{(n)}$. We use two cyclic indices i_{A} and $i_{\Omega} = i_{\text{A}} + n/2$, which represent the beginning and the ending of a candidate error interval. For each error interval the corresponding estimation for μ is given by $(\phi_{(i_{\text{A}})} + \phi_{(i_{\Omega})})/2$. The estimation $\hat{\mu}$ is computed as the centre of the shortest candidate error interval and can be found by one sweep.

3.3 Estimating the portion of noise

After estimating the mean direction $\hat{\mu}$, we utilise the characteristic that the variation of the data is less in comparison to the domain. Therefore, we assume that all observed angles in a small interval centred at the antipole $\tilde{\mu} := \hat{\mu} + \pi/2$ of $\hat{\mu}$ are introduced by noise. In our implementation, we choose an interval length of $0.2 \cdot \pi$, which gives appropriate results. Since we assume equally distributed noise on the whole domain, we can easily estimate the overall portion of noise: Let n be the size of the observation set and c be the counted angles in the interval $]\tilde{\mu} - 0.1\pi, \tilde{\mu} + 0.1\pi]$ the portion of noise could be estimated as $\hat{\alpha} = 5c/n$.

3.4 Estimating the shape parameter

Finally, we estimate the shape parameter, which is κ in the von Mises case. This can been done by inspecting the minimised error $err := \mathrm{err}(\hat{\mu})$ (see Eq. (13)). We compute $\hat{\kappa}$ as the unique solution of the equation

$$0.5 = \int_{-err}^{err} p_{\hat{\mu},\hat{\kappa},\hat{\alpha}}(\phi)\ \mathrm{d}\phi\ . \tag{15}$$

Unfortunately, this equation has no analytical solution. We found κ by using the fact that the right side of the objective functional is increasing in $\hat{\kappa}$, so that we can guarantee to find κ by a binary search algorithm. This estimation technique can be adapted to other distributions, which have one shape parameter (e.g. the Gaussian model).

4 Results

Our image database consists of 45 images showing heart tissue. All slices are arranged in the common coordinate system, such that the epicard is parallel to the upper border of the image. We have estimated the parameter of the proposed model for all images and compute the match score (see Table 1). Furthermore, the average of the single results are plotted in the row ’Average’ and the analysis of the whole observation set, which consists of the union of all observation sets are presented in the row ’Complete Set’. This combination is valid because all slices are referenced to the same context.

The first observation is that in every case the von Mises model outperforms the Gaussian model by mostly over 0.5 match score points. The average match score of the von Mises model is with 92.19% reasonable high (98.24% in the case of simulated data) and validates our model. In Figure 7 we have plotted the observation histogram against the derived PDF (normalised to the same integration area) of the sample slice. The approximation of the model with the observation is good, except near the mean orientation. In some other slices, we obtain local marginal inaccuracies other where, but not at a fixed position. This phenomenon could be explained with local differences in the heart tissue, which exhibit different structures at different locations of the heart muscle. Since the preparations in our test study are located all over the heart, combining all results eliminate local phenomena. In Figure 8 the observation histogram of the aggregated observations and the derived PDF are overlaid. The visible model-observation coherency in this plot demonstrates the overall goodness of our model. This is also confirmed by the match score indicator of the complete evaluation of 94.87%. Here, the match score 93.98% of the Gaussian estimation is clearly lower.

The estimations of μ and κ correspond to the expectancy of the medical experts and in their opinion confirm the need of a more accurate model of the heart [6].

Preparation	Size	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\alpha}$	MS von Mises	MS Gaussian
Slice 01	57728	-3.08	10.80	2.97%	91.24%	91.09%
Slice 02	54608	10.07	7.61	4.07%	91.05%	90.58%
Slice 03	61283	7.42	9.08	3.26%	92.40%	92.00%
Slice 04	69057	16.72	12.97	1.77%	91.62%	91.36%
Slice 05	60890	17.83	7.35	2.94%	93.29%	92.83%
Slice 06	58330	-3.44	5.86	7.02%	92.54%	91.88%
Slice 07	66060	3.69	8.66	3.84%	92.40%	91.96%
Slice 08	61291	-1.79	7.30	5.42%	92.33%	91.85%
Slice 09	55559	-11.93	4.98	7.50%	93.25%	92.56%
Slice 10	60250	0.15	7.53	6.43%	91.85%	91.39%
Slice 11	56364	-5.46	4.82	11.34%	92.09%	91.31%
Slice 12	60266	-1.22	4.70	12.05%	92.00%	91.20%
Slice 13	56292	4.78	4.74	6.47%	93.31%	92.45%
Slice 14	57900	-0.01	5.41	7.38%	90.84%	90.16%
Slice 15	54972	6.40	4.14	8.53%	92.58%	91.61%
Slice 16	64023	-3.82	7.45	5.40%	90.29%	89.80%
Slice 17	58538	-7.73	4.15	5.55%	93.72%	92.91%
Slice 18	60359	-0.43	5.79	6.20%	93.01%	92.39%
Slice 19	62503	2.18	8.38	3.18%	93.61%	93.22%
Slice 20	59616	0.09	5.21	4.34%	92.45%	91.76%
Slice 21	60453	-1.60	5.47	8.92%	91.74%	91.08%
Slice 22	60085	3.74	6.77	6.92%	91.59%	91.03%
Slice 23	61623	-5.41	3.76	10.73%	93.11%	92.10%
Slice 24	60643	-1.25	4.03	4.37%	93.34%	92.36%
Slice 25	61079	2.63	5.97	6.20%	92.62%	91.96%
Slice 26	63194	17.74	5.15	8.37%	92.85%	92.13%
Slice 27	58143	1.53	4.29	10.27%	92.62%	91.75%
Slice 28	59888	2.48	9.66	6.13%	91.29%	90.93%
Slice 29	60537	1.30	4.62	12.31%	90.92%	90.25%
Slice 30	57753	5.82	5.39	9.51%	89.81%	89.10%
Slice 31	56389	0.76	6.30	12.16%	91.54%	91.00%
Slice 32	54771	0.55	6.28	8.90%	90.59%	89.98%
Slice 33	64169	-2.17	6.50	5.71%	91.05%	90.51%
Slice 34	66315	-0.72	4.47	4.62%	92.04%	91.20%
Slice 35	54963	-11.74	5.44	8.15%	93.65%	93.00%
Slice 36	58668	-0.10	4.00	7.74%	92.70%	91.74%
Slice 37	63878	-8.31	6.11	6.46%	90.94%	90.39%
Slice 38	58285	-1.47	7.20	4.90%	92.07%	91.63%
Slice 39	56380	-0.88	4.53	7.05%	93.42%	92.63%
Slice 40	54885	-0.35	6.16	10.06%	92.53%	91.99%
Slice 41	60458	-12.58	5.44	7.05%	93.86%	93.20%
Slice 42	53136	1.31	3.29	12.62%	92.33%	91.17%
Slice 43	47940	4.96	8.58	6.39%	93.73%	93.34%
Slice 44	62925	17.67	10.12	2.33%	91.56%	91.21%
Slice 45	66382	8.60	13.71	2.33%	90.78%	90.51%
Average	59530	1.18	6.45	6.80%	92.19%	91.57%
Complete Set	2678831	1.17	4.27	6.75%	94.87%	93.98%

Table 1. Results of the statistical analysis on the whole image data base

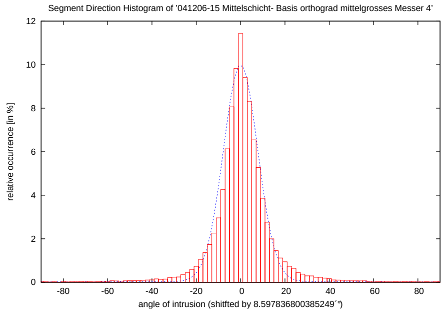


Fig. 7. Angle of intrusion histogram for the example in Figure 1.

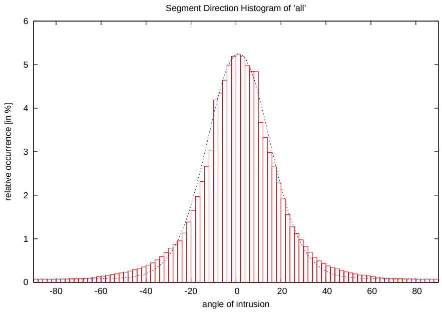


Fig. 8. Angle of intrusion histogram collected in 45 slices.

5 Conclusions and further work

In this paper, we have presented a completely automatic method for analysing the angle of intrusion of myocyte strings in heart tissue slices. Firstly, we have described an algorithm to extract the myocyte string and then given a method to measure the angle of intrusion at a multitude of sample points. Furthermore, we have developed a statistical model for the angle of intrusion distribution and validated this model experimentally.

Motivated by the results we want to advance the improvement of the heart model, by extracting the structure of the myocyte strings, which are connected to each other. Moreover, we are discussing with other research groups if a simulation of a heartbeat based on such image material could be possible.

References

1. N. I. Fisher. *Statistical analysis of circular data*. Cambridge University Press, 1993.
2. W. J. Karlon, J. W. Covell, A. D. McCulloch, J. J. Hunter, and J. H. Omens. Automated measurement of myofiber disarray in transgenic mice with ventricular erxpression of ras. *The Anatomical Record*, 252(4):612–625, 1998.
3. A. M. López, F. Lumbreras, and J. Serrat. Creaseness from level set extrinsic curvature. In H. Burkhardt and B. Neumann, editors, *Proc. of the 5th ECCV*, pages 156–169, 1998.
4. A. M. López, F. Lumbreras, J. Serrat, and J. Villanueva. Evaluation of methods for ridge and valley detection. *IEEE Trans. on PAMI*, 21(4):327–335, April 1999.
5. P. P. Lunkenheimer, K. Redmann, J. Florek, U. Fassnacht, C. W. Cryer, F. Wübbeling, P. Niederer, and R. H. Anderson. The forces generated within the musculature of the left ventricular wall. *Heart*, 90:200–207, February 2004. PMID: 14729798.
6. P. P. Lunkenheimer, K. Redmann, N. Kling, K. Rothaus, X. Jiang, C. W. Cryer, F. Wübbeling, P. Niederer, S. Y. Ho, and R. H. Anderson. The three-dimensional architecture of the left ventricular myocardium. *Anat. Rec. (in press)*, 2006.
7. K. Rothaus and X. Jiang. Multi-scale midline extraction using creaseness. In *Proc. of the ICAPR*, volume 2, pages 502–511, August 2005. LNCS 3687.
8. K. Rothaus and X. Jiang. Comparison of methods for hyperspherical data averaging and parameter estimation. In *Proc. of the ICPR (accepted for publication)*, August 2006.
9. Zeyun Yu and Chandrajit L. Bajaj. A fast and adaptive method for image contrast enhancement. In *Proc. of the ICIP*, volume 2, pages 1001–1004, October 2004.

Statistical analysis of electrophoresis time series for improving basecalling in DNA sequencing

Anna Tonazzini and Luigi Bedini [*]

Istituto di Scienza e Tecnologie dell'Informazione - CNR
Via G. Moruzzi, 1, I-56124 PISA, Italy
anna.tonazzini@isti.cnr.it

Abstract. In automated DNA sequencing, the final algorithmic phase, referred to as basecalling, consists of the translation of four time signals in the form of peak sequences (electropherogram) to the corresponding sequence of bases. Commercial basecallers detect the peaks based on heuristics, and are very efficient when the peaks are distinct and regular in spread, amplitude and spacing. Unfortunately, in the practice the signals are subject to several degradations, among which peak superposition and peak merging are the most frequent. In these cases the experiment must be repeated and human intervention is required. Recently, there have been attempts to provide methodological foundations to the problem and to use statistical models for solving it. In this paper, we exploit a priori information and Bayesian estimation to remove degradations and recover the signals in an impulsive form which makes basecalling straightforward.

1 Introduction

In automated sequencing, a reaction of extension from the initial primer of a given DNA strand generates a complete set of fragments in which the last base is marked with a fluorescent dye out of four different types, one for each type of base. Fragments are then sorted by length by means of electrophoresis and detected, as they pass under a laser, by four optical sensors, capturing the emission in the distinct wavelength ranges where the four dyes emit. The result is an electropherogram, that is four time series in the form of peak sequences, each representing the variation with time of the concentration of DNA fragments ending with the same base. Each peak in the four signals represents a base, its size is related to the number of DNA fragments of a given length, while its time location is related to the specific length and reflects the position of the base in the DNA strand under consideration. Basecalling is the final algorithmic phase of automated sequencing, and consists in obtaining the DNA base sequence from the electropherogram by the ordered reading of the peaks.

[*] This work has been supported by the European project Network of Excellence MUSCLE FP6-507752 (Multimedia Understanding through Semantics, Computation and Learning)

The most popular commercial basecallers are the software developed by ABI [1], running on the ABI Prism sequencers, and Phred [5] [6], which has been used in the Genome Project. Both are based on peak detection algorithms refined with heuristics, and are very efficient when the peaks are well distinct and quite regular in spread, amplitude and spacing.

Unfortunately, in the practice data production is subject to several processes that lead to degradations of the electropherograms, particularly near the end of the sequence. Among those, the most important and frequent are peak superposition, known as cross-talk, due to the spectral overlapping between fluorescent dyes, and peak merging, known as diffusion, due to mobility shifts and deviations of the fragments in the gel. Signal leakage may also occur, resulting in secondary peaks. These degradations may seriously affect the performance of basecalling algorithms, and, in the current practice, they entail repeating the experiment, comparing the base sequence with that of the complementary strand [7], and manual editing.

The availability of economic DNA sequencers and reliable and fast basecalling algorithms, which allow to reduce as much as possible the intervention of human operators, is still an open issue and is especially important in order to cope with large scale sequencing of whole genomes, sequencing of the genomes of as many as possible species, comparative genomics and evolutionary studies, and the increasing diffusion of sequencing of individual DNA segments in the clinical practice. Furthermore, accuracy up to a single base would be essential for reliable locations of SNPs, and for the efficiency of gene prediction software, e.g. to avoid premature termination due to false stop codons.

In the literature, there have been several attempts to provide methodological foundations to basecalling, and to use statistical models which allow the incorporation of prior knowledge about the structure of the problem and the data directly into the basecalling algorithm, without resorting to heuristics [3], [8], [10], [11], [12]. In particular, in [8] hidden Markov models and Markov chain Monte Carlo methods are used.

In this paper, the problem of removing cross-talk and diffusion in electropherograms is formulated as one of joint blind source separation and blind deconvolution. In particular, Bayesian estimation and a priori information are exploited to recover the signals in an impulsive form which makes the task of basecalling straightforward.

2 Problem formulation and Bayesian estimation

In ideal conditions (i.e. same velocity for all fragments of a given length, fluorescence emission in four separated wavelength ranges) the ideal electrophoretic signal s_j, $j = 1, 2, 3, 4$, would be an impulse train, where the impulse locations identify the mutual positions of the bases of type j out of four types (namely, A, T, C and G) along the DNA strand under consideration, and the impulse magnitudes vary in time according to the changing color concentration, i.e the changing number of fragments of a given length. It is immediate to see that in such a case the task of basecalling would be straightforward. Conversely, the fragment mobility is a random process, subject to variations due to the nature of the experiment, and the $j-th$ measured signal x_j, $j=1,2,3,4$, represents the intensity of fluorescence emitted in one of four wavelength ranges, where the emission spectra overlap. The electropherograms can thus be considered as the result of the application to the ideal signals of two operators in cascade: a convolution with a kernel related to the mobility distribution (diffusion), followed by a mixture of the four signals, modelling the superposition of the fluorescence emission spectra in four wavelength ranges (cross-talk). The data model we consider is thus:

$$x_i(t) = \sum_{j=1}^{4} A_{ij} \left(h_j * s_j\right)(t) + b_i(t) \qquad \forall t, \quad i=1,2,3,4 \tag{1}$$

where b_i is a noise term incorporating the left error sources, A is the 4×4 unknown cross-talk matrix, and h_j is the unknown impulse response which models the diffusion effect. This is related to the peak shape, depending on the casuality of fragment mobility. In general, h_j can be considered a Gaussian function, with unknown and time-varying variance. Indeed, longer fragments are more prone to mobility variations, so that it is expected that the variance of the impulse response slowly increases with time. Asymmetric, heavy-tailed peaks, due to deviations in the gel of long fragments, could be modelled as a mixture of Gaussians.

In the current practice, noise removal, cross-talk correction (also referred to as color separation) and deconvolution are performed off-line and separately, as steps of pre-processing of the electropherogram, prior applying basecaller software. As per the noise, this is assumed as constituted of two terms. A white, Gaussian term is suitable for modelling error sources such as fluorescent impurity in the gel, electronics and light scattering, and, since the actual DNA fluorescence is a very slowly varying signal, low pass filtering is usually employed to filter this noise out. Another noise term is the baseline, i.e. an error term due to a constant value of background fluorescence, which is modelled as a slowly increasing function of time. The correction of the baseline error is usually performed by removing a roughly constant waveform from the recorded signals.

To eliminate the cross-talk between the four channels of the electropherogram, usually a linear operator is applied. As already said, cross-talk is due to the overlapping of the outputs of the optical filters that separate the fluorescence from each of the four tags. This overlapping is linear and can be modelled through a mixing matrix. When this matrix is known, the cross-talk can be eliminated by applying its inverse to the data. The mixing matrix, however, is not known, and must be determined. To this purpose, techniques mostly based on analysis of the second order statistics of the signal have been proposed.

Given the data model of eq. 1, our aim is instead to jointly perform estimation of mixing and diffusion, and color separation and deconvolution, using a priori knowledge that one might have about the problem. Thus, recovering the ideal signals, i.e. removing the cross-talk and peak spreading effects, is seen as a problem of joint blind source separation and blind deconvolution. In a Bayesian framework, we propose a Maximum A Posteriori (MAP) estimate for the unknowns of the problem:

$$(\hat{\mathbf{s}}, \hat{A}, \hat{\mathbf{h}}) = arg \max_{\mathbf{s},A,\mathbf{h}} P(\mathbf{s}, A, \mathbf{h}|\mathbf{x}) = arg \max_{\mathbf{s},A,\mathbf{h}} P(\mathbf{x}|\mathbf{s}, A, \mathbf{h})P(\mathbf{s})P(A)P(\mathbf{h}) \quad (2)$$

where $P(\mathbf{x}|\mathbf{s}, A, \mathbf{h})$ is the noise distribution, and $P(\mathbf{s})$, $P(A)$, and $P(\mathbf{h})$ are the prior distributions for the three independent sets of variables. At present, we consider the noise term to be in its whole a white, Gaussian and stationary process. Although in our approach the baseline is considered incorporated in a generic Gaussian noise term, from experiments conducted on synthetic data we have seen that the proposed method is robust enough against non-stationary noise, whose variance is slowly increasing with time.

The prior adopted for the signals has been chosen on the basis of the minimum number of constraints one may reasonably enforce on the expected, ideal output of the electrophoresis process. In blind source separation, when as in our case both the mixing and the sources are unknown, a typical constraint which is enforced to sort out a solution from the infinite ones which fit the data, is statistical independence of the sources. This approach has given rise to a number of very efficient methods and algorithms known as independent component analysis (ICA) [2] [4] [9]. In our case, however, ICA is not suitable, since we know that the four electrophoretic signals should not be superimposed to each other. This means that the sources are actually dependent, but, at the same time, this information provides us with a very powerful constraint for efficiently bounding the problem. Thus, in our method, to obtain separation, at each time t only one signal out of the four is allowed to be non-zero. For deconvolution, we enforce positivity and minimum energy of the signals. Indeed, these constraints used together are able to produce super-resolution, and then are very effective for the deconvolution of impulsive signals. With respect to the estimation of the mixing and diffusion operators, we considered generic constraints for both A and $\mathbf{h}$. In particular, the adopted prior for A constrains its elements to be positive, while $\mathbf{h}$ is modelled as a Gaussian function and bounds on its variance are used.

The joint MAP estimation of eq. 2 is usually approached by means of alternating componentwise maximization with respect to the three sets of variables in turn:

$$\hat{\mathbf{h}} = arg \max_{\mathbf{h}} P(\mathbf{x}|\mathbf{s}, A, \mathbf{h})P(\mathbf{h}) \quad (3)$$

$$\hat{A} = arg \max_{A} P(\mathbf{x}|\mathbf{s}, A, \mathbf{h})P(A) \quad (4)$$

$$\hat{\mathbf{s}} = arg \max_{\mathbf{s}} P(\mathbf{x}|\mathbf{s}, A, \mathbf{h})P(\mathbf{s}). \quad (5)$$

where the priors $P(\mathbf{h})$, $P(A)$ and $P(\mathbf{s})$ are chosen in such a way to probabilistically enforce the over-mentioned constraints. We solve the above scheme via a Simulated Annealing algorithm in A and $\mathbf{h}$, alternated with deterministic updates for $\mathbf{s}$, based on gradient ascent.

3 Experimental results

To quantitatively measure the performance of the proposed method, we carried out a number of experiments on synthetically generated DNA electropherograms. Two of such experiments are illustrated in Figures 1-3, for the noiseless and noisy cases, respectively. The data were generated by convolving four non superimposed impulse trains with Gaussian impulse responses, and then linearly mixing the four resulting signals. For each impulse train, the number of impulses, their locations and amplitudes, were chosen randomly, and the standard deviation of the corresponding impulse response was kept fixed along the sequence, in the assumption that diffusion can be considered stationary for short sequences.

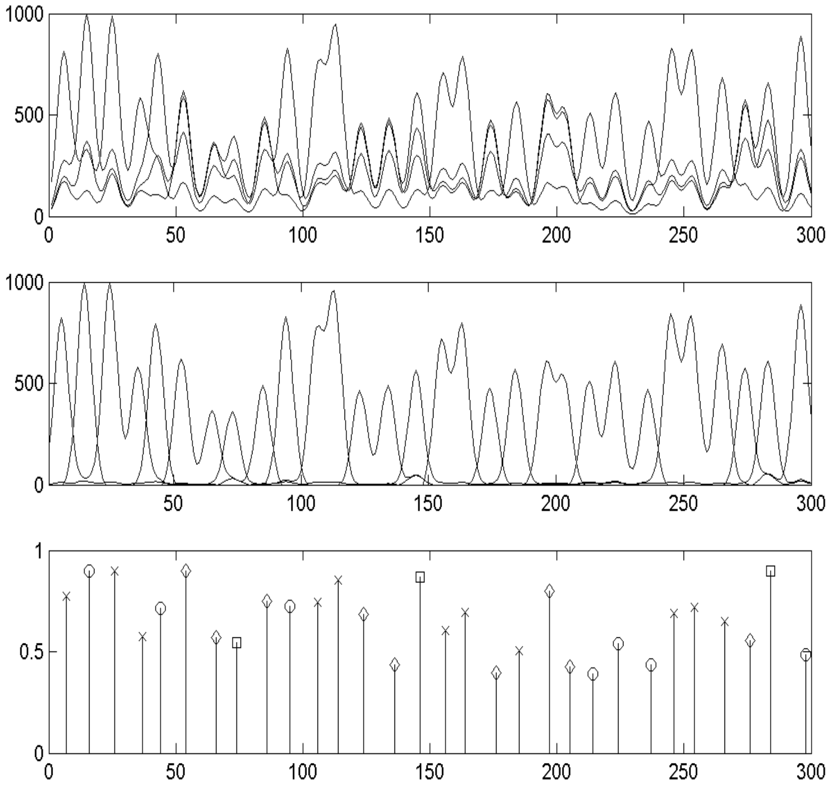


Fig. 1. Top: noiseless synthetic DNA sequencing data; Middle: color corrected data; Bottom: output from joint separation and deconvolution.

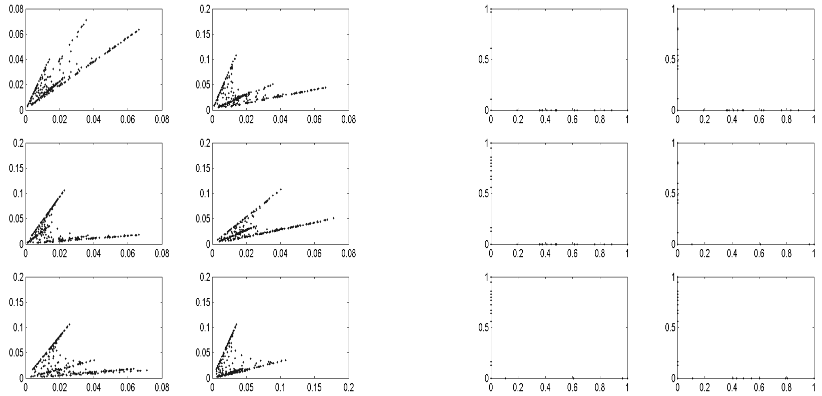


Fig. 2. Left: data scatterplot; Right: reconstruction scatterplot.

Figure 1 shows the results of the method in the noiseless case. In particular, the top panel shows the very bad quality electropherogram considered as data, the middle panel shows the intermediate result of blind color separation, and the bottom panel shows the final signals reconstructed after both color separation and blind deconvolution. Note, however, that the algorithm directly produces the final reconstruction starting from the data alone, and the intermediate color separation result has been obtained by multiplying the data with the inverse of the estimated mixing matrix. In this case, the final reconstruction exactly reproduces the positions of the original impulse trains considered, apart from scale factors in the amplitudes. The symbols marking the different impulses indicate the four kinds of DNA bases (A, T, C, G). The original mixing matrix adopted for generating the data was:

$$A_{true} = \begin{bmatrix} 1.0000 & 0.4976 & 0.1277 & 0.2129 \\ 0.9536 & 1.0000 & 0.3723 & 0.2415 \\ 0.6725 & 0.7184 & 1.0000 & 0.3345 \\ 0.2725 & 0.2136 & 0.3266 & 1.0000 \end{bmatrix}$$

while the estimated one was:

$$A_{est} = \begin{bmatrix} 1.0000 & 0.4675 & 0.1348 & 0.2185 \\ 0.9525 & 1.0000 & 0.3697 & 0.2422 \\ 0.6862 & 0.6683 & 1.0000 & 0.3453 \\ 0.2717 & 0.2693 & 0.3014 & 1.0000 \end{bmatrix}$$

For comparison purposes, the two matrices has been rescaled by dividing each column for its highest value. The mean square error between A_{true} and A_{est} was 0.0217. The standard deviations of the four Gaussian impulse responses were estimated up to an accuracy of 0.001. Figure 2 shows the scatterplots of the data (left panel) and of the reconstructed signals (right panel). While a high degree of correlation is present between each couple of data signals, the reconstructions are perfectly uncorrelated.

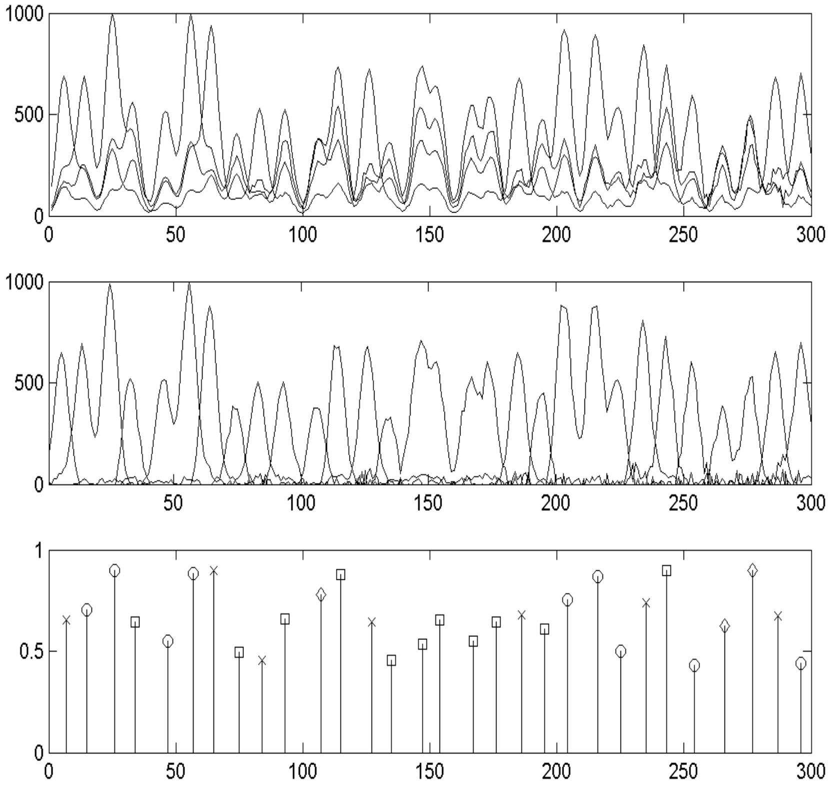


Fig. 3. Top: noisy synthetic DNA sequencing data; Middle: color corrected data; Bottom: output from joint separation and deconvolution.

In another experiment, shown in Figure 3, we added same noise to the convolved and mixed signals. This was a white, Gaussian process, with standard deviation slowly increasing with time, to simulate the baseline error. Also in this case the reconstructions of the signals, the mixing matrix and the impulse response standard deviations were more than satisfactory, showing that the method is robust enough even against non-stationary noise.

Other experiments were conducted on real data, for which results from automatic sequencing machines were available. In particular, we performed tests on several segments of the genome of a *Gymnochlora* sp. alga. With our method, we obtained some improvements with respect to the performance of the commercial basecallers, even for high quality electropherograms. Figure 3 shows the result obtained on a segment for which the reliability of the calls of the commercial basecaller was very low. For this segment, the highly reliable sequencing of the complementary strand was available. We could thus perform a biological validation of the results, based on an estimate of the true sequence, obtained for complementarity from the dual strand. In particular, we observed that the sequence provided by the commercial sequencer contained seven errors (mainly missing bases, i.e. deletions), while ours only three errors. Finally, Figure 4 shows a short sequence of satisfactory quality where, however, the software running on the commercial automated sequencing machine produced an error in the interval 100-150 where the sequence "ATA" was recognized. In fact, according to the other strand, considered reliable by the biologists, the middle "T" should instead be an "A". As shown in the bottom panel of Figure 4, in the same position our algorithm correctly recognized an "A".

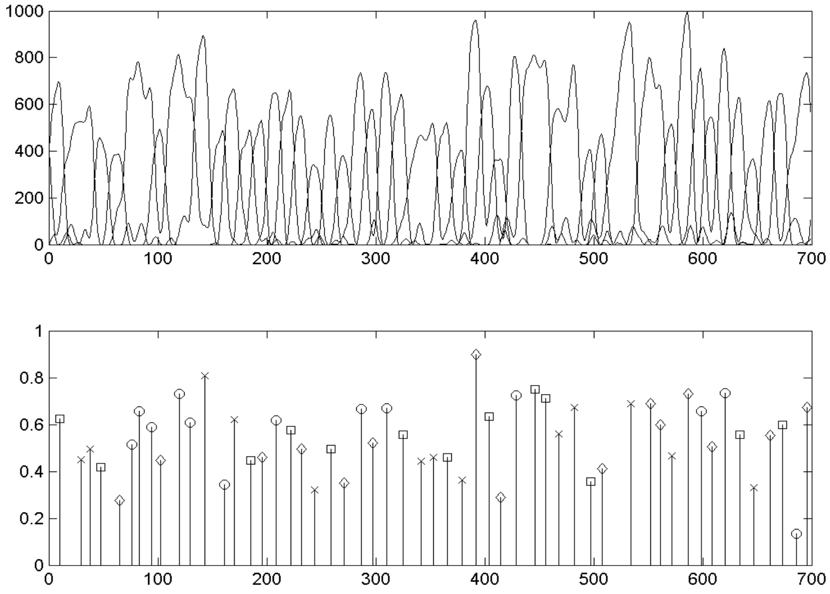


Fig. 4. Top: real DNA sequencing data; Bottom: output from joint separation and deconvolution.

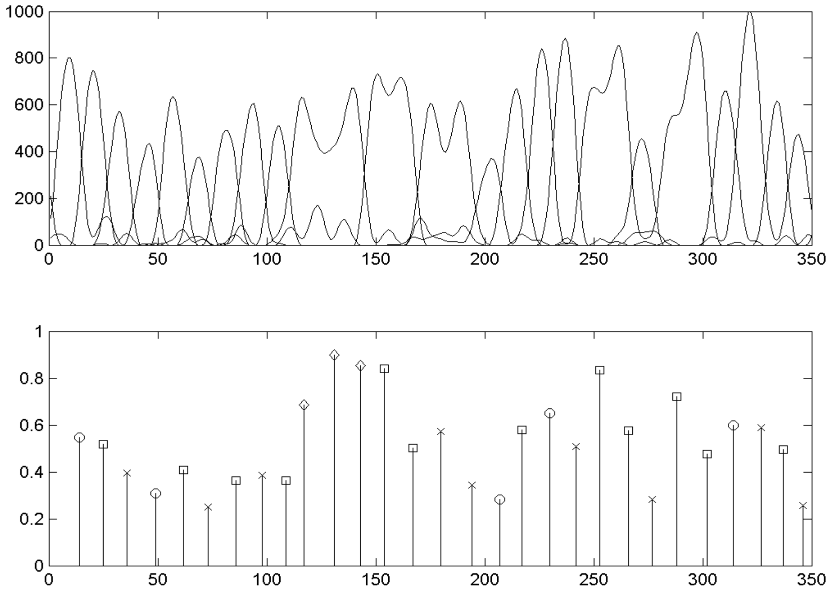


Fig. 5. Top: real DNA sequencing data; Bottom: output from joint separation and deconvolution.

4 Conclusions

We have proposed a method based on statistical models for the processing of electrophoretic time series produced in automated DNA sequencing, with the aim at removing typical degradations and improving basecalling. The degradations we considered are the most frequent ones, that is peak superposition (cross-talk) and peak merging (diffusion). We formulate the problem in a Bayesian estimation framework as one of joint blind source separation and blind deconvolution. In particular, the a priori information we exploited allows to obtain the restored electropherograms in an impulsive form which makes basecalling straightforward. Preliminary results on synthetically generated data and real DNA sequences showed the promising performance of the method, even against very bad or noisy electropherograms. In addition, the method is suitable for handling any kind of constraint. Further improvements could thus be obtained, for instance, by including constraints on the number of allowed bases and bounds on the spacing between bases. Another possible extension could consists in enforcing the bases to be complementary with those of the other strand, when, as it often happens, this is available as well.

5 Acknowledgements

We are grateful to Paolo Gualtieri and Anna Frassanito of the Biophysics Institute, CNR, Pisa for providing us with the data and assisting us in the biological validation of the results.

References

1. ABI. ABI Prism.: DNA sequencing analysis software, User's Manual (1996) Perkin Elmer Applied Biosystems, Foster City, CA .
2. Amari, S., and Cichocki, A.: Adaptive blind signal processing - neural network approaches. Proc. IEEE **86** (1998) 2026–2048.
3. Boufounos, P., El-Difrawy, S., and Ehrlich, D.: Basecalling using hidden Markov models. Journal of the Franklin Institute **341** (2004) 23–36.
4. Comon, P.: Independent Component Analysis, a new concept?. Signal Processing **36** (1994) 287–314.
5. Ewing, B., Hillier, L., Wendl, M., and Green, P.: Base-calling of automated sequencer traces using Phred. I, Accuracy assessment. Genome Res. **8** (1998) 175–185.
6. Ewing, B. and Green, P.: Base-calling of automated sequencer traces using Phred. II, Error probabilities. Genome Res. **8** (1998) 186–194.
7. Freschi, V. and Bugliolo, A.: Computer-aided DNA base calling from forward and riverse electropherograms. Trans. on Computat. Systems Biol. **LNBI 3737** (2005) 1–13.
8. Haan, N. M. and Godsill, S. J.: Modelling electropherogram data for DNA sequencing using variable dimension MCMC. Proc. Int. Conf. on Acoustics Speech and Signal Processing - ICASSP (Instanbul, Turkey, 2000) 3542–3545.
9. Hyvärinen, A.: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Trans. Neural Networks **10** (1999) 626–634.
10. Li, L. and Speed, T. P.: An estimate of the cross-talk matrix in four-dye fluorescence-based DNA sequencing. Electrophoresis **20** (2000) 1433–1442.
11. Li, L.: DNA sequencing and parametric deconvolution. Statistica Sinica **12** (2001) 179–202.
12. Pereira, M., Andrade, L., El-Difrawy, S. and Manolakos, E. S.: Statistical learning formulation of the DNA base-calling problem and its solution using a Bayesian EM framework. Discrete Appl. Math. **104** (2000) 229–258.

Automatic Segmentation of Unstained Living Cells in Bright-Field Microscope Images

Marko Tscherepanow, Frank Zöllner, and Franz Kummert

Applied Computer Science, Faculty of Technology,
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld
{marko, fzoellne, franz}@techfak.uni-bielefeld.de

Abstract. The automatic subcellular localisation of proteins in living cells is a critical step to determine their function. The evaluation of fluorescence images constitutes a common method of localising these proteins. For this, additional knowledge about the position of the considered cells within an image is required. In an automated system, it is advantageous to locate and segment these cells in bright-field microscope images taken in parallel with the fluorescence micrographs. Unfortunately, currently available cell segmentation methods are only of limited use within the context of protein localisation, since they frequently require microscopy techniques that enable images of higher contrast (e.g. phase contrast microscopy or additional dyes) or can merely be employed with too small magnifications. Therefore, this article introduces a novel approach for the robust automatic segmentation of unstained living cells in bright-field microscope images.

1 Introduction

The genomes of a variety of species have been decoded in recent years. Knowledge about genes enables the analysis of corresponding, frequently unknown proteins and their functions, e.g. in order to gain knowledge about cell processes or to develop efficient drugs. A common approach of determining the function of proteins is the examination of fluorescence microscope images [1–3], which is especially well-suited for the analysis of subcellular locations in intact cells.

In order to localise them, the considered proteins are tagged with a fluorescence dye, for instance with the green fluorescent protein (GFP) or one of its spectral variants [4]. Unfortunately, the surrounding cells themselves are almost invisible in these fluorescence images (see Fig. 1). Thus, additional information are required to associate fluorescent spots with specific cells. Commonly applied methods for the acquisition of these information consist in manual segmentation [1] or the usage of specially stained cells [2].

In contrast to that, our approach enables an automatic segmentation of *Spodoptera frugiperda* cells (Sf9) without applying additional dyes. A bright-field microscope image, taken in parallel with each fluorescence image, is used for the identification of cells (see Fig. 2) which constitute the basis for the analysis of the corresponding fluorescence image.

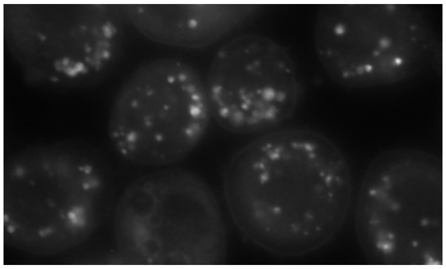


Fig. 1. Fluorescence micrograph showing Sf9 cells with stained lysosomes.

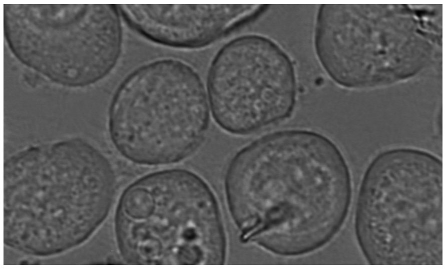


Fig. 2. Bright-field image taken simultaneously with the micrograph of Fig. 1.

The bright-field images are segmented by applying an active contour approach briefly outlined in [5]. After a discussion of relevant literature (see Section 2), this technique as well as required methods for the automatic determination of initial segments are described in Section 3. Section 4 proposes various enhancements that are relevant for a practical application of our approach. These methods are evaluated in Section 5. Eventually, conclusions are drawn and a short outlook is given in Section 6.

2 Related Work

Numerous articles have been published regarding automatic cell segmentation. Unfortunately, the subcellular localisation of proteins in living cells imposes special limitations that prohibit the application of most of these methods.

A large number of approaches such as [6–8] employs phase contrast microscopy in order to improve the contrast of acquired images. This technique requires special objectives that reduce the amplitude of incident light. As we take a fluorescence image in parallel, the light of fluorescent objects would be attenuated, as well. An alternation of the objective between the acquisition of the images causes further problems, since it modifies the optical path. Consequently, an association of corresponding pixels of both images would be hampered. Other approaches require special dyes [2, 9, 10] that are frequently applied to dead cells or tissues. If they were utilised with living cells they might interfere with examined proteins or even kill the cells.

Unlike phase contrast microscopy and the application of dyes, bright-field microscopy can be applied to the segmentation of cells in the context of protein localisation without adversely influencing the outcome of the investigation. Unfortunately, bright-field images are usually of low contrast, intensity-variant, and unevenly illuminated [11]. Furthermore, they show a great variety of cell appearances [12]. In addition, a relatively high magnification (60×, 1 μm equals 6.45 pixels) is necessary, as subcellular structures are to be localised. So, the considered Sf9 cells comprise between 10,000 and 80,000 pixels. Therefore, methods utilising small search windows in order to detect whole cells (cf. [9, 12]) cannot be employed.

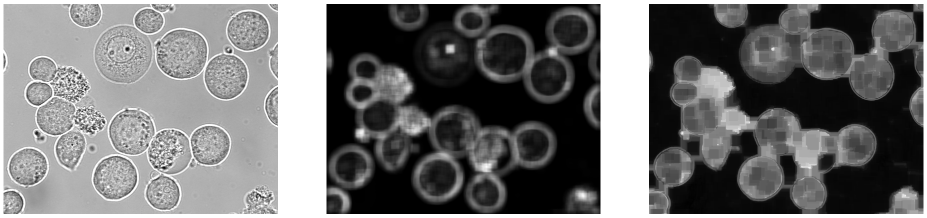


Fig. 3. Local intensity variation in a bright-field image (left). The result of the self-complementary top-hat (right) allows a noticeably better recognition of the image foreground than the variance map (centre) using a neighbourhood of 41 × 41 pixels.

As cells in bright-field microscope images are separated from other cells and their surrounding by their membrane, it is beneficial to include this information into the segmentation procedure. Parametric active contours or rather snakes have proven advantageous for that purpose [7, 8]. So, we have developed a snake algorithm for the robust segmentation of Sf9 cells in bright-field microscope images as well as methods for its automatic initialisation which are described in the following section.

3 Cell Segmentation in Bright-Field Images

Before actually segmenting cells in bright-field images, extensive preprocessing is essential to obtain a sufficient initialisation. Firstly, a separation between image areas containing cells (foreground) and other regions (background) occurs (see Section 3.1). Secondly, probable cell membrane pixels are detected (see Section 3.2) so as to enable a separation of neighbouring cells. Afterwards, cell markers, i.e. small regions within possible cells are determined (see Section 3.3) which are subsequently applied as initialisation of the segmentation procedure (see Section 3.4).

3.1 Separation of Image Foreground and Background

Wu et al. have shown that the local intensity variation is a valuable feature for the separation of foreground and background in bright-field images [11]. Instead of computing the local variation defined by the variance within a square neighbourhood, we take advantage of a morphological operator: the self-complementary top-hat [13]. Figure 3 depicts the result of the application of the self-complementary top-hat to an exemplary bright-field image as well as the corresponding variance map if a square neighbourhood of 41 × 41 pixels (suggested by Wu et al.) is considered. In order to increase the computational efficiency we employ structuring elements with a size of 25 × 25 pixels.

The bimodal distribution of the local intensity variations determined by the self-complementary top-hat is considerably more distinctive than the one computed by analysing the variance. Hence, the automatic separation of image foreground and background, which is performed by minimum error thresholding [14], is alleviated.

3.2 Detection of Probable Cell Membrane Pixels

Probable cell membrane pixels are determined by utilising morphological operators, as well. As the cell membrane possesses a linear shape that is less curved than other cell compartments and is characterised by a substantial change of pixel intensities, linear structuring elements are applied to the gradient magnitude image. All image structures that cannot contain this linear structuring element such as dirt, noise, and intracellular objects are removed by a morphological opening. In order to get closed contours, this operation is repeated for seven additional orientations. The resulting images are fused by computing the point-wise maximum, which constitutes an algebraic opening [13].

The length l of the linear structuring elements is crucial to the result of the algebraic opening. If it is chosen too small, irrelevant image structures will remain; if the value is too high, cell membrane pixels will disappear. Hence, we have developed an automatic procedure for the determination of an optimal value (see Section 4.1).

3.3 Determination of Cell Markers

On the basis of the computed image background and cell membrane pixels small regions within probable cells are identified – the cell markers (see Fig. 4). It is assumed that points possessing a great distance to the image background and membrane pixels lie inside cells. These points are determined by computing the local maxima of the distance transform [13].

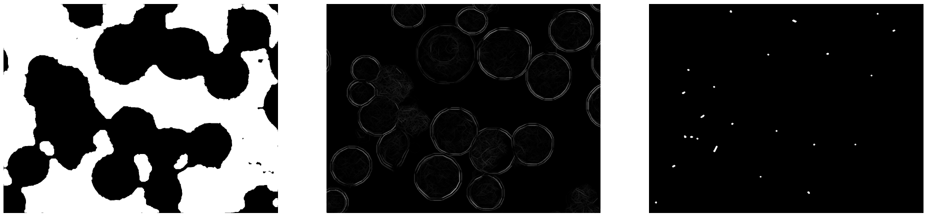


Fig. 4. Computation of cell markers. The cell markers (right) are determined in such a way that they maximise the distance to the image background (left) and membrane pixels (centre).

In order to obtain an appropriate initialisation for the segmentation step, these regions are dilated by a small circular structuring element (diameter: 5% of the maximal cell radius, 9 pixels). Afterwards, the contours are traced so as to obtain a polygonal representation that comprises only the start and end point of adjoining lines.

3.4 Cell Segmentation by Active Contours

Active contours have several advantages with respect to the segmentation of cells. Firstly, they always yield closed contours even if the corresponding cell membrane is hardly visible. Secondly, they enable the inclusion of context specific knowledge such as membrane curvature and cell size. So, the robustness can be improved.

Several approaches have been proposed for the computation of active contours, e.g. variational calculus, dynamic programming, and greedy methods. We have decided to apply a greedy approach [15] due to its efficient computability, stability, and flexibility. Since our approach aims at complete independence from user interactions while processing images, special requirements have to be fulfilled. In particular, the determined cell markers instead of close approximations of the resulting contour should be applicable as initialisations.

Cohen [16] proposed a method to realise the growth of snakes by introducing an inflation force. This technique applies normal vectors of the contour in order to determine the direction of extension. As a result, the contour might overlap if it is initialised with a concave cell marker. Hence, we have decided to utilise an alternative basis for the growth of the contour – the minimal distance to the initial contour. Equation (1) shows the corresponding energy functional E^*_{snake} of a parametric curve $v\big(x(s), y(s)\big)$ with arc length s.

$$E^*_{\mathrm{snake}} = \int_0^1 \big[\alpha E_{\mathrm{cont}} + \beta E_{\mathrm{curv}} + \gamma(E_{\mathrm{dist}})E_{\mathrm{ao}} + \delta(E_{\mathrm{dist}})E_{\mathrm{dist}}\big]\,ds \tag{1}$$

E_{cont} and E_{curv} control the continuity and curvature, respectively. Moreover, E_{cont} fosters equal spacing between points [15]. E_{ao} represents the resulting image of the algebraic opening (see Section 3.2) and E_{dist} the distance from the initial contour. As the energies are minimised, the image as well as the distance have to be inverted. Thus, a maximal considered distance $\varDelta_{\mathrm{max}}$ is required. We have set it to the maximal cell radius increased by a tolerance interval of 20% (198 pixels in total).

The parameters α, β, γ, and δ control the influence of the respective energy terms. At this, γ and δ are modified dependent on E_{dist}.

$$\gamma(E_{\mathrm{dist}}) = \gamma_0 \cdot \frac{\varDelta_{\mathrm{max}} - E_{\mathrm{dist}}}{\varDelta_{\mathrm{max}}} \tag{2}$$

$$\delta(E_{\mathrm{dist}}) = \delta_0 + \gamma_0 - \gamma(E_{\mathrm{dist}}) \tag{3}$$

According to (2), $\gamma(E_{\mathrm{dist}})$ yields high values if E_{dist} is small, i.e. if the snake has a great distance to its initialisation. By this, high pixel values near the cell markers, within the cells are suppressed. Equation (3) ensures that the sum of $\gamma(E_{\mathrm{dist}})$ and $\delta(E_{\mathrm{dist}})$ equals the sum of its base values γ_0 and δ_0, respectively. So, the extending force is reduced if the snake reaches a distance from its cell marker where the probability of membrane pixels is high. Additionally, background pixels receive a high value of E_{dist} in order to avoid an extension of the snake in this region.

4 Enhancements

In the introduction of our segmentation approach in Section 3 several questions were left open although they are crucial for the correct function. They have been topics of current research and are answered in the following. Section 4.1 outlines a method that enables the automatic determination of the optimal length l for the linear structuring elements which are applied during the algebraic opening. A further problem consists in the parametrisation of the snakes. As they are growing, new points have to be inserted (see Section 4.2).

4.1 Improvement of the Detection of Membrane Pixels

The basis for the automatic determination of the length l of the linear structuring elements is provided by $n = 499$ cell masks manually extracted by biological experts. Besides the mask of a cell i itself, the points of a tube with a diameter of 5% of the mean cell radius that is centred at the mask boundary are considered in order to detect the intensities of membrane pixels. The sets of the corresponding points p are denoted by $\mathcal{M}_i$ (mask) and $\mathcal{T}_i$ (tube), respectively. According to (6), then an optimal value l for the length of the line elements is computed by iterating over all possible values up to $\Delta_{\max}$.

$$I_l^{\mathcal{T}} = \sum_{i=1}^{n} \sum_{\forall p \in \mathcal{T}_i} I_l\big(x_p, y_p\big)^2 \tag{4}$$

$$I_l^{\mathcal{M}} = \sum_{i=1}^{n} \sum_{\forall p \in \mathcal{M}_i} I_l\big(x_p, y_p\big)^2 \Delta\big(x_p, y_p\big) \tag{5}$$

$$l_{\text{opt}} = \arg\max_{\forall l} \left(\frac{I_l^{\mathcal{T}}}{\max\limits_{\forall l} I_l^{\mathcal{T}}} - \frac{I_l^{\mathcal{M}}}{\max\limits_{\forall l} I_l^{\mathcal{M}}} \right) \tag{6}$$

$I_l\big(x_p, y_p\big)$ constitutes the image generated by an algebraic opening with a structuring element of length l. The consideration of squared pixel values results in a reduced influence of small intensities that have less negative effects on the segmentation than high ones. Moreover, the points of the mask image are weighted by their minimal distance $\Delta\big(x_p, y_p\big)$ to the boundary. l_{opt} is optimal in a sense that it maximises the difference of the intensities (scaled to fit into the interval $[0, 1]$) within both examined image regions in order to enhance the contrast.

4.2 Insertion of New Points

The segmentation consists in the extension of snakes starting from small regions within probable cells. So, the distances between adjoining points are increased and resampling of the snake, i.e. the insertion of new points is necessary. On the

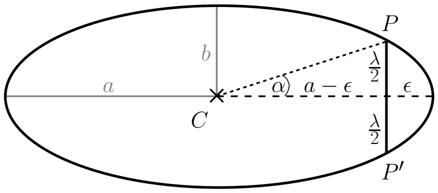


Fig. 5. Approximation of an ellipse by line segments. A line segment of length λ connecting the points P and P' causes an approximation error ϵ if it is equally divided by the major axis. As the distance between the ellipse and its centre C is maximal there, ϵ is maximal, as well. Thus, ϵ constitutes the worst case value.

other hand, too high a number of points results in an increased computational effort. Thus, some kind of compromise has to be reached. Since Sf9 cells are almost elliptically shaped, an ellipse approximation of the current snake is performed [17]. This yields the lengths of the semiminor axis b and of the semimajor axis a as well as the centre C. On the basis of these values the approximation error ϵ occurring if the ellipse is approximated by a line segment of length λ is computed (see Fig. 5).

An ellipse can be described by $x = a \cdot \cos\alpha$ and $y = b \cdot \sin\alpha$. Inserting the coordinates $x_P = a - \epsilon$ and $y_P = \frac{\lambda}{2}$ of point P and fusing the results leads to (7) which enables the determination of λ.

$$\lambda = 2b \cdot \sin\left(\arccos\frac{a-\epsilon}{a}\right) \tag{7}$$

Instead of computing the ellipse approximation after every iteration step of the snake algorithm (variable split length, VSL), it can be applied to the determination of a constant split length λ^* (CSL). For this purpose, all manually extracted cells are approximated by an ellipse and λ^* is set to the minimal value of λ. So, a correct approximation of all cells with an error less than ϵ can be guaranteed, as well.

5 Results

We evaluated our methods on a dataset containing 499 cells manually extracted from 45 images by biological experts. In order to enable investigations regarding different foci, the dataset comprises images of the same specimen at three manually adjusted focal planes (A, B, and C) showing the cell characteristics depicted in Fig. 6. All 499 manually extracted cells were automatically marked during the preprocessing step (see Section 3.3) and each cell mask was associated with the marker closest to its centre. Furthermore, the length of the linear structuring elements for the algebraic opening was automatically set to $l_{\text{opt}} = 31$ according to Section 4.1.

In order to assess the segmentation, the manually extracted cell masks were compared with the corresponding automatically segmented cells by performing a

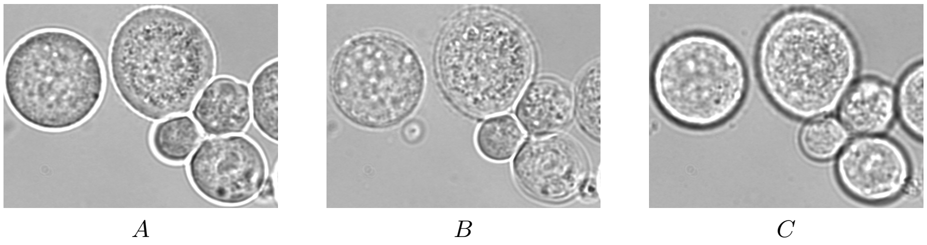


Fig. 6. Cells at different focal planes. The appearance of the examined cells varies if the focus is modified.

15-fold cross test. The energy weights were chosen in such a way as to minimise the error term $\bar{d}^{\text{err}}$ for all except one of the images of a focal plane (see (8)).

$$\bar{d}^{\text{err}} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i^{\max}}{b_i} \tag{8}$$

$d_i^{\max}$ denotes the maximal distance of corresponding manually and automatically determined contours of a cell i. These values are normed to the current manually determined cell size represented by the length of the semiminor axis b_i of the cell's approximation by an ellipse.

After computing the energy weights, the remaining image was segmented in order to measure the test errors. $\bar{d}_A^{\text{test}}$, $\bar{d}_B^{\text{test}}$, and $\bar{d}_C^{\text{test}}$ denote the mean of these test errors over all images (see Tab. 1). Additionally, the mean point number per snake $\bar{p}$ and the average processing time[1] per image $\bar{t}$ on an AMD Athlon 64 CPU (2GHz) were determined.

Table 1. Comparison of the segmentation if variable split length (VSL), constant split length (CSL), and no resampling are applied. The dash denotes parameters that were not available.

method	ϵ	λ^*	$\bar{d}_A^{\text{test}}$	$\bar{d}_B^{\text{test}}$	$\bar{d}_C^{\text{test}}$	$\bar{p}$	$\bar{t}$
VSL	0.5	–	0.104	0.118	0.142	33.9	1.038s
	0.125	–	0.088	0.109	0.139	59.2	1.200s
CSL	0.5	18	0.094	0.130	0.143	45.2	0.802s
	0.125	9	0.102	0.116	0.141	89.6	0.980s
no resampling	–	–	0.109	0.123	0.146	23.4	0.708s

The results of all methods show that the choice of the focal plane has a considerable effect on the quality of the segmentation. The errors rise from plane A to plane C. These results originate in less distinctive cell membranes (B) and stronger intracellular intensity variations (C), respectively (cf. Fig. 6)

Both reparametrisation methods attained smaller segmentation errors than the original approach which does not perform resampling. Since CSL utilises

[1] excluding the time for the computation of the cell markers

a minimal value of the split length λ that is sufficient for all cells, it requires additional points in comparison to VSL. These unnecessary additional points seem to deteriorate the segmentation compared to VSL (e.g. for $\epsilon = 0.125$). The lowest errors were reached by VSL with $\epsilon = 0.125$ which required significantly more processing time than the other methods because of the determination of λ during the actual segmentation. So, if enough time is available VSL should be employed. Otherwise, the original approach and CSL, especially with $\epsilon = 0.5$, are beneficial.

In order to assess our results, the manually extracted segments of 363 cells determined by five persons were compared pairwisely. The corresponding contours possess a mean maximal distance of 5% of the cell size with a standard deviation of 2.5%, as the cell membranes cannot always be determined unambiguously. Thus, we conclude that our methods achieve a very high accuracy.

6 Conclusion

We have presented an approach for the automatic segmentation of cells in bright-field microscope images. Furthermore, several enhancements with respect to the quality of the preprocessing as well as the segmentation have been introduced. The evaluation occurred on images at three different focal planes in order to enable the choice of an optimal one. At this focal plane, all methods yielded excellent results insofar as the segmentation error is only slightly higher (difference $< 10\%$) than the deviations of segments manually determined by different persons. Figure 7 depicts the segmentation of a bright-field image if VSL with $\epsilon = 0.125$ is applied.

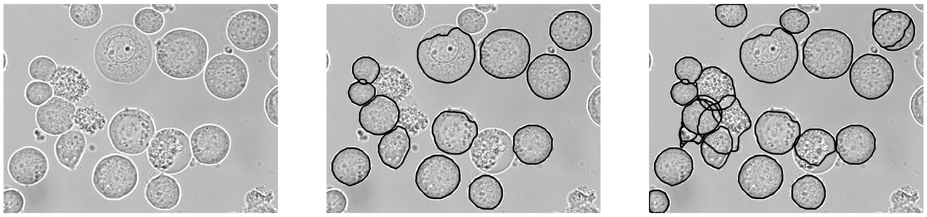


Fig. 7. Segmentation of a bright-field image (left). The segments are depicted as black contours that are equal to the final snakes. The central image shows only segments, that could be associated with manually extracted cell masks whereas the right picture comprises all snakes.

Since the segmentation yields contours that represent dead cells, dirt and other image structures (see Fig. 7), as well, a classification of the segments has to be performed [18]. On the basis of the recognised cells, our research will now be directed towards the analysis of the corresponding fluorescence images.

References

1. Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O'Shea, E.K.: Global analysis of protein localization in budding yeast. Nature **425** (2003) 686–691
2. Liebel, U., Starkuviene, V., Erfle, H., Simpson, J.C., Poustka, A., Wiemann, S., Pepperkok, R.: A microscope-based screening platform for large-scale functional protein analysis in intact cells. FEBS Letters **554** (2003) 394–398
3. Murphy, R.F., Velliste, M., Porreca, G.: Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. Journal of VLSI Signal Processing **35** (2003) 311–321
4. Tsien, R.Y.: The green fluorescent protein. Annual Review of Biochemistry **67** (1998) 509–544
5. Tscherepanow, M., Zöllner, F., Kummert, F.: Aktive Konturen für die robuste Lokalisation von Zellen. In: Bildverarbeitung für die Medizin. (2005) 375–379
6. Debeir, O., Ham, P.V., Kiss, R., Decaestecker, C.: Tracking of migrating cells under phase-contrast video microscopy with combined mean-shift processes. IEEE Transactions on Medical Imaging **24** (2005) 697–711
7. Ray, N., Acton, S.T., Ley, K.: Tracking leukocytes in vivo with shape and size constrained active contours. IEEE Transactions on Medical Imaging **21** (2002) 1222–1235
8. Zimmer, C., Labruyère, E., Meas-Yedid, V., Guillén, N., Olivo-Marin, J.C.: Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: A tool for cell-based drug testing. IEEE Transactions on Medical Imaging **21** (2002) 1212–1221
9. Nattkemper, T.W., Wersing, H., Ritter, H., Schubert, W.: A neural network architecture for automatic segmentation of fluorescence micrographs. Neurocomputing **48** (2002) 357–367
10. Reddy, C.K., Dazzo, F.B.: Computer-assisted segmentation of bacteria in color micrographs. Microscopy and Analysis **91** (2004) 17–19 European edition.
11. Wu, K., Gauthier, D., Levine, M.: Live cell image segmentation. IEEE Transactions on Biomedical Engineering **42** (1995) 1–12
12. Long, X., Cleveland, W.L., Yao, Y.L.: Automatic detection of unstained viable cells in bright field images using a support vector machine with an improved training procedure. Computers in Biology and Medicine **6** (2006) 339–362
13. Soille, P.: Morphological Image Analysis. Springer (2003)
14. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recognition **19** (1986) 41–47
15. Williams, D.J., Shah, M.: A fast algorithm for active contours and curvature estimation. Computer Vision, Graphics, and Image Processing: Image Understanding **55** (1992) 14–26
16. Cohen, L.D.: Note: On active contour models and balloons. Computer Vision, Graphics, and Image Processing: Image Understanding **53** (1991) 211–218
17. Fitzgibbon, A.W., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. IEEE Trans. Pattern Anal. Mach. Intell. **21** (1999) 476–480
18. Tscherepanow, M., Zöllner, F., Kummert, F.: Classification of segmented regions in brightfield microscope images. In: Proceedings of the International Conference on Pattern Recognition (ICPR). (2006) to appear.

ibai Institute of Computer Vision
and Applied Computer Sciences
Dr. Petra Perner

Institute of Computer Vision and Applied Computer Sciences IBaI

Director: Dr. Petra Perner

Address: Körnerstr. 10
04107 Leipzig
Germany

Phone: +49 341 8612273
FAX: +49 341 8612275
E-Mail: info@ibai-institut.de

Personal Homepage:
www.ibai-research.de

Institute`s Homepage:
www.ibai-institut.de

International Conference on Data Mining and Machine Learning MLDM
www.mldm.de

Industrial Conference on Data Mining ICDM
www.data-mining-forum.de

BioMedVision Center
www.biomedvision.de

Data Minng Tutorial
www.data-mining-tutorial.de

dmls2006

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □□ □ □

Workshop on Data Mining in Life Science

Petra Perner (Ed.)

Workshop Proceedings

IBaI CD-Report ISSN 1617-2671

Published July, 2006

www.data-mining-forum.de

Organized by:

Institute of Computer Vision
and Applied Computer Sciences
Dr. Petra Perner

Preface

The availability of growing amounts of data and multimedia documents in biological, medical, and natural sciences has motivated a strong research interest in novel methods and systems for automatically extracting and synthesizing previously unknown and interesting knowledge. This is also the motivation of this workshop to focus on applications of data mining spanning the full range of natural and biomedical sciences domains.

Following the explosion of interest in natural and biomedical sciences applications of data mining, this workshop on data mining in the Life Sciences is the first of its kind to be held at ICDM-06, in Leipzig, Germany.

Six papers are to be presented at this First Workshop on Data Mining in the Life Sciences. These papers represent the research and experience of authors working in five different countries on a wide range of problems and projects. They illustrate some of the major trends of current research in Data Mining in the Life Sciences.

The first paper focuses on *biological knowledge extraction, integration, and dissemination* by making available to the scientific community an integrated and comprehensive knowledge-base for biomolecular modeling, to which data mining actively contributes. The second paper surveys *biological sequences mining*, which is a very salient area of research given the growing amount of biological sequences available. The paper also provides several innovative ways of mining for invariants within these sequences. The third paper tackles the problem of *discovering higher-level knowledge structures from biomedical literature* in the form of prototypical cases that can be added to the knowledge-base of a case-based reasoning system for decision-support assistance for example. It builds on recent advances in text mining from literature, especially concept mining and relationship mining. The fourth paper proposes to *mine images for specific parameters in food*, and deals with image mining for intelligent interpretation. The system mines for hygiene-relevant parameters from images of grains of cereal crops. The fifth paper deals with *colony image segmentation*, and presents a novel algorithm particularly suitable for this kind of image segmentation. The algorithm is based on kernel spatial fuzzy c-means. The last paper also proposes an image mining algorithm for the *segmentation of densely packed rock fragments* with uneven illumination.

Overall, these papers represent an excellent sample of the most recent advances of data mining in the Life Sciences, and promise very interesting discussions and interaction between the major researchers in this niche of data mining research.

June 2006 *Isabelle Bichindaritz*

Citation

Petra Perner (Ed.)
Workshop on Data Mining in Life Science, DMLS2006, Workshop Proceedings,
IBaI CD- Report, ISSN 1617-2671, July 2006

Example Paper:

Gabor Bereczki and Masaru Tomita
An interaction knowledgebase for biomolecular modelling In Workshop Proceedings: Petra Perner (Ed.), Workshop on Data Mining in Life Science, DMLS2006
IBaI CD- Report, ISSN 1617-2671, July 2006, p. 1-13.

Table of contents

An interaction knowledgebase for biomolecular modeling

Gabor Bereczki and Masaru Tomita

Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan

Abstract. Many modeling problems in cell simulation require access to information provided by various public biological databases. Providing unified access to these databases as well as integrating them with model building software presents a standing challenge for designers of modeling systems. We have set up a knowledgebase integrating several of the most influential public biomolecular databases, a web services front end for browsing and searching the unified data warehouse, SOAP RPC services for application interface to the data and a Java model editing client for graphical model editing.

1 Introduction

The problem of modeling whole cells, the fundamental building blocks of life, poses a significant challenge, ever since its emergence little before the turn of the millennium. The problem set involves modeling metabolic reactions, gene expression, translation, transcription, signal transduction, polymerization, 3D scaffold formation, et cetera.

E-cell is a research project at the Institute for Advanced Biosciences, Keio University, Japan aimed at the modeling of a whole living cell. The project was started up by Masaru Tomita and Koichi Takahashi in 1997 by creating E-cell1[1], an ODE solver simulation engine for metabolic and gene expression processes. In 2003 E-cell3[2] was released, an improved simulation environment, with capabilities to integrate different algorithms .

As part of the project, the E-cell 3 Modeling Environment was set up with the aim of helping model creation utilizing a graphical user interface. The ultimate objective of a modeling environment is to facilitate the research cycle by eliminating the bottleneck of reconstructing biochemical networks [3].

Recent trends in the field suggest that the focus in modeling has shifted to the small scale experimental work where models comprising of a several reactions and substances are intensively investigated through experimental work and the results are published as models. Large scale models can then be built through a combination of these small models. Hence there is a need for software tools that provide:

- An interface to various public databases containing information related to biochemical reactions.
- A simple and concise user interface for drawing up simple models from databases.
- Methods to recombine, test, debug and estimate parameters of the various small scale models.

The first requirement can be satisfied with the newly developed knowledge base that integrates various databases, the second requirement by the web services interface together with the Java editing tool now under development and the third requirement by the traditional desktop application Model Editor which is part of the E-cell 3 software package.

According to Galperin [4] there are over 850 public databases as of October 15 2005. Of great importance to the modeling community are those databases that contain molecular interactions and background information on important molecular biology entities such as genes and proteins. Thus databases of primary interest are those that contain information on:

- metabolomic interactions
- protein-protein interactions
- transcriptional regulatory interactions
- gene expression information
- genes and nucleotide sequence catalogs
- protein catalogs
- biochemical molecule catalogs

A closer examination at the databases of this subset of molecular phenomena suggests that many of the existing hundreds of databases are small and redundant. The most significant databases are introduced in Figure 1.

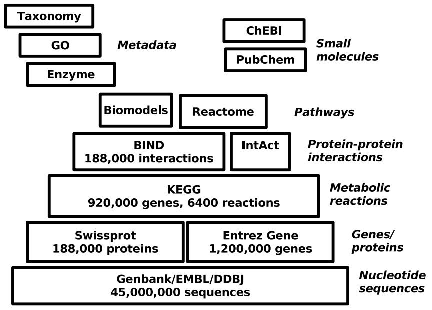


Fig. 1. Several of the most significant databases for systems biology. Record counts are taken from own research at the end of summer 2005.

Similar project in the field usually fall into two categories (i) federated query systems, such as Biomart[10] or BioMediator[12] and (ii) datawarehouses most notably BioSilico[11] and and Biozon[13] . GEM System [5] developed at our instute aims at genome wide reconstruction of metabolic pathways. Compared to these works our proposed knowledgebase presents a datawarehouse system with a conceptual model compatible to SBML[14] built upon a simple and flexible relational data model. The datawarehouse as well as the front end focuses on creation of models instead of mere visualization of pathway data. With the currently released version of the user interface, SBML models can be created in an interactive and semiautomated way utilizing user defined rules to infer stoichiometry models which can be later extended with mass balance differential equations and parameters. Both ODE solvers and parameter estimating tools are provided by the E-cell 3 project.

2 Methods

To achieve the goals laid down in the preceding section, the following tasks must be completed:

- obtaining the data from public databases
- parsing the raw data into our data warehouse
- integrating the data from different databases
- storing the data
- distributing the integrated data over a network
- creating client tools to utilize the data

To be able to integrate the different -omics concepts, the different physical data representations of various databases as well as allowing an incremental, database by database build-up, maximum flexibility and extendability is required. A dynamic data model is planned henceforth. Facing an enormous quantity of data from heterogeneous sources it is advisable to retrieve high quality data only, however not restrict the scope of the stored data by any other criteria (such as organism or biological process) for the sake of versatility. For example in the case of Genbank sequences, only the non redundant set of sequences is processed excluding the majority of data contained i.e. ESTs, STSs and the similar high throughput - low annotation entities. Given the scarcity of resources and to retain scientific integrity no content processing or change in the data is allowed only curated resources are used, this implies that new interactions are not inferred either by own computational or experimental research. The data parsing should focus on maximum recovery of cross references and interactions. The data warehouse is a repository of many of the attributes that an entity has in the original database, but not all attributes. Name, Identifier and Description class attributes are mainly parsed, as they serve the purpose of cross referencing and text search. Data provenance is taken care of by the data warehouse keeping track of the origin of every single record even after entities are merged. To maintain the integrity of the data a very simple update uproach is utilized: building the database from scratch. This approach improves robustness by removing complexity from the implementation and the extra computational costs are only in the repeated parsing as the entity integration and relationship origination algorithms must run on the whole database.

The data model defines the representation of knowledge in our data warehouse. The representation of knowledge is carried out in the form of an attributed, undirected graph, where nodes are *entities* and edges are *relationships*.

Logical data model: The four major modeling objects are: *entities*, *attributes*, *relationships*, *rules*. An entity can be a gene, protein, reaction or any kind of biomolecular phenomena (entity class) that databases or models may contain. Entities belonging to the same class can be merged without loss of information if they have matching globally unique identifiers. The knowledge base only contains strong (e.g. gene, protein) and associative (e.g reaction, pathway) entities. Weak entities (e.g.

publications) are processed as attributes in order to reduce data model complexity. Attributes are any properties that the biomolecular object represented by the entity may possess. Attributes are categorized as cross references (used for merging entities, creating relationships), names (for synonym search), free text fields (for text search), or other information (e.g. formulas, parameters). Relationships describe any means by which two entities can be connected. Example for relationships are part of, generalization, type of, substrate of, product of. Relationships are generated from entities and attributes using rules and are therefore redundant. Relationships provide the backbone of the meta interaction network and the means of navigation in the database. Rules govern data processing steps other than parsing. Simple rules are created for semantic labeling, merging of entities, and creation of relationships. The conceptual logical data model is not fixed, but basically aimed at SBML[14] level 2 compatibility. Thus entities should conform to SBML components wherever it is possible. Metadata, such as GO terms and taxonomy data are also processed in the entity-relationship form and referenced to biomolecular data.

Physical data model: There are two databases in the knowledgebase;

- *temp;* for preprocessing tables for raw parsed data
- *data;* for storage of entities, attributes , relationships and rules indexed for easy navigation, search and data mining

The knowledge base issues an artificial unique identifier for every entity, attribute and relationship and connects those tables using these IDs. Entities, relationships and attributes have types, which are defined in the type table. The main advantage of this approach is that (i) different entities from different data sources can be easily incorporated, as there is no need for a separate table definition for each new entity type and (ii) multi value attributes are easily resolved without the need for introducing new entities in a hierarchical fashion.

The files are downloaded from the database servers using the ftp protocol and then parsed into preprocessing tables in the *temp* area. All databases under integration have their set of preprocessing tables. Every preprocessing table represents a different type of entity. During parsing, multi-value attributes are normalized into multiple rows. The columns of a preprocessing table are similar to those of the attribute table, with the difference that attribute types are parsed as is from the source database therefore not globally unique. During load the preprocessed data is transferred from the *temp* database to the *data* database. Attributes are semantically labeled, that is, every attribute is assigned an attribute type code by using a mapping table. Artificially unique identifiers of entities and attribute rows are also assigned in the load phase. Semantic labeling of properties happens by rules that specifies the label that should be given to a specific attribute in a specific database. The most important goal for semantic labeling is to find common identifiers and names in different databases and label them with the same attribute type ID for purposes of cross-referencing. Most frequent common identifiers are: EC number, GI ID, Gene ID, Uniprot ID and various KEGG IDs. In current implementation there are 933 labeling rules for 33 preprocessing tables (data sources).

Integration of databases practically means merging entities of the same type from different (or the same) database source. Because attributes are made uniform during the semantic labeling procedure merging happens without regard to the original source of the data. Integration happens in 2 steps:

Merging of entities is performed by rules stored in the merge_rules table. If the values of a certain type of identifier match in different instances of a certain type of entity, the entity instances will be merged. Merging actually means overwriting the entity ID of the attributes belonging to the other (to be merged) entities with the surviving entity ID. The surviving entity ID is chosen randomly from the matching ones. Merge by attributes is implemented in SQL stored procedures and is heavily optimized, because there are millions of rows that needs to be compared and can be computationally very expensive. There are 9 merge rules in the current implementation to integrate entities by identifier matching.

Establishing relationships. Relationships (such as type of, part of, participates, specialization of etc.) are determined by another rule table: relationship_rules. A relationship between two entities can be set up if they belong to a certain type and the value of attribute1 of entity1 matches attribute2 of entity2. This problem is thus very similar to that of merging by attributes and is implemented in a similar fashion. There are 45 relationship rules present into the data warehouse.

Integrated data should be made available to users in many different forms. Since it is not recommended to expose the underlying database tables to users directly hence a communication layer was established to function as an interface between the user and the knowledgebase. HTTP was chosen as the protocol of choice for communication between the data warehouse and the client side due to its is universality and negligible security challenges. XML was chosen as the media of communication as it is widely used for data exchange purposes and both HTML and SOAP RPC is built upon XML technology.

The client side consists of a web services interface that was designed with simplicity in mind and was inspired by the Google philosophy in order to provide the core access to the knowledgebase. All the attribute values are indexed for keyword search and the database tables (entity,attributes, relationships) are made browseable through a user friendly HTML interface. A two dimensional modeling tool has been developed which communicates with both the data warehouse server and the web surface. This modeling tool facilitates pathway construction from the knowledgebase. The 4 problems that need to be addressed are:

1. Traversing the edges of the semantic graph: Traversal is governed by rules that determine which edges (relationships) and nodes(entities) can be walked through.
2. Cutting out subgraphs from an infinite graph: A version of the shortest path algorithm is to be applied in which given two nodes (source and destination) a subgraph must be found which contains a minimum number of edges.
3. Mapping datawarehouse entities into SBML objects: To convert database entities into SBML, a straightforward mapping can be used; reactions can be coded as SBML Reactions, most type of relationships can be translated into SBML SpeciesReferences and most other entities can be interpreted as SBML Species.
4. Laying out subgraph as an aesthetic 2D diagram: A slightly modified version of the Kamada- Kamai [15] graph layouting algorithm is implemented in which edges are imaginary springs, which pull connected nodes together and nodes are imaginary charged spheres which repel each other. Ordinary differential equations can be written up for the forces and the system can be relaxed through iteration.

The *unified datawarehouse* is implemented on a MySQL 5.1.16 server running on Linux Fedora Core 3 platform using traditional relational database technology. The *user interfaces* are powered by the Apache Cocoon servlet which uses pure XML technology to create any publication formats from XML format input source. The Cocoon servlet connects to the MySQL database via JDBC connection. The client side editor tool is implemented in Java utilizing the Graphical Editing Framework (GEF) and Sun SOAP RPC libraries . All software components used are freely extendable and distributable under the lesser or greater GPL license. Methods used for the project are summarized on Figure 2.

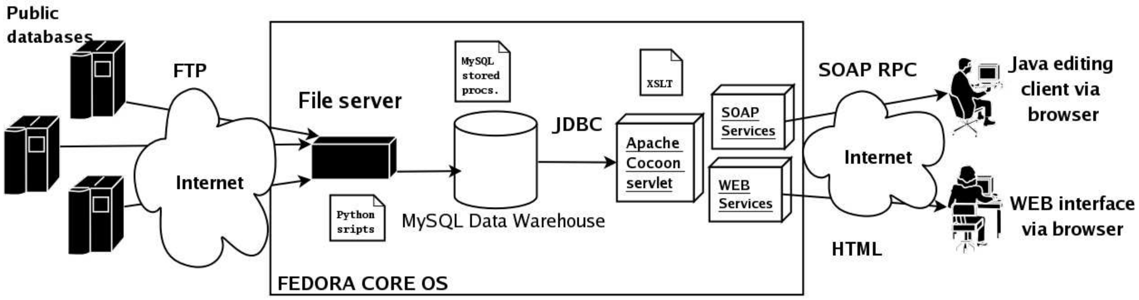


Fig. 2. Methods used for the whole project

3 Results

Data content and storage: The datawarehouse contains 13,383,335 unique entities, 198,923,486 attributes and 12,626,326 relations between entities. The database occupies around 12.3 GB of disk space without indexes, which consume another 14.5GB. This is due to heavy indexing utilized by the full text search features of MySQL. The indexes contain redundant, but performance increasing information.

Biological scope: The datawarehouse combines data from a relevant and representative set of proteomics, metabolomics, genomics, interactome and ontology databases in the hope that new relationships and networks can be incurred from the otherways unconnected data. The data warehouse currently includes data from:

- KEGG [8]: A comprehensive database of metabolic pathways, reactions, compounds, participating genes, pathways.
- The Biomolecular Interaction Network Database (BIND)[9]: A collection of records documenting molecular interactions.
- NCBI Genbank[16]: An annotated collection of all publicly available DNA sequences. In the data warehouse Genbank is parsed as genes, DNA sequences, RNA sequences entities.
- The NCBI Reference Sequence Project[17] (RefSeq): An effort to provide the best single collection of naturally occurring biomolecules, representative of the central dogma, for each major organism. The database is a collection of DNA, RNA and Protein sequences and genes.

- NCBI (Entrez) Gene[18]: An implementation to organize information about genes, serving as a hub between databases internal and external to NCBI. The database structure is built around the central dogma cross referencing genes, DNA, RNA and protein sequences.
- The NCBI Taxonomy[19]: A compilation of the names and IDs to all organisms that are represented in genetic databases with at least one nucleotide or protein sequence.
- Enzyme nomenclature database[20]: A repository of information related to the nomenclature of enzymes. The database contains only enzyme entities.
- The UniProtKB/Swiss-Prot[21] Protein Knowledgebase: A curated protein sequence database that provides a high level of annotation (such as the description of protein function, domains structure, post-translational modifications, variants, etc.).Contains only protein entities.
- The Gene Ontology (GO)[22] project: A collaborative effort to address the need for consistent descriptions of gene products in different databases. The database contains ontology entries of biological process, cellular component and function.

The original number of entities was 15,560,534 of which 2,177,199 (14%) was merged thus leading to the current number of entities. The number of attributes did not change in the integration process as all attributes were inherited by the merged entities for maintaining data provenance (source of data is stored at the attribute level). The nearly 200 million attributes currently in the database can be classified into 273 different types of which the most abundant attributes are NCBI Taxonomy ID, NCBI GI ID, NCBI Gene ID, Pubmed ID, Description.

Using the statistical tables (data not shown), a superficial analysis of the data contained by the database shows that sequence databases are overrepresented compared to interaction databases, the reason for this is that low added value information (sequencing information) is most abundant in databases - in the range of tens of millions of sequences, while high power information (reactions, pathways) are less frequently found, perhaps a few hundred thousand partially overlapping reactions can be found. Although in the entity and attributes datasets the sequence information is overrepresented, relatively few relationships were uncovered until the last load of the database . An explanation might be that the data in the source databases is so inefficiently cross referenced that very few additional information can be incurred from gene expression related data. An example can be seen on Fig 3 where the connectivity of entities in preprocessing tables by frequently used gene and protein identifiers were studied.

User interfaces to the knowledgebase currently comprise of a web services interface and a Java editing tool. *Web services* facilitate easy search & browse of the database. The user can perform keyword search on a Google like interface and returns the result in group of ten entities. By clicking on one of the hits, detailed results are presented. The detailed results contain

- the type of the entity
- the relationships the entity participates
- the detailed attribute list
- links to outside references (if any)
- data sources

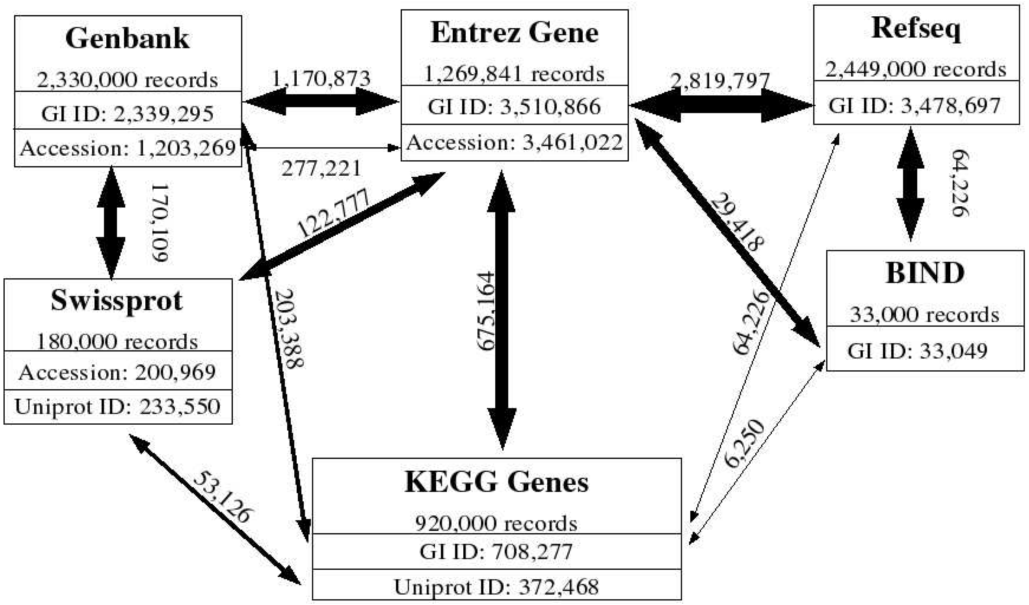
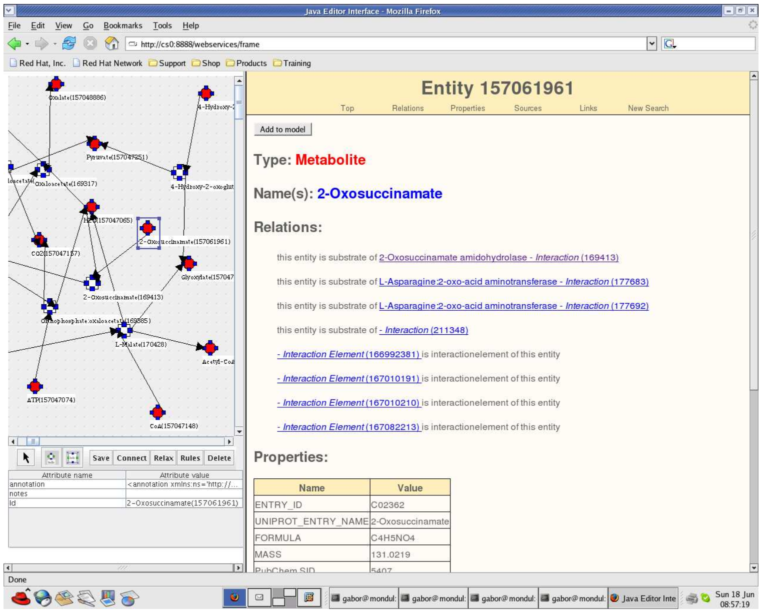


Fig. 3. Connectivity of databases. The numbers on arrows show the number of common ID fields. The thickness of arrows demonstrates the degree of the chances of linking the two databases by the given ID.



Attribute name	Attribute value
annotation	<annotation xmlns:ns='http://...
notes	
Id	2-Oxosuccinamate(157061961)

Entity 157061961

Type: Metabolite

Name(s): 2-Oxosuccinamate

Relations:

this entity is substrate of 2-Oxosuccinamate amidohydrolase - *Interaction* (169413)

this entity is substrate of L-Asparagine:2-oxo-acid aminotransferase - *Interaction* (177683)

this entity is substrate of L-Asparagine:2-oxo-acid aminotransferase - *Interaction* (177692)

this entity is substrate of - *Interaction* (211348)

- *Interaction Element* (166992381) is interactionelement of this entity

- *Interaction Element* (167010191) is interactionelement of this entity

- *Interaction Element* (167010210) is interactionelement of this entity

- *Interaction Element* (167082213) is interactionelement of this entity

Properties:

Name	Value
ENTRY_ID	C02362
UNIPROT_ENTRY_NAME	2-Oxosuccinamate
FORMULA	C4H5NO4
MASS	131.0219
PubChem SID	5407

Fig. 4. The most recent version of the user interface enables interactive browsing in the knowledgebase as well computer aided build up of on bespoke models

See Figure 4 for sample screenshot. The related entities are hypertext referenced and thus a whole network of interacting molecules can be explored within a short space of time. To make the browsing through the biomolecular network more user friendly a Java editing tool is being developed. The Java editing tool is a simplified SBML editor which performs the following tasks:

- keeps track of the entity information pages the user has visited
- builds a graphical representation of the network walked through upon request
- extends this graph by automatic actions
- saves the graph, which in fact is a model skeleton, in SBML format
- annotates SBML model skeleton with database identifier information

The user can then perform various other modeling tasks with the saved model skeleton (SBML model not necessarily containing reaction parameters), improve it into a true model, fit parameters, etc, with classical modeling tools of their choice.

4 Conclusion and future work

In this paper we have presented a comprehensive model editing environment for biomolecular modeling based on information from publicly available databases. Further work on integrating the data warehouse into the modeling environment is needed as the database services must be made available to the E-cell's native Model Editor. The most important use cases for this integration are: annotation of modeled species and reactions with database identifiers, and merging of small models from different sources, which might need database driven reconciliation of names and Ids.

The biological scope of the current database needs to be extended with explicit transcriptional binding information (other than that of BIND dataset) and kinetic information, all of these extensions fit well into the current SBML compatible framework.

A limitation of the current implementation is that it does not accommodate updates from the original databases. The datawarehouse needs to be built from scratch every time an update is carried out. There are no design problems that would not allow an incremental update approach since the information about the data source and its primary ID is stored in the attributes table and hence can be deleted or updated. However, the heavy indexing of tables required to speed up servicing user requests may prevent performing incremental databases on a regular basis.

Connectivity issues outlined in the data analysis section raise the need for obtaining relationship information from homology orthology investigations though performing real time mass runs from high throughput BLAST runs.

5 Acknowledgments

I would like to express my thanks to Kouichi Takahashi for advising me at the startup of this project, Sakurada Takeshi for providing the hardware and network and Arun Krishnan for precious advice.

References

1. Tomita et al.: E-CELL: Software for whole cell simulation Bioinformatics **15** (1999) 72-84
2. Takahashi et al.: A multi-algorithm, multi-timescale method for cell simulation Bioinformatics **20** (2004) 538-46
3. Kouichi Takahashi: Multi-algorithm and multi-timescale cell biology simulation PhD Thesis (2004)
4. Michael Y. Galperin: The Molecular Biology Database Collection: 2005 update Nucl. Acids Res. **33** (2005) D5-D24
5. Arakawa et. al: GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes BMC Bioinformatics **7** (2006)
6. Catherine Brooksbank, Graham Cameron and Janet Thornton : The European Bioinformatics Institute's data resources: towards systems biology Nucl. Acids Res. **33** (2005) D46-D53
7. Wheeler et al.: Database resources of the National Center for Biotechnology Information Nucl. Acids Res. **33** (2005) D39-D45
8. Kanehisa et al. The KEGG resources for deciphering the genome Nucl. Acids Res. **32** (2004) D277-D280
9. Alfarano et al.: The Biomolecular Interaction Network Database and related tools 2005 update Nucl. Acids Res. **33** (2005) D418-D424
10. http://www.ebi.ac.uk/biomart/
11. Hou et al. BioSilico: an integrated metabolic database system Bioinformatics **20** () 3270-3272
12. Mork et al. The Multiple Roles of Ontologies in the BioMediator Data Integration System Data Integration in the Life Sciences LNCS **3615** (2005)
13. Aaron Birkland and Golan Yona: BIOZON: a system for unification, management and analysis of heterogenous biological data BMC Bioinformatics **7** (2006) 70
14. Hucka et al.: The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models Bioinformatics **19** (2003) 524-31
15. Kamada T., Kawai S.: An algorithm for drawing general undirected graphs Inf. Process. Lett. **31** (1989)
16. Dennis A. et al. Genbank Nucl. Acids Res. **33** (2005) D34-D38
17. Pruitt et al. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins Nucl. Acids Res. **33** (2005) D501-504
18. Maglott et al. Entrez Gene: gene-centered information at NCBI Nucl. Acids Res. **33** (2005) D54-D58
19. Wheeler et al. Database resources of the National Center for Biotechnology Information Nucl. Acids Res. **28** (2000) 10-4
20. Bairoch A. : The ENZYME database in 2000 Nucl. Acids Res. **28** (2000) 304-305
21. Bairoch et al. The Universal Protein Resource (UniProt) Nucl. Acids Res. **33** (2005) D154-D159
22. The Gene Ontology consortium: Gene Ontology: tool for the unification of biology. Nature Genet. **25** (2000) 25-29.

Mining Invariants in Biological Sequences

Yingwei Wang[1], Lawrence Hale[2], Kathleen Hill[3], and Shiva Singh[3]

[1] Department of Computer Science and Information Technology,
University of Prince Edward Island,
Charlottetown, Prince Edward Island, Canada C1A 4P3
ywang@upei.ca

[2] Department of Biology, University of Prince Edward Island,
Charlottetown, Prince Edward Island, Canada C1A 4P3
lhale@upei.ca

[3] Department of Biology, University of Western Ontario,
London, Ontario, Canada N6A 5B7
{ssingh,khill22}@uwo.ca

Abstract. In this paper we first discuss the existence and importance of invariants in biological sequences, and then explore the features and approaches in mining these invariants. We show some interesting invariants in biological sequences, including Dinucleotide Relative Abundance Profiles (DRAPs) and Chaos Game Representations (CGRs). We also discuss the unique features and possible approaches of invariability mining.

1 Introduction

Data mining is defined as the process of analyzing data to identify patterns or relationships. In early days, the objects of data mining were mainly database tables. Later, the objects of data mining were expanded to text, and more recently, to biological sequences [1][2][3][4].

Biological sequences include DNA, RNA, and protein. With the continuing progress in molecular biology and bioinformatics, huge numbers of biological sequences are becoming available to data mining. As these sequences are rich with patterns and relationships, mining biological sequences holds great promise.

Patterns and relationships can be mined from biological sequences in many different ways; in this paper, we focus on mining a specific pattern: invariability.

2 Invariants As Functions

First, we review the concepts of invariability and invariant. Invariability is defined as a quality of uniformity and lack of variation. Invariant is defined as a feature (quantity or property or function) that remains unchanged when a particular transformation is applied to it.

For example, let us consider the invariant π. When we say π is an invariant, we do not mean merely the number beginning with 3.1415926. Rather, we mean that the underlying quantity (i.e. the ratio of the circumference of a circle to its diameter) is an invariant. The fact that the ratio of the circumference of a circle to its diameter is always the same value shows the relationship between circumference and diameter of a circle.

Based on the above understanding, we may define an invariant in biological sequences as a function on sequences that does not change or changes in a very small range.

We may use an example to show an invariant in biological sequences: Suppose function $n(s)$ is the number of different nucleotides in DNA sequence s. We know that $n(s) \leq 4$ because there are only 4 possible nucleotides: A, C, G, and T.

If s is an artificially synthesized DNA sequence, $n(s)$ could be 1, 2, 3, or 4. If s is shorter than 4, $n(s)$ cannot be 4. If s is a naturally existing sequence and is long enough (such as longer than 10,000 base pairs), it is likely that $n(s) = 4$ holds. We understand that counterexamples may exist, but this does not really matter because we only use $n(s)$ as an example to illustrate the concept of invariants.

Suppose $n(s) = 4$ holds for any naturally existing sequence s where s is longer than 10,000 base pairs, we may say $n(s)$ is an invariant. This invariant is very simple and it is only used to illustrate the concept of invariants. Usually, an invariant in biological sequences only holds within a particular *scope*. In the $n(s)$ example, the scope is naturally existing DNA sequences that are longer than 10,000 base pairs.

3 Some Invariants in Biological Sequences

In this section, we describe a few more invariants in biological sequences. We have two goals in mind when we introduce these invariants. The first goal is to get familiar with, and to expand, the concept of invariants; the second goal is to set these invariants as the foundation of the discovery of more invariants.

3.1 DRAP

For a sequence s, the Dinucleotide Relative Abundance Profile [5] DRAP(s) is an array $\{\rho_{XY} = f_{XY}/f_X f_Y\}$, where XY stands for all possible dinucleotide combinations, f_X denotes the frequency of the mononucleotide X in s and f_{XY} the frequency of the dinucleotide XY in s.

Karlin and Burge observed that DRAPs of different DNA sequence samples from the same organism are generally much more similar to each other than to those of sequences from other organisms. In addition, closely related organisms generally have more similar DRAPs than distantly related organisms.

$DRAP(s)$ is a function on DNA sequences, but it is different from $n(s)$. For a DNA sequence s, $n(s)$ produces a value, but $DRAP(s)$ produces an array that contains 16 values. This reminds us that in the definition of invariants we should accommodate not only single value functions, but also functions which produce other data formats, such as arrays, matrices, and so on.

Another point we should notice is that $DRAP(s)$ is not a constant. For different DNA sequence samples from the same organism, $DRAP(s)$ will produce arrays that are similar, but not exactly the same. From this point of view, $DRAP(s)$ is not an invariant because it does change. In the biological world, quantities that do not change at all are very rare. However, quantities that change in a small range can be considered as invariants for practical purposes.

The scope of invariant $DRAP(s)$ is also different from the scope of invariant $n(s)$. For $DRAP(s)$, the scope is only for sequences from the same organism, which is much smaller than the scope of $n(s)$.

Invariant $DRAP(s)$ are unique within a specific species. It was concluded that the DRAP values constitute a genomic signature of an species.

3.2 CGR

In 1990, Jeffrey proposed using *Chaos Game Representation (CGR)* to visualize DNA primary sequence organization [6]. A CGR is plotted in a square, the four vertices of which are labelled by the nucleotides A, C, G, T, respectively. The plotting procedure can be described by the following steps: the first nucleotide of the sequence is plotted halfway between the centre of the square and the vertex representing this nucleotide; successive nucleotides in the sequence are plotted halfway between the previous plotted point and the vertex representing the nucleotide being plotted. The major advantage of CGR is that it is a two-dimensional plot that can provide a visual representation of primary DNA sequence organization for a sequence of any length, including entire genomes.

Various geometric patterns, such as parallel lines, squares, rectangles, and triangles are among the interesting patterns that can be found in CGRs. Some CGRs even show a complex fractal geometrical pattern which is very similar to the Sierpinsky Triangle [7]. These interesting features relevant to the DNA sequence organization attracted further research in CGR [8][9][10][11].

CGR is not just a visualization tool. In fact, CGR is itself another interesting invariant. Experiments showed that variation between CGR images within a genome was smaller than variation among genomes [12]. Fig. 1 shows the similarity of CGR images within the same genome. Fig. 2 shows the dissimilarity of CGR images in different genomes.

CGRs in their original form are not easily processed by a computer. Thus another form of CGR was introduced: FCGR (the Frequency matrix extracted from a CGR). The structure of FCGR was introduced in [12] and the name FCGR was proposed in [13]. A kth-order FCGR is defined as follows.

A kth-order FCGR of a sequence s, denoted by $FCGR_k(s)$, is a $2^k \times 2^k$ matrix. To obtain this FCGR, we first plot a CGR from s, then divide this CGR by a $2^k \times 2^k$ grid so that each grid square corresponds to an element in the matrix. We then count the number of points *inside* each grid square, and use the number of points as the matrix element corresponding to the grid square. We do not count those points on the grid square lines because they represent the length $k-1$ oligonucleotide at the beginning of the DNA sequence, and we can omit these $k-1$ points as long as the DNA sequence is much longer than

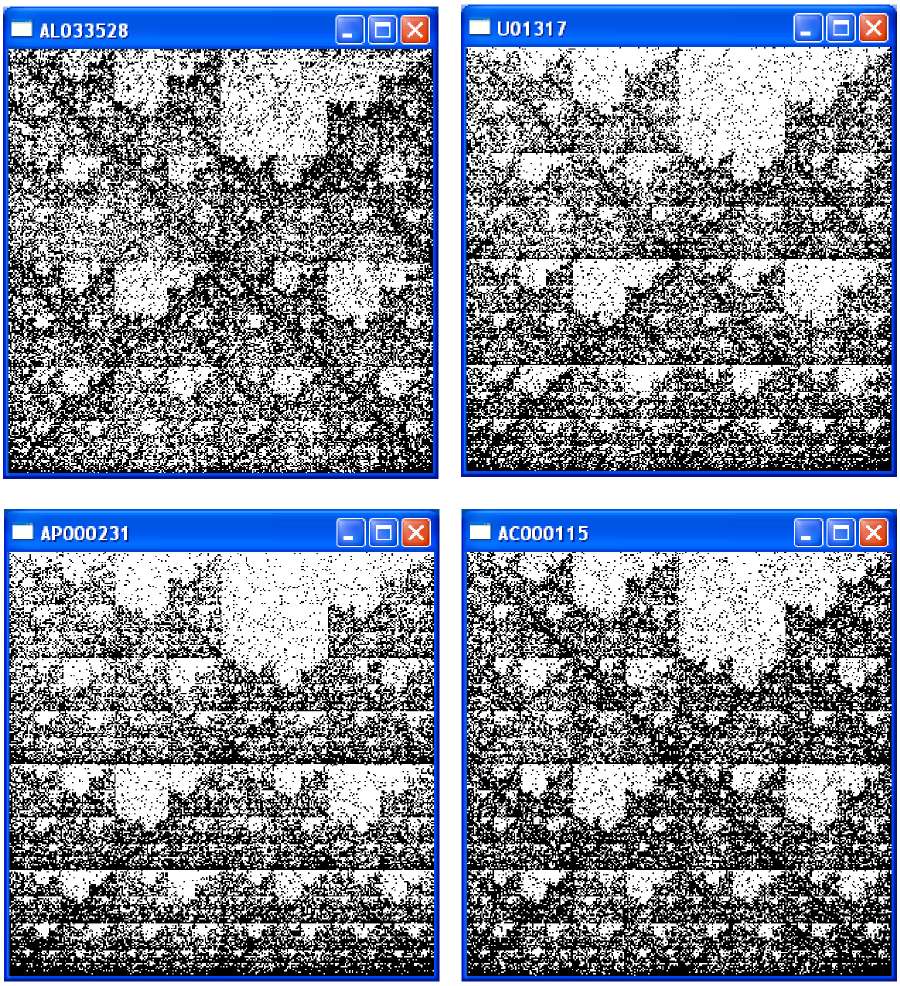


Fig. 1. These four CGRs are all plotted from DNA sequences from the human genome. AL033528 (83111 bp) is from Chromosome 1; U01317 (73308 bp) is from Chromosome 11; AP000231 (77728 bp) is from Chromosome 21; AC000115 (95855 bp) is from Chromosome X. These sequences have different lengths so that some of them are darker and others are lighter. They are very similar in the general patterns.

k. Note that, instead of being a graphical representation like CGR, a FCGR is a numerical matrix.

A FCGR can also be constructed directly from a sequence instead of plotting a CGR first and then converting the CGR into a FCGR. We can construct a FCGR directly by counting the number of occurrences of each length k oligonucleotide in the sequence and putting this number into the appropriate place of

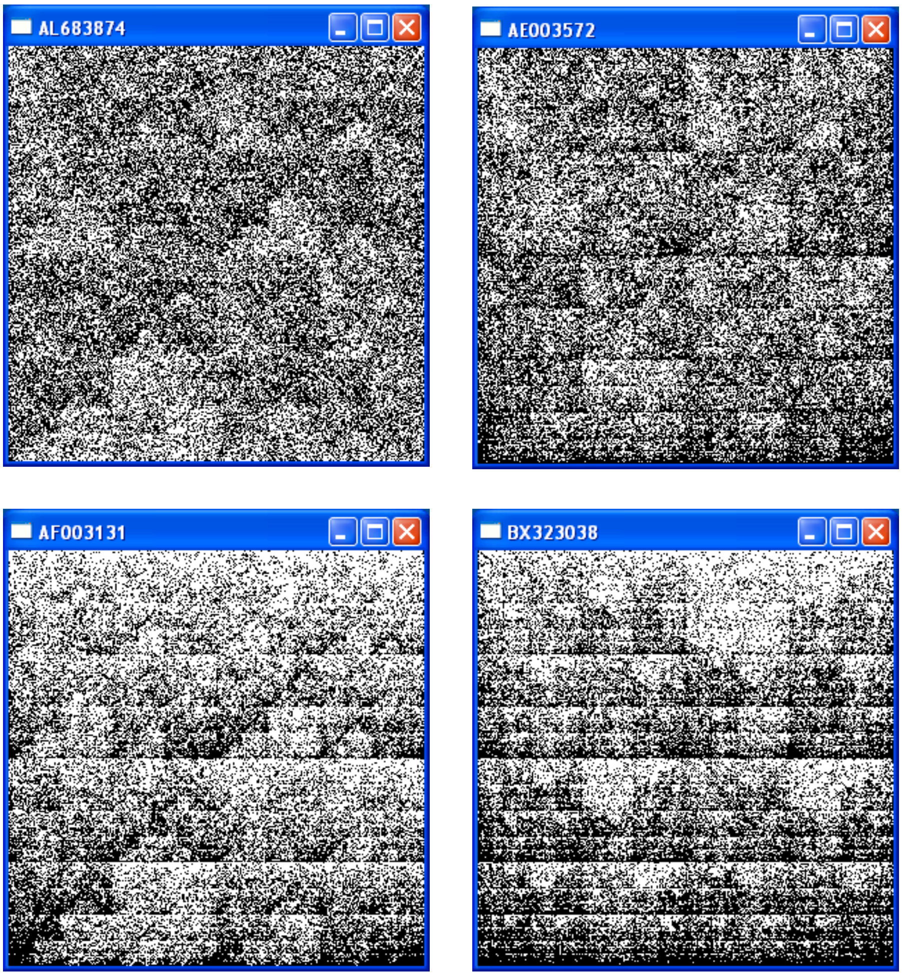


Fig. 2. These four CGRs are plotted from DNA sequences from different genomes. AL683874 (68100 bp) is from a fungus; AE003572 (91383 bp) is from fruit fly; AF003131 (65649 bp) is from nematode; BX323038 (87548 bp) is from zebrafish. These sequences have different lengths so that some of them are darker and others are lighter. They are very different in the general patterns.

the FCGR matrix, according to the correspondence between a length k oligonucleotide and a CGR grid square.

A first-order FCGR and a second-order FCGR have the structure shown below, where N_w is the number of occurrences of the oligonucleotide w in the sequence s.

$$FCGR_1(s) = \begin{pmatrix} N_C & N_G \\ N_A & N_T \end{pmatrix}$$

$$FCGR_2(s) = \begin{pmatrix} N_{CC} & N_{GC} & N_{CG} & N_{GG} \\ N_{AC} & N_{TC} & N_{AG} & N_{TG} \\ N_{CA} & N_{GA} & N_{CT} & N_{GT} \\ N_{AA} & N_{TA} & N_{AT} & N_{TT} \end{pmatrix}$$

The definition of $FCGR_{k+1}(s)$ can be obtained by replacing each element N_X in $FCGR_k(s)$ with 4 elements

$$\begin{matrix} N_{CX} & N_{GX} \\ N_{AX} & N_{TX} \end{matrix}$$

Here we omit the definition of CGR resolution and detailed mathematical discussions, and give the following conclusion directly: a kth-order FCGR is equivalent to a CGR of resolution $\frac{1}{2^k}$. This conclusion describes the relationship between CGRs and FCGRs. CGRs and FCGRs have been thoroughly explored in [11].

CGRs and FCGRs are important invariants in DNA sequences. The scope of each invariant in this category is that the DNA sequences are within a specific genome, and that the sequence should be not too short (at least 1k bp).

These invariants have very important biological meanings. To some extent, a CGR plotted using a DNA sequence reflects a unique identity of the species from which the DNA sequence was extracted. CGRs and FCGRs are also important starting points in searching for more invariants in DNA sequences.

4 Features and Approaches of Invariability Mining

From the analysis and illustrations in Section 2 and Section 3, we know that invariants do exist in DNA sequences. In this section we discuss the features and approaches of mining invariants from DNA sequences.

The first feature of invariability mining is that this is totally a new area. We refer to data mining on database tables as traditional data mining (We use this term only for convenience; there is no negative meaning). Although invariability mining is also a process of analyzing data to identify patterns and relationships, it is different from traditional data mining. As such, the techniques developed in traditional data mining usually cannot be applied to invariability mining directly. There are two major differences. First, traditional data mining deals with database tables, whereas invariability mining deals directly with biological sequences. Second, the mining goals are different. Invariability mining is interested in a specific type of pattern–invariants, whereas traditional data mining tries to find association relationships among variables. These differences pose difficulties and challenges for researchers in this area, but are ultimately very exciting.

The second feature of invariability mining is that this area is tightly related to biological discoveries. An invariant found in biological sequences usually reveals important biological regularities. The interpretation of such invariants could be an important topic. At the same time, the discovery of invariants relies heavily on a priori biological knowledge. The ultimate reasons of the existence of biological invariants are the natural regularities. There is no such area as invariability mining in general databases. Without natural regularities, invariability mining is meaningless.

Efficient methods and algorithms in the area of invariability mining of biological sequences are yet to be discovered. Here we describe two of the major approaches that can be used in this new area.

The first approach in invariability mining is transformation. Existing invariants (including the ones described in Section 3) are good starting points. We may analyze these invariants, find the components that reflect the regularities, and propose new possible invariants.

The second approach in invariability mining is search with prediction. Often-used functions in existing invariants may be accumulated and tested with different sequence groups. Traditional data mining techniques, such as clustering, can be used in determining the promising sequence groups.

5 Conclusion

Mining invariants in biological sequences is a new area which fits the broad definition of data mining, but has features that are different from traditional data mining.

In this paper, we first explore the concepts in this new area, point out that an invariant can be described by a function and defining the scope of this function.

We describe some important invariants existing in biological sequences, such as DRAPs and CGRs. These invariants not only give us concrete examples of invariants, but also help us to clarify definitions. These invariants also serve as starting points in further mining processes.

The general features and approaches of invariability mining are also discussed. We point out that invariability mining is a new area and it is tightly related to biological discoveries. We may use the transformation approach to propose new invariants from existing invariants, or use search with prediction approach to find new invariants.

In conclusion, mining invariants in biological sequences is an emerging and promising area, and it is worth further exploration.

References

1. Li, J., Wong, L., Yang, Q.: Data mining in bioinformatics. IEEE Intelligent Systems **20** (2005) 16–18
2. Evans, S., Lemon, S., Deters, C., Fusaro, R., Durham, C., Snyder, C., Lynch, H.: Using data mining to characterize DNA mutations by patient clinical features. In: Proceedings AMIA Annu Fall Symp. (1997) 253–257

3. Yin, M.M., Wang, J.T.L.: Mining genes in DNA using genescout. In: Proceedings of second IEEE International Conference on Data Mining. (2002) 733–733
4. Hoffman, P., Grinstein, G., Marx, K., Grosse, I., Stanley, E.: DNA visual and analytic data mining. In: Proceedings Visualization'97. (1997) 437–441
5. Karlin, S., Burge, C.: Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics **11** (1995) 283–290
6. Jeffrey, H.J.: Chaos game representation of gene structure. Nucleic Acids Research **18** (1990) 2163–2170
7. Mandelbrot, B.: The Fractal Geometry of Nature (2nd edition). W. H. Freeman and Co., San Francisco, California (1982)
8. Dutta, C., Das, J.: Mathematical characterization of chaos game representation. Journal of Molecular Biology **228** (1992) 715–719
9. Hill, K.A., Schisler, N.J., Singh, S.M.: Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. Journal of Molecular Evolution **35** (1992) 261–269
10. Oliver, J.L., Bernaola-Galvan, P., Guerrero-Garcia, J., Roman-Raldan, R.: Entropic profiles of DNA sequences through chaos-game-derived images. Journal of Theoretical Biology **160** (1993) 457–470
11. Wang, Y., Hill, K., Singh, S., Kari, L.: The spectrum of genomic signatures: from dinucleotides to chaos game representations. GENE **346** (2005) 173–185
12. Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B.: Genomic signuture: Characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. **16** (1999) 1391–1399
13. Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A., Fletcher, M.: Analysis of genomic sequences by chaos game representation. Bioinformatics **17** (2001) 429–437

Case Mining from Medical Literature

Isabelle Bichindaritz

University of Washington, 1900 Commerce Street, Box 358426,
Tacoma, WA 98402, USA
ibichind@u.washington.edu

Abstract. This article addresses the task of mining cases from biomedical literature to automatically build an initial case base for a case-based reasoning (CBR) system. This research takes place within the Mémoire project, which has for goal to provide a framework to facilitate building CBR systems in biology and medicine. By analyzing medical literature, the CaseMiner system mines for medical concepts such as diseases, signs and symptoms, laboratory tests, and treatment plans all connected together in a given medical domain. It then organizes these concepts in a structure called a case. This case mining component provides a definite help to start-up the creation of a biomedical CBR system case base, composed of both concrete cases and prototypical cases. Most cases learnt by CaseMiner are prototypical case, but some of the cases learnt from medical case studies really correspond to actual patients' cases. This article validates the approach by presenting a comparison between the prototypical cases learnt from stem-cell transplantation domain with those created by a team of experts in the domain.

1 Introduction

Case-based reasoning (CBR) systems in biomedicine rely on patients' cases to propose diagnosis assessment and treatment recommendations in case-based decision-support systems. Often these systems have reported that the cases readily available in electronic format are incomplete at best, and have resorted to multimodal reasoning systems to complement the cases with knowledge bases expressed in models and/or rules. Many times, cases are not even available in electronic format, which requires a tremendous amount of time entering data into the CBR systems just to bootstrap it. This context prompts for the design of advanced automatic knowledge elicitation tools to provide CBR systems with the adequate knowledge they need for reasoning, without spending years elicitating this knowledge from experts. CaseMiner system presented in this article builds on a current trend to develop case mining systems to take advantage of electronically available knowledge sources that may be mined for cases.

The idea of mining cases from medical literature comes from current trends in text mining research from medical literature. Recently, the fast growing number of biomedical publications has motivated the development of innovative information extraction and data mining systems and tools [8, 9]. A new field in information extraction aims at discovering knowledge from literature in the form of unknown and meaningful relationships between concepts found in biomedical bibliographic databases [11, 15] such as Medline [24]. This idea of discovering new relations from a bibliographic database was pioneered by Swanson [23, 24], who proposes a text mining system that made seven medical discoveries that have been later published in relevant medical journals. Following in his tracks, methods and systems have been developed to mine for new knowledge from literature as novel concepts and relationships.

The system presented here proposes to automate the process of mining for cases from biomedical literature. It builds on a concept miner learning for relationships between concepts, such as the relationship between caloric restriction and aging, and not for isolated concepts. The system presented combines some information extraction features and some data mining features. First, it extracts the different parts of a document such as title, keywords, abstract, and body, then it mines for concepts and relationships in the document, before aggregating these in a prototypical case structure.

The next section presents case representation for the case mining system. The third section introduces the Unified Medical Language System (UMLS) as the ontology guiding the discovery process. The fourth section sets forth the CaseMiner system architecture and different components. The fifth section presents an evaluation of the system. It is followed by a discussion and a conclusion.

2 Case Representation

The goal of the Mémoire project [3] at the University of Washington is to provide a framework for the creation and interchange of cases, concepts, and CBR systems in biology and medicine. Its approach is to generalize from previous CBR systems built in biomedicine, mostly from Carepartner [3] at first.

The cornerstone of the knowledge acquisition process has been the conception of prototypical cases, called clinical pathways in this system. This prototypical case structure is important for this article because it is also the prototypical case structure proposed in Mémoire as a generic prototypical case representation structure. Prototypical cases serve as generic cases, and consequently, this is the kind of cases that CaseMiner is mining for. The clinical pathways, 91 of them having been implemented in the test version of the system, correspond to clinical diagnostic categories for the most part, some of them corresponding also to essential signs and symptoms requiring specific assessment or treatment actions. The clinical pathways are knowledge structures represented from the ontology described above, namely: all diseases, functions (also known as signs and symptoms), labs, procedures, medications, sites, and planning actions. Most of the terms naming these objects are standardized using the Unified Medical Language System (UMLS) ontology. Only the terms not corresponding to objects in the UMLS have been added to the domain specific ontology. In particular, the planning actions used in the Treatment part of a prototypical case did not exist in the UMLS and were all created for the system.

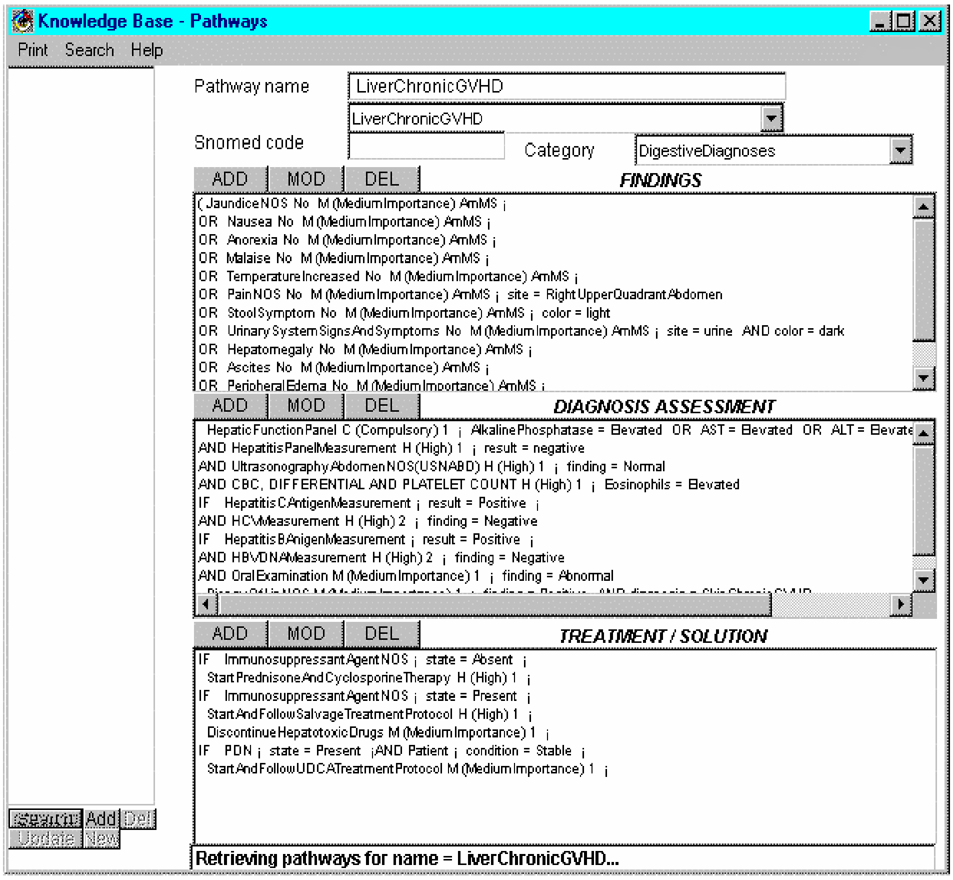
An example of a prototypical case is provided on Figure 1 (chronic graft versus host disease - a complication of stem-cell transplantation). It shows that a prototypical case comprises three parts:

1. A list of findings, corresponding to signs and symptoms.
2. A diagnosis assessment plan, which is a plan to follow for confirming (or informing) the diagnosis.
3. A treatment/solution plan, which is a plan to follow for treating this disease when confirmed, or a solution when the pathway does not correspond to a disease.

The diagnosis assessment part and the treatment part can also be seen as simplified algorithms, since they use IF THEN ELSE structures, and LOOP structures, as well as SEQUENCE structures of actions in time, which, when instantiated with actual patients' data, provide a diagnosis assessment plan, or a treatment plan, tailored to a specific patient. In this way, this knowledge structure allows for sophisticated adaptation when reusing a prototypical case.

3 UMLS Project

The “Unified Medical Language System” (UMLS) from the National Library of Medicine (NLM) [18], a specialized thesaurus in biomedicine, provides standardized concepts for the creation of a controlled domain vocabulary. The UMLS provides a very powerful resource for rapidly creating a robust scientific thesaurus in support of precision searching, and a starting point for an ontology of the medical domain. Further, the semantic type descriptors for each concept and semantic network may offer some interesting opportunities for intelligent searching and mapping of concepts representing research findings, and their relationships.



Pathway name: LiverChronicGVHD

LiverChronicGVHD

Snomed code:

Category: DigestiveDiagnoses

FINDINGS

```
( JaundiceNOS No  M (MediumImportance) AmMS ¡
OR  Nausea No  M (MediumImportance) AmMS ¡
OR  Anorexia No  M (MediumImportance) AmMS ¡
OR  Malaise No  M (MediumImportance) AmMS ¡
OR  TemperatureIncreased No  M (MediumImportance) AmMS ¡
OR  PainNOS No  M (MediumImportance) AmMS ¡  site = RightUpperQuadrantAbdomen
OR  StoolSymptom No  M (MediumImportance) AmMS ¡  color = light
OR  UrinarySystemSignsAndSymptoms No  M (MediumImportance) AmMS ¡  site = urine  AND color = dark
OR  Hepatomegaly No  M (MediumImportance) AmMS ¡
OR  Ascites No  M (MediumImportance) AmMS ¡
OR  PeripheralEdema No  M (MediumImportance) AmMS ¡
```

DIAGNOSIS ASSESSMENT

```
  HepaticFunctionPanel C (Compulsory) 1  ¡  AlkalinePhosphatase = Elevated  OR  AST = Elevated  OR  ALT = Elevate
AND HepatitisPanelMeasurement H (High) 1  ¡  result = negative
AND UltrasonographyAbdomenNOS(USNABD) H (High) 1  ¡  finding = Normal
AND CBC, DIFFERENTIAL AND PLATELET COUNT H (High) 1  ¡  Eosinophils = Elevated
IF   HepatitisCAntigenMeasurement ¡  result = Positive  ¡
AND HCVMeasurement H (High) 2  ¡  finding = Negative
IF   HepatitisBAnigenMeasurement ¡  result = Positive  ¡
AND HBVDNAMeasurement H (High) 2  ¡  finding = Negative
AND OralExamination M (MediumImportance) 1  ¡  finding = Abnormal
```

TREATMENT / SOLUTION

```
IF   ImmunosuppressantAgentNOS ¡  state = Absent  ¡
  StartPrednisoneAndCyclosporineTherapy  H (High) 1  ¡
IF   ImmunosuppressantAgentNOS ¡  state = Present  ¡
  StartAndFollowSalvageTreatmentProtocol  H (High) 1  ¡
  DiscontinueHepatotoxicDrugs M (MediumImportance) 1  ¡
IF   PDN ¡  state = Present  ¡AND Patient ¡  condition = Stable  ¡
  StartAndFollowUDCATreatmentProtocol M (MediumImportance) 1  ¡
```

Retrieving pathways for name = LiverChronicGVHD...

Figure 1. A clinical pathway, corresponding to a prototypical case, for chronic graft versus host disease (CGVHD)

Syntactic and semantic analysis tools for automated Natural Language Processing (NLP) are also provided by the National Library of Medicine’s UMLS project [16, 17]. UMLS ultimate goal is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health.

By navigating the semantic network provided, it is possible to know which concepts extracted by the NLM tools from biomedical documents correspond to diseases, which correspond to findings, which correspond to medications, etc.. It is also possible to know which relationships connect different concepts. There is a total of 135 semantic types and 54 relationships provided by the UMLS semantic network. Additionally, it is possible to extend the semantic network, for our purpose with a semantic network of planning actions that can be connected with a 'treat' relationship with other concepts.

4 CaseMiner case miner

CaseMiner system mines for cases and prototypical cases from biomedical literature. A selection of documents for a given medical domain is the input to this system. Pertinent documents may be literature articles, but also textual clinical practice guidelines, and medical case studies. It is important that such documents should all be related to a given domain, such as in our example stem-cell transplantation.

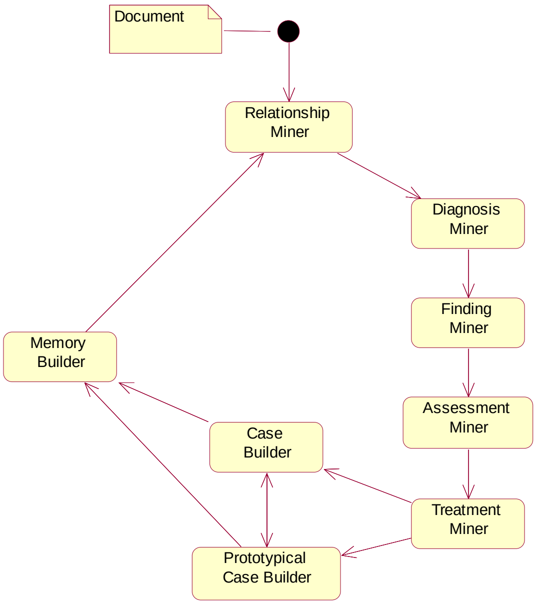


Figure. 2. CaseMiner architecture

4.1 Architecture

ConceptMiner core component is the RelationshipMiner, which mines for triples <*concept1-1, relationship-1,2, concept-2*> from a document. It also attaches a condition to a triple when it finds it to represent the information that IF a condition occurs, then an action or test is undertaken. This can be represented as <*concept-1, relationship-1,2, concept-2*> *IF* <*concept-3, relationship-3,4, concept-4*>. An example can be <*Patient, startTreatment, PrednisoneAndCyclosporineTherapy*> *IF* <*absent, property_of, Immuno-suppressantAgentNOS*>. This structure is called a triple pair.

CaseMiner interprets the results from RelationshipMiner by successively mining for diagnoses in DiagnosisMiner, findings in FindingMiner, assessments in AssessmentMiner, and treatments in TreatmentMiner. Following, it builds cases from these results in CaseBuilder or PrototypicalCaseBuilder. The order between these two components can be altered since in some cases, learnt relationships will be associated with conditions, which signals a prototypical case, and in others there will not be any of these conditions, which signals a practice case. Generally, from medical articles and clinical practice guidelines, the learnt artifact will be a prototypical case. From clinical case studies, the learnt artifact will be a practice case. The previous steps deal with prototypical cases and practice cases built from scratch from a single document. A next step is to consolidate learning results across documents. This step is called MemoryBuilder.

Figure 2 represents the architecture of the system with its different components.

4.2 Relationship miner

The RelationshipMiner component is a precursor system of CaseMiner developed for the Telemakus system [10], which consists of a set of domain documents (current focus is the biology of aging), a conceptual schema to represent the main components of each document, and a set of tools to query, visualize, maintain, and map the set of documents through their concepts and research findings [10]. For that purpose, this system mines and maps research findings from research literature. At present, knowledge extraction resorts to systems with both manual and automated components. A key area of current work is to move towards automating the research concept identification process, through data mining [10]. This is exactly why RelationshipMiner was developed.

The author's research team developed an automated system to mine for concepts linked by relationships from biomedical literature [4]. This system originally kept only the pairs of concepts in relationship, for the indexing purposes of Telemakus project, and not the relationships between these. The system since then has been improved both in its concept mining features, and a relationship mining feature has been added [5].

The RelationshipMiner involves two knowledge bases, UMLS database, and domain specific database (DSDB), which in particular stores the pre-processed documents that will serve as the input to the system. Within DSDB, the domain specific thesaurus represents the standardized vocabulary of the domain. Concept mining involves processing articles already stored in domain-specific database (DSDB) by parsing the different parts of a document from their Medline structure. These articles comprise the full text of the original articles, parsed in several parts, such as title, summary, section part, figure and table legends, and so forth.

The components of the system are described in Bichindaritz and Akkineni [4] and Bichindaritz [5]. Following steps of parsing and analyzing literature articles using in particular NLM's syntactic and semantic analysis components [16, 17], RelationshipMiner produces triples of the form <*first concept, relationship, second concept*> , such as for example <*caloric restriction, effects_of, young age rats*>.

4.3 Diagnosis, finding, assessment, and treatment miner

From the relationships triples and pairs of triples discovered by RelationshipMiner, diagnoses are extracted by mapping the concepts within these relationships with the UMLS semantic network ‘disease’ concept type. Similarly, findings are mapped to the UMLS semantic network ‘finding’ concept type, assessment with ‘laboratory or test’ concept type, and treatment with ‘clinical drug’, ‘substance’, ‘food’, or ‘planning action’.

4.4 Case builder

After triples are built, and if no triple pair is associated with a specific diagnosis, a case can be built by connecting in a case structure the triples associating a patient with his/her list of diagnoses, list of findings, list of assessment results, and list of treatment actions. A mined case is represented on figure 3.

Connector	Finding Name	(Properties, Values)
	diagnosis	Name= LiverChronicGVHD
AN	Nausea	
AN	Anorexi	
AN	PainNOS	site=RightUpperQuadrantAbdomen
AN	Stoo	color=light
AN	ImmunosuppressantAgentNOS	status=absent

Diagnosis Assessment

Connector	Procedure Name	(Properties, Values)

	HepaticFunctionPanel	finding=AlkalinePhosphataseMeasurement(ALKP) result=elevate
		finding=ASTMeasurement(AST) result=elevated
		finding=ALTMeasurement(ALT)result=elevate
		finding=LDHMeasurement(LDH)result=elevated

AN	UltrasonographyAbdome nNOS(USNABD	finding=Normal
AN	HepatitisCAntigenMeasurement	result = Negative
AN	OralExamination	finding=abnormal
AN	RequestGIConsul	finding=LiverChronicGVHD

Treatment Plan

Condition/Connector	Planning Action Name
	StartPDNCSPTherapy

Figure 3. Practice case mined

4.5 Prototypical case builder

After triples are built, and if some triple pairs are associated with a specific diagnosis, a prototypical case can be built for this diagnosis by connecting in a prototypical case structure the triples associating a patient with his/her list of diagnoses, list of findings, list of assessment results, and list of treatment actions.

A summarized example can look like the clinical pathway provided on figure 1, although some elements are not learnt now, such as the importance of a finding.

4.6 Memory builder

Memory builder is the processing step that takes place after cases and prototypical cases have been built from a document. Generally, prototypical cases can be described in different documents, often partially. For example, one document describes the symptoms of a disease, while another document describes the assessment process, while yet another describes the treatment plan. The link between the different parts is established from the knowledge that all these parts are related to a particular diagnosis. If a prototype for a disease already exists in memory, the following operations take place:

Merge the list of findings that are not contradictory. If some findings are contradictory, connect them with an OR.

Choose the list of assessment recommendations which is the most complete, or which comes from the most authoritative source.

Choose the list of treatment recommendations which is the most complete, or which comes from the most authoritative source.

These steps deal with the conflicts that may arise during the incremental construction of the prototypes when documents are processed one by one. Other mechanisms of dealing with conflicts will be studied for the future, such as weighing documents based on their pertinence for the domain.

For clinical practice cases, the cases are simply added to the memory, since they all correspond to different, non identifiable, patients.

5 Evaluation

The CaseMiner system was evaluated by comparison with the knowledge-base developed for the FHCRC Carepartner system. This knowledge-base comprises in particular 91 prototypical cases developed through the course of two years. This evaluation deals with the prototypical case mining feature of the system.

The documents processed are all related to the hematopoietic stem-cell transplantation domain. Therefore the papers were selected manually for their pertinence for this domain. Moreover, only the authors from FHCRC were selected, and topics related to diagnosis and treatment, meta analysis articles, and clinical practice guidelines available online. A total of 500 articles were selected from over 5,000 thousand for their pertinence and coverage of the task at hand.

The success of the system is determined by the recall and precision ratios. Precision is the ratio of matching features to the total number of features identified. Recall is the ratio of matching features to the total number of features identified by the manual process. The precision and recall are calculated both at the prototypical case level, and at the level of features matched within each prototype. For example, a match between a finding of JaundiceNOS between the actual LiverChronicGVHD clinical pathway, and the learnt LiverChronicGVHD prototypical case corresponds to a feature match. Results are aggregated by averaging the results across all the prototypical cases and all their features.

The system is evaluated for all the 500 articles. The average values of recall and precision for the documents are shown in Table 1. It shows very encouraging results, even though the process of learning prototypical cases is very complex. These results show that the system definitely learns structures very related with what the experts came up with – and so in a much shorter time (one day of processing versus 2 years with the knowledge elicitation meetings with the team members). Even though the prototypical cases may show some missing information, they provide a sound starting point for case-based reasoning, and will be complemented later on by actual clinical cases, the features of which are more complete and can compensate for the knowledge gaps in the prototypes.

Table 1. Precision and recall ratios

Number of Documents	Prototypical case Recall	Prototypical case Precision	Feature Recall	Feature Precision
500	**95%**	**62%**	**70%**	**65%**

These results also show that if the system does not learn as much as the experts, it would provide an excellent starting point for further refinements with the medical team. The time saved is very significant, and moreover the approach presents major advantages.

Some of the advantages are that the system can be trained continuously and thus could regular update its knowledge base. In fact, an explanation of why the results are somewhat different from the clinical team creation, is that knowledge has evolved since then, and it may simply be that the current prototypes are more current than the ones from the knowledge-base, which date five years now. Another explanation is that not all the articles were processed, and that a more judicious choice or more complete choice would provide better results. Finally, there is no evidence that the prototypes provided by the human experts are always better. What would be an interesting supplemental evaluation would be to compare recommendations from the system for both knowledge bases. Unfortunately, this is not possible in this particular domain at this time, the grant having ended, but will be attempted in other domains.

6 Discussion

Important previous work has been attempted in CBR to either retrieve textual cases [1], or to apply CBR to information retrieval [6]. In reality, the present work does not reason from textual cases, as the other textual CBR systems. A future extension of this work is to apply CBR to biomedical cases described in textual format, where textual CBR will become very pertinent. Pertinent previous work relates to case mining, feature mining, and prototype mining. These are addresses in successive paragraphs.

Case mining refers to the process of mining potentially large data sets for cases [27]. Researchers have often noticed that cases simply do not exist in electronic format, that databases do not contain well-defined cases, and that the cases need to be created before CBR can be applied. Another option is to start CBR with an empty case base. When large databases are available, preprocessing these to learn cases for future CBR permits to capitalize on the experience dormant in these databases. [27] propose to learn cases by linking several database tables.

Feature mining refers to the process of mining data sets for features. Many CBR systems select the features for their cases, and/or generalize them. [26] for example focus on dimension reduction and/or feature selection, and shows that this improves the classification and CBR accuracy. In biomedical domains, in particular when data vary continuously, the need to abstract features from streams of data is particularly prevalent. Recent, and notable, examples include [14, 20] who reduce their cases time series dimensions through *Discrete Fourier Transform*. [19] propose an original method for generalizing features, using both *clustering* techniques to group the cases into clusters containing similar cases, and feature selection techniques.

Generalized case mining refers to the process of mining databases for generalized and/or abstract cases. Generalized cases are called in varied ways, such as prototypical cases, abstract cases, prototypes, stereotypes, templates, classes, categories, concepts, and scripts. Although all these terms refer to slightly different concepts, they represent structures that have been abstracted or generalized from real cases either by the CBR system, or by an expert. Examples of such systems include systems described in [2, 12, 13, 21, 22].

Finally, many authors learn *concepts*, and refer to conceptual clustering as their learning methodology [25]. [7] use *formal concept analysis* (FCA) – a mathematical method from data analysis - as another induction method for extracting knowledge from case bases, in the form of *concepts*.

The abundance of literature in case, feature, and prototype mining shows that this question is essential to CBR, as a machine learning methodology. CaseMiner is mostly related to case mining, but differs from previous approaches [27] by mining from literature. It does not mine from textual cases as in [26]. One of the main advantages of the method proposed here is that it will facilitate the bootstrapping of CBR systems in biomedicine by providing a starting case base of mostly prototypical cases, which will render the methodology readily applicable to a much wider range of domains, in particular those where electronic cases are not available, like Carepartner [3].

7 Conclusion

Case mining from medical literature is a very promising approach to building case bases. It has the potential of spreading the development of CBR systems in many domains where either electronic cases are not available, or they are incomplete, which is most frequent, or experts are not available for months or years of knowledge elicitation tasks. Moreover, it offers new opportunities for updating case bases from recent medical advances, and for leveraging multiple domain CBR. Research ahead in this direction involves automatically selecting the body of documents the most adequate for feeding the case mining system, detecting out-of-scope documents automatically, studying performance improvement and stability of the system, learning more complex case structures and features, combining case mining from databases and from literature, and studying the knowledge discovering process in itself from both the case-based approach, the rule-based approach, and the model-based approach.

References

1. Ashley, K.D., Lenz, M. (eds.),: Textual Case-Based Reasoning. In: AAAI-98 Workshop, Technical Report WS-98-12, AAAI Press, Menlo Park, CA, 1998.
2. Bellazzi R., Montani S., Portinale L.: Retrieval in a Prototype-Based Case Library: A Case Study in Diabetes Therapy Revision. In: Smyth, B., Cunningham, P. (eds.): Proceedings of ECCBR 98. Lecture Notes in Artificial Intelligence, Vol. 1488. Springer-Verlag, , Berlin, Heidelberg, New York (1995) 64-75
3. Bichindaritz, I.: Mémoire: Case-based Reasoning Meets the Semantic Web in Biology and Medicine. In: Funk, P., Gonzàlez Calero, P.A. (eds.): Proceedings of ECCBR 2004. Lecture Notes in Artificial Intelligence, Vol. 3155. Springer-Verlag, Berlin, Heidelberg, New York (2004) 47-61

4. Bichindaritz I., Akineni S.: Concept Mining from Biomedical Literature. In: Perner, P., Imiya, A. (eds.): Proceedings of MLDM 05. Lecture Notes in Artificial Intelligence, Vol. 3587. Springer-Verlag, Berlin, Heidelberg, New York (2005) 682-691
5. Bichindaritz I.: Name Relationship Mining from Biomedical Literature. In: Perner, P., Imiya, A. (eds.): Proceedings of ICDM 06. Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg, New York (2006) (in press)
6. Daniels, J.J., Rissland, E.L.: A Case-Based Approach to Intelligent Information Retrieval. In: Proceedings of SIGIR 95, ACM Press, New York, NY, (1995) 238-245
7. Dìaz-Agudo, B., Gonzàlez-Calero, P.: Classification Based Retrieval Using Formal Concept Analysis. In: Aha, D., Watson, I. (Eds.): Proceedings of ICCBR 01, Lecture Notes in Artificial Intelligence 2080, Springer-Verlag, Berlin, Heidelberg, New York (1995) 173-188
8. Dorre, J., Gerstl, P., Seiffert, R.: Text mining: finding nuggets in mountains of textual data. In: Chaudhuri, S., Madigan, D., and Fayyad, U. (eds.): Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM press, New York (1999) 398-401
9. Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A.: GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17, Suppl 1 (2001) S74-S82
10. Fuller, S., Revere, D., Bugni, P., Martin, G.M.: A knowledgebase system to enhance scientific discovery: Telemakus. Biomed Digit Libr. Sep 21;1(1):2 (2004)
11. Hearst, M.A.: Untangling Text Data Mining. In: Dale, R., Church, K. (eds.): Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, (1999) 3-10
12. Malek M., Rialle, V.: A Case-Based Reasoning System Applied to Neuropathy Diagnosis. In: Keane, M., Haton, J.-P., Manago, M. (eds.): Proceedings of EWCBR 94. Acknosoft Press, Paris, (1994) 329-336
13. Maximini, K., Maximini, R., Bergmann, R.: An Investigation of Generalized Cases. In: Ashley, K.D., Bridge, D.G. (eds.): Proceedings of ICCBR 03. Lecture Notes in Artificial Intelligence 2689, Springer-Verlag, Berlin, Heidelberg, New York (2003) 261-275
14. Montani, S., Portinale, L., Bellazzi, R., Leornardi, G.: RHENE: A Case Retrieval System for Hemodialysis Cases with Dynamically Monitored Parameters. In: Funk, P., Gonzàlez Calero, P. (eds.): Proceedings of ECCBR 04. Lecture Notes in Artificial Intelligence 3155, Springer-Verlag, Berlin, Heidelberg, New York (2004) 659-672
15. Nasukawa T., Nagano, T.: Text Analysis and Knowledge Mining System. Knowledge management Special Issue. IBM systems journal Vol. 40 (2001) 967-984
16. National Library of Medicine: The Specialist NLP Tools. http://specialist.nlm.nih.gov [Last access: 2005-04-01] (2004)
17. National Library of Medicine: MetaMap Transfer (MMTx), http://mmtx.nlm.nih.gov [Last access: 2005-04-01] (2005)
18. National Library of Medicine: The Unified Medical Language System. http://umls.nlm.nih.gov [Last access: 2005-04-01] (2005)
19. Niloofar, A., Jurisica, I.: Maintaining Case-Based Reasoning Systems: A Machine Learning Approach. In: Funk, P., Gonzàlez Calero, P. (eds.): Proceedings of ECCBR 04. Lecture Notes in Artificial Intelligence 3155, Springer-Verlag, Berlin, Heidelberg, New York (2004) 17-31
20. Nilsson, M., Funk, P.: A Case-Based Classification of Respiratory sinus Arrhythmia. In: Funk, P., Gonzàlez Calero, P. (eds.): Proceedings of ECCBR 04. Lecture Notes in Artificial Intelligence 3155, Springer-Verlag, Berlin, Heidelberg, New York (2004) 673-685
21. Perner, P.: Different Learning Strategies in a Case-Based Reasoning System for Image Interpretation. In: Smyth, B., Cunningham, P. (eds.): Proceedings of ECCBR 98. Lecture Notes in Artificial Intelligence, Vol. 1488. Springer-Verlag, , Berlin, Heidelberg, New York (1995) 251-261
22. Yang, Q., Hong, C.: Case Mining from Large Databases. In: In: Ashley, K.D., Bridge, D.G. (eds.): Proceedings of ICCBR 03. Lecture Notes in Artificial Intelligence 2689, Springer-Verlag, Berlin, Heidelberg, New York (2003) 691-702

23. Schmidt, R., Gierl, L.: Experiences with Prototype Designs and Retrieval Methods in Medical Case-Based Reasoning Systems. In: Smyth, B., Cunningham, P. (eds.): Proceedings of ECCBR 98. Lecture Notes in Artificial Intelligence, Vol. 1488. Springer-Verlag, , Berlin, Heidelberg, New York (1995) 370-381
24. Swanson, D.R.: Information discovery from complementary literatures: Categorizing viruses as potential weapons. Journal of the American Society for Information Science Vol. 52(10) (2001) 797-812
25. Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence Vol.9 (1997), 183-203
26. Wilson, D.C., Leake, D.B.: Mainting Case-Based Reasoners: Dimensions and Directions. Computational Intelligence Journal, Vo. 17, No. 2, May 2001
27. Wiratunga, N., Koychev, I., Massie, S.: Feature Selection and Generalisation for Retrieval of Textual Cases. In: Funk, P., Gonzàlez Calero, P. (eds.): Proceedings of ECCBR 04. Lecture Notes in Artificial Intelligence 3155, Springer-Verlag, Berlin, Heidelberg, New York (2004) 806-820

Detection of Hygiene-relevant Parameters from Cereal Grains based on Intelligent Image Interpretation and Data Mining

Petra Perner and Thomas Günther*

Institute of Computer Vision and Applied Computer Sciences, IBaI
04107 Leipzig, Körnerstr. 10,
pperner@ibai-institut.de, www.ibai-institute.de
* JenaBios GmbH, D-07749 Jena, Loebstedter Str. 78

Abstract. We are going on to develop a novel method for the detection of hygiene-relevant parameters from grains of cereal crops based on intelligent image acquisition and interpretation methods as well as data mining methods. We present our first case study that describes the data acquisition, the planned image analysis and interpretation method as well as the reasoning methods that can map the automatic acquired parameters of grain to the relevant hygiene parameters. The preliminary results show that with the new computer science methods it is possible to come up with new insights into the quality control of food stuff.

1 Introduction

Fungal contamination of cereals is a serious economic problem throughout the world. Several fungi cause a reduction of grain quality, especially changes in color and taste [Müller et al., 1997], [Herrman et al, 1998], and [Rodeman, 2003]. However the main riks of fungal damage arise from the production of toxic compounds, known as mycotoxins. Mycotoxins can cause serious adverse health effects. Toxigenic fungi that produce mycotoxins in grains of cereals or oil seeds belong to the genera Aspergillus, Alternaria, Fusarium and Penicillium. The control of this problem is therefore of particularly interest in food safety and quality control programs.

The aim of the research is the development of an automatic image acquisition and image interpretation system for the fast recognition of cereal grains damaged by fungi. Thereby should be developed a data acquisition unit that allows taking the coverage from the grain and allows to place it under a microscope for the acquisition of a digital image. This image should be used in order to automatically determine the number and the kind of fungi spores contained on the grain. For that we have to develop suitable intelligent image analysis and interpretation methods. Based on the enumeration of fungal spore classes we have to develop a method that can map this information to the hygiene-relevant parameters. The work we present here reports the results of our first case study. They show that the proposed methods based on intelligent image analysis and data mining are very suitable to capture the desired information and allow recognizing formerly unknown information that can be helpful to determine the quality of food stuff.

2 Material

For the study have been used different quality classes of wheat grains:

1. visual optical perfect grains from a charge where no fungal grains were included,
2. fungal damaged grains,
3. gall-mosquito damaged grains, and
4. visual optical perfect grains taken from a charge of fungal damaged grains.

In total we had 10 samples from each class. Thirty single grains were taken from each sample for further evaluation.

3 Image Data Acquisition

The main problem was to make the coverage on the grains visible under the microscope and make it usable for further digital processing. Therefore we have developed a procedure for taking the coverage from grains and bring it onto a medium that can be placed under a microscope. From there can be acquired a digital image with the help of a digital camera connected with the microscope.

The method of choice was a water-based extraction method. The grains were placed into a boil together with stones. This water-filled boil was shaken for 2 minutes, then the water was filled into a centrifuge and the sediment was put on a slide. This slide was placed under the microscope and a digital image was taken. There are other methods for extracting the coverage from the grain possible but this should not be the main topic of this paper. The resulting digital images are shown in Figure 1a-4a.

4 Intelligent Image Analysis and Interpretation

4.1 Image Analysis

The main aim of the image analysis was to recognize possible fungi spores and process them further for determination of the type of fungi spore. Here we used our novel case-based object recognition method [Perner et al., 2005] developed for recognizing biological objects with high variation. For the architecture of such a system see Figure 5. The case-based object recognition method uses cases that generalize the original contour of the objects and matches these cases against the contour of the objects in the image. During the match a score is calculated that describes the goodness of the fit between the object and the case. Note the result of this process is not the information about what type of fungi spore is contained in the image. The resulting information tells us only if it is highly likely that the considered object is a fungi spore or not. Further evaluation is necessary to determine the kind of fungi spore. This demonstrates the result in the images, see Figure 1b-4b. One of the main problems of such a case-based object recognition method is to fill up the case base with a sufficient large enough number of cases. We used our procedure described in [Perner et al., 2004] for that. For the study we have 10 different cases, which is not enough as we can see in the image but it allows us to demonstrate the applicability of the method. The method has to be adapted to the specific image quality to show better results as well as more cases have to be learnt by our case acquisition procedure.

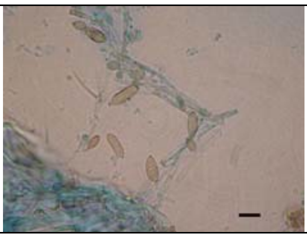
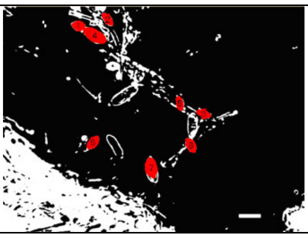
	

Fig. 1a Coverage of Grain with Cladosporium	**Fig. 1b** Segmented Image

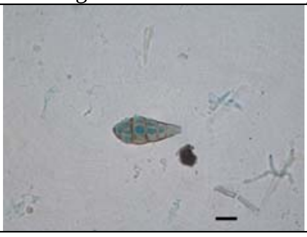
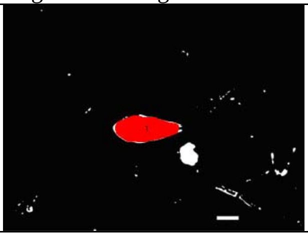

	

Fig. 2a Coverage of Grain with Alternaria Alternata	**Fig. 2b** Segmented Image

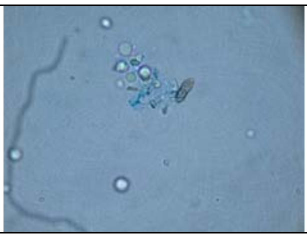
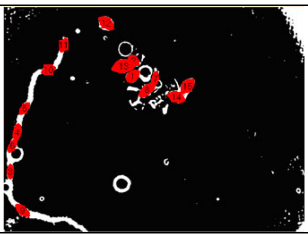

	

Fig. 3a Coverage of Grain	**Fig. 3b** Segmented Image

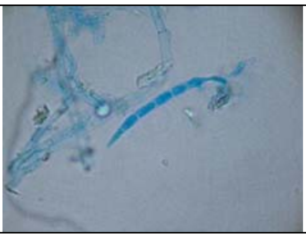
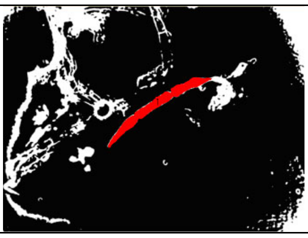
	

Fig. 4a Coverage of Grain with Fusarium Spore	**Fig. 4b** Segmented Image

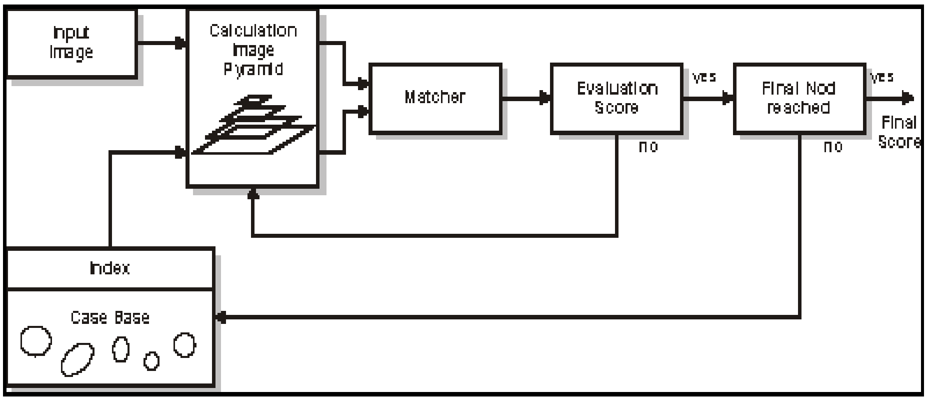


Fig. 5 Architecture of a Case-Based Object Recognition System

4.2 Image Interpretation and Data Mining

After the methods have recognized potential objects that are likely to be fungi spores we have to extract more features from the objects that distinguish the object from the background and different fungi spores. Of course one feature is already the shape information used in the matching process but that is not enough for more detailed recognition. The features that have to be calculated for this kind of objects are the inner structure, texture and gray level information. We haven't done that for this kind of objects considered in this publication yet. But we know from our past research on airborne fungi that it is possible to find automatic extractable features to describe fungi spores and use them for classification into different kinds of fungi spores. It is left to future work to find the right features for the considered fungi spores in this application and to build the feature extraction procedure for them. Based on this feature set we can construct the classifier. We use decision tree induction based on our tool Decision Master [Perner, 2003]. This gives us a good classifier.

As the result we will get the information about the kind of fungi spores contained in the image and the number of fungi spores versa the kind of fungi spores.

5 Mapping of Image Information to Hygiene Relevant Parameter with Data Mining

In this study the kind and the number of fungi spores was determined manually since it was a case study and we haven't developed the fully automatic system yet. The aim of the study was to figure out if the proposed methods can bring out information about hygiene-relevant parameters and besides that new information that can be used to control the quality of food stuff. From the 4x10 different samples a data base was created where the columns of each entry show the class, that is the optical visual inspection label, the number of Fusarium spores, the number of Alternaria/Ulocladium, the number of Aspergillus/Penicillium, the number of Cladosporium, the number of fungi spores with unknown classification and the total number of fungi spores. In addition to the enumeration of fungal spores the concentration of a main mycotoxin of the genus Fusarium deoxynivalenol (DON) was determined by a commercial enzyme immunoassay screening (ELISA test).

Table 1-4 shows that there is a significant difference in the number and the kind of fungi spores for the different charges. Figure 6 shows that DON value corresponds to the visually determined class labels. Grain with a low number of Fusarium spores have low DON values and grain charges with high number of Fusariam spores have high DON values.

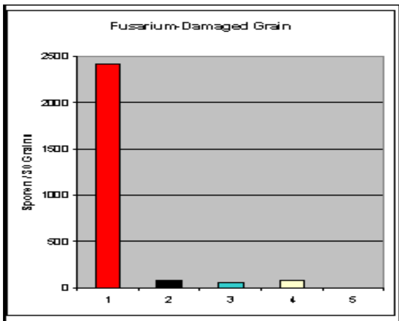


Table 1 Fusarien Damaged Graind

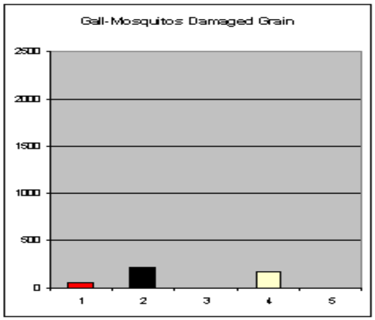


Table 2 Gall-Mosquitos Damaged Grain

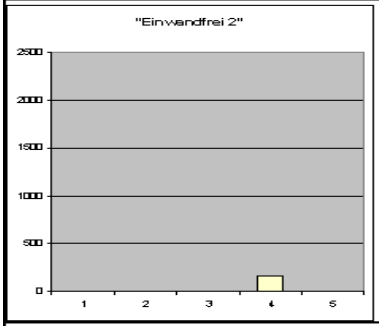


Table 3 Charge "einwandfrei 2"

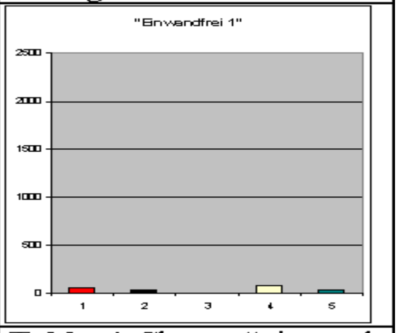


Table 4 Charge "einwandfrei 1"

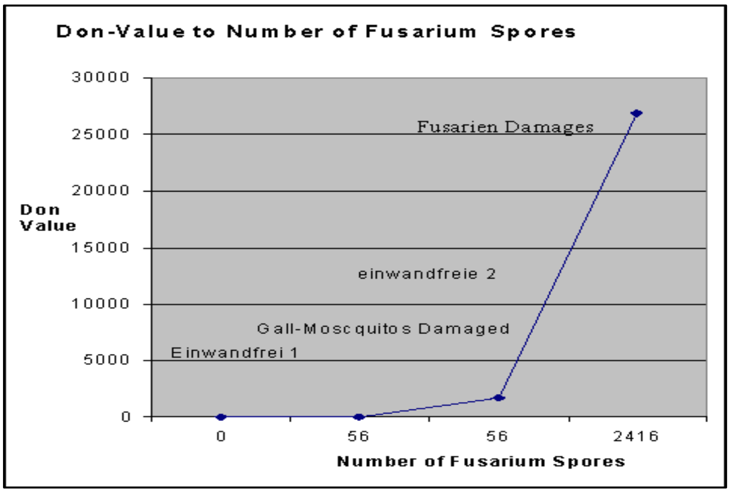


Fig. 6 Don Value to Number of Fusarium Spores

Decision tree induction with Decision Master [Perner, 2003] on an entropy-based criterion was performed in order to find out the relation between the coverage of fungi spores and the class label (mycotoxin value). The induction experiment shows that there is a relation between the number of Cladosporium spores and Fusarium spores respective the class, see Figure 7. It says that grain charges with a high number of Cladosporium spores will have a low number of Fusarium spores. That means these charges are either perfect charges or gall-mosquitoes damaged charges. Whereas charges with low Cladosporium spores can be either charges with a high number of Fusarium spores or a low number of Fusarium spores. Note that charge “einwandfrei 2” (visual perfect grains) has been taken out from a sample with Fusarium damaged grains. It seems that the number of Cladosporium spores indicates this fact. The number of Alternaria and Aspergillus spores did not have a significant influence in this experiment.

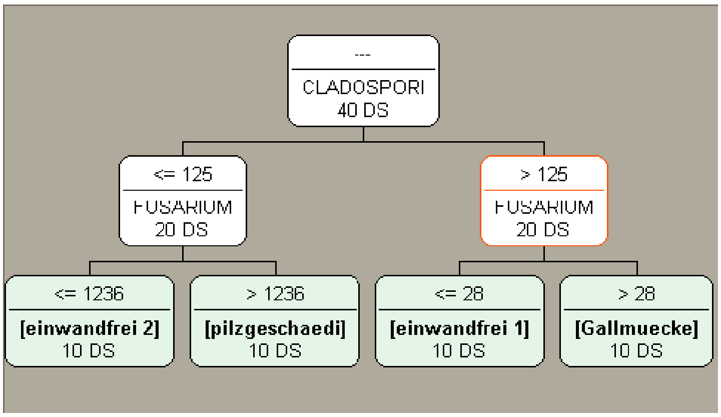


Fig. 7 Decision Tree for the Determination of Grain Quality based on Number and Type of Fusarium Spores

6 Conclusions

We have presented our first results on our case study for the detection of hygiene-relevant parameters from cereal grains based on intelligent image acquisition and interpretation methods as well as data mining method. It is a joint work between a computer scientist, food experts and microbiologists. We have shown that data acquisition is an important task and that it has to do with more than data base construction as it is in many data mining experiments. The image acquisition method we have demonstrated in this paper works well and can be fully automated. It can also be constructed in such a way that the coverage from each single grain can be taken off and evaluated based on the intelligent image interpretation and data mining methods. The image analysis on case-based object recognition works well for this task but has to be tuned so that a better object recognition rate can be achieved. From each single object can be extracted image features and these features can be used for classification. It is preferable to construct the classifier based on decision tree induction methods. Once the type and number of fungi spores has been determined this information can be set into relation with the hygiene-relevant parameters. We have shown that the number of Fusarium spores correlates with the DON levels which is a value used for the determination of the mycotoxin concentration. However when considering this experiment as a data mining experiment and applying decision tree induction to the created data base some other important information can be extracted which are more or less hidden before. The next steps of our work will be to improve the image interpretation methods. When we have a fully automatic algorithm we will apply our method to a large number of grain samples. The aim is to come up with a new measurement method for the determination of hygiene-relevant parameters on grains. Besides that we would like to discover formerly unknown relations or information based on the material in the coverage of the grain such as different types of fungi spores.

Acknowledgement

We like to thank the researchers from the Bundesanstalt für Bundesforschungsanstalt für Ernährung und Lebensmittel, Institut für Getreide-, Kartoffel- und Stärketechnologie, in particular Prof. Lindhauer for providing some of the examples of the material and for encouraging advices and useful discussions on the direction of the research.

References

[Müller et al., 1997] Müller, H.M. et al. (1997): Fusarium toxins in wheat harvested during six years in an area of Southwest Germany. Natural Toxins 5, 25-30

[Herrman et al,, 1998] Hermann, W.; Kübler, E. und Aufhammer, W. (1998): Ährenbefall mit Fusarien und Toxingehalt im Korngut bei verschiedenen Wintergetreidearten. Pflanzenwirtschaften 2, 97-107

[Rodeman, 2003] Rodeman, B. (2003): Auf resistente Sorten setzen. DLG Mitteilungen 3, 44-46

[Perner *et al*., 2005] Petra Perner, Silke Jähnichen, and Horst Perner. *Case-Based Object Recognition for Airborne Fungi Recognition.* Intern. Journal on Artificial Intelligence in Medicine, to appear 2005

[Perner, 2003]. Petra Perner. Data Mining on Multimedia Data, Springer Verlag 2003

[Perner et al., 2004] Petra Perner and Silke Jähnichen. *Case Acquisition and Case Mining for Case-Based Object Recognition*, In: Peter Funk, Pedro A. González Calero (Eds.), Advances in Case-Based Reasoning, Proceedings of the ECCBR 2004, Madrid/Spain, Springer Verlag 2004, Vol. 3155, pp. 616-629

A new image segmentation algorithm based on kernel spatial fuzzy c-means

Weixing Wang[1], BingCui[1]

[1] Computer department of sicence and technology, Chongqing University of Posts and Telecommunications,
400065 Chongqing, China
wangwx@cqupt.edu.cn

Abstract. Fuzzy c-means clustering with spatial constraints (FCM_S) can consider spatial information of images. But it lacks of enough robust to noise and outliers. And computation complexity of FCM_S is high. In this paper, a new objective function is presented. Weighted average pixel value of pixel x_k ($\bar{x}_k$) is substituted for x_k in FCM. $\bar{x}_k$ denotes an integrative value of x_k and its neighbor pixels. So the impact of noise or outliers will be reduced in the process of clustering. Then image pixels are mapped from the original space into a higher dimensional feature space by using Mercer kernel functions. As a result, c-means clustering can be performed efficiently in the feature space for that kernel functions can induce robust distance measures while the computational complexity is low. Some experiments are conducted on images with or without noise. The results show that this algorithm is suitable and robust for these images, and the running time decreased sharply.

Keywords: image; segmentation; fuzzy c-means clustering; spatial; kernel

1 Introduction

Image segmentation plays an important role in many fields such as Image processing、Machine vision [1]. The results of segmentation determine the results of image analysis and process. And image segmentation is one of classical problems in the field of image process. There are more than one thousand algorithms about it. Many scholars have classified them [2]. But there is not a common used method that can segment all kinds of images efficiently.

Fuzzy c-means clustering is a method that can classify n samples into c clusters [3]. In last decades, it is wildly used for image segmentation [4]. Image clustering is to partition pixels into c-clusters, so members of same cluster are more similar to one another than to members of another cluster. Jianwu Liu et al. have used this method to segment MRI brain tumor images [5]. Although FCM algorithm can do well in segmenting most noise-free images, it fails to segment images with noise and outliers. It may lead to nonrobust results due to the use of nonrobust Euclidean distance and disregard of spatial information of images. As a result, many scholars have did some research in these problems. For the first problem, R. J. Hathaway and J.C Bezdek proposed that fuzzy c-means clustering uses L_p distance norm not L_2 norm in FCM objective function when the distance between one sample and another one is calculated [6]. For the second problem, D. L.Pham [7] and M. N. Ahmed et al. [8] presented that spatial information should be incorporated into original FCM objective function. Meanwhile, the introduction of L_p norm distance and spatial constraints increases computational complexity. And the robustness is still poor.

The rest of this paper is organized as follows: In section 2, we introduce fuzzy c-means clustering. In section 3, spatial information is incorporated into fuzzy c-means clustering. In section 4, we describe the kernel function. Kernel-based fuzzy clustering with spatial constraints (KFCM_S) is derived. A new objective function is presented in section 5. We use above algorithms to segment colony images in section 6, and give conclusions in section 7.

2 Fuzzy c-means Clustering

Assume that there are N pixels in a colony image. Let these N pixels be the dataset $X = \{x_1, x_2, \cdots, x_N\}$. Every pixel sample has three attributes, i.e., the values of R、G、B. If the segmentation results need to be very good, many other attributes, such as texture information、grads information, can be added to.

In order to partition these N pixels into c clusters, we just need to minimize the value of objective function $J_m = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^m \| x_k - v_i \|^2$ mathematically. Where $\| x_k - v_i \|$ means Euclidean distance d_{ki}; v_i is the centroid of the i th cluster; And the array $U = \{u_{ik}\}$ is the fuzzy partition matrix whose element u_{ik} denotes the membership of the k th pixel to the i th cluster. U satisfies $\sum_{i=1}^{c} u_{ik} = 1, k = 1,2\cdots N$ and $0 < \sum_{k=1}^{N} u_{ik} < N, i = 1,2\cdots c$; m denotes a weighting exponent on each fuzzy membership. The minimizing value of J_m can be derived by evaluating the centroids v_i and membership functions u_{ik} that satisfy a zero gradient condition. So, we can obtain that:

$$u_{ik} = \frac{\left(\| x_k - v_i \|^2\right)^{-\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\| x_k - v_j \|^2\right)^{-\frac{1}{m-1}}}, i = 1,2\cdots c, k = 1,2\cdots N,$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^m x_k}{\sum_{k=1}^{N} u_{ik}^m}, i = 1,2,\cdots c.$$

3 Fuzzy c-means Clustering with Spatial Constraints

M. N. Ahmed, S. M. Yamany, N. Mohamed proposed a modification to FCM [8]. They introduced a term that allows the labeling of a pixel to be influenced by the labels in its immediate neighborhood. In other words, the modification considers spatial constraints and aims at keeping continuity on neighboring pixel values around a pixel. Such a regularization is useful in segmenting scans corrupted by salt and pepper noises. The modified objective function is given by

$$JS_m = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^m \| x_k - v_i \|^2 + \frac{\alpha}{N_R}\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^m\left(\sum_{x_r \in N_K} \| x_r - v_i \|^2\right)$$

Where N_k denotes the set of neighbors that exit in a window around x_k; N_R is the cardinality of N_k. The effect of the neighbor term is controlled by the parameter α. The relative importance of the regularizing term is inversely proportional to the signal-to-noise ratio (SNR) of colony images. A lower SNR requires a higher value of α.

The objective function JS_m can be minimized under the constraint of U as stated in section 3. Similarly, we can get the following equation:

$$u_{ik} = \frac{\left(\| x_k - v_i \|^2 + \frac{\alpha}{N_R}\sum_{r \in N_k} \| x_r - v_i \|^2\right)^{-\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\| x_k - v_j \|^2 + \frac{\alpha}{N_R}\sum_{r \in N_k} \| x_r - v_j \|^2\right)^{-\frac{1}{m-1}}}, i = 1, 2\cdots c, k = 1, 2\cdots N$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^m (x_k + \frac{\alpha}{N_R} \sum_{r \in N_k} x_r)}{(1+\alpha)\sum_{k=1}^{N} u_{ik}^m}, i = 1,2,\cdots c\,.$$

4 Kernel-based Fuzzy c-means Clustering with Spatial Constraints

4.1 Kernel Function

The success of support vector machine has greatly drawn people's attention in kernel function. The spirit of kernel function is a mapping from an input space to a feature space with higher dimension. If the operation of mapped vectors is only inner product, we need not to know the detailed mapping form. We can reduce the operation and overcome curse of dimensionality by substituting kernel function for inner product. Every liner algorithm that only uses inner products can be easily extended to a nonlinear version only through the kernels satisfying Mercer's conditions [9].

Assume that samples in colony image input space are $x_i, i = 1,2,\cdots,N$, these pixel samples are mapped into a feature space by a nonlinear mapping Φ. Then we derive $\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_N)$. Inner product in input space can be denoted by Mercer kernel in feature space, i.e. $K(x_i, x_j) = (\Phi(x_i) \bullet \Phi(x_j))$ [10].

There are some common used Mercer kernel functions:

Polynomial kernel: $K(x, y) = (x \bullet y + 1)^d$, where d is a user-defined parameter.

Radial basis function: $K(x, y) = \exp\left(\frac{-\left(\sum_{i=1}^{d} | x_i - y_i |^a \right)^b}{\sigma^2} \right)$, where d is the dimension of vectors x, $a \ge 0$, $1 \le b \le 2$. If $a = 2$, $b = 1$, we can obtain:

Gaussian kernel: $K(x, y) = \exp\left(- \beta \| x - y \|^2 \right), \beta > 0$ is a user-defined parameter.

Some more kernel functions can be found in [11]. In this paper, we use the Gaussian kernel for its robustness [12].

4.2 Kernel-based Fuzzy c-means Clustering with Spatial Constraints

We can construct the kernelized version of fuzzy c-means clustering with spatial constraints algorithm and modify its objective function by the mapping Φ as follows:

$$\|\Phi(x_k)-\Phi(v_i)\|^2=(\Phi(x_k)-\Phi(v_i))^T(\Phi(x_k)-\Phi(v_i))$$
$$=\left(\Phi(x_k)^T\Phi(x_k)\right)-\left(\Phi(v_i)^T\Phi(x_k)\right)-\left(\Phi(x_k)^T\Phi(v_i)\right)+\left(\Phi(v_i)^T\Phi(v_i)\right)$$
$$=K(x_k,x_k)-K(v_i,v_i)-2K(x_k,v_i)$$
$$=2(1-K(x_k,v_i))$$

Similarly, $\|\Phi(x_r)-\Phi(v_i)\|^2=2(1-K(x_r,v_i))$.

Thus,

$$JS_m^{\Phi}=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^m(1-K(x_k,v_i))+\frac{\alpha}{N_R}\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^m\left(\sum_{x_r\in N_K}(1-K(x_r,v_i))\right)$$

Parameters used in JS_m^{Φ} stand for the same meaning as they are used in JS_m .

And,

$$u_{ik}=\frac{\left((1-K(x_k,v_i))+\frac{\alpha}{N_R}\sum_{r\in N_k}(1-K(x_r,v_i))\right)^{-\frac{1}{m-1}}}{\sum_{j=1}^{c}\left((1-K(x_k,v_j))+\frac{\alpha}{N_R}\sum_{r\in N_k}(1-K(x_r,v_j))\right)^{-\frac{1}{m-1}}},i=1,2\cdots c,k=1,2\cdots N$$

$$v_i=\frac{\sum_{k=1}^{N}u_{ik}^m\left(K(x_k,v_i)x_k+\frac{\alpha}{N_R}\sum_{r\in N_k}(K(x_r,v_i)x_r)\right)}{\sum_{k=1}^{N}u_{ik}^m\left(K(x_k,v_i)+\frac{\alpha}{N_R}\sum_{r\in N_k}K(x_r,v_i)\right)},i=1,2,\cdots c$$

The algorithm can be presented in the following steps:

(1) Fix the number c of these clusters and initial centroids, set ε >0 to a very small value;

(2) Calculate the value of u_{ik} , and update the partition matrix U ;

(3) Calculate the value of v_i , and update the centroids V ;

(4) Repeat steps (2) 、 (3) until the termination criterion is statisfied : $\|V_{new}-V_{old}\|<\varepsilon$.

We segmented some colony images by KFCM_S. Fig.1 is a colony image, Fig.2 is the segmented image by KFCM_S. And Fig.3 is a colony image with salt and pepper noise, Fig.4 is its' segmented image. Here the parameters c =2, ε =0.001, *m*=2.0, α =2.0, β =0.00002. From the results, we can see that colony parts can be segmented from the whole image well and the algorithm performs well on the image with salt and pepper noise.

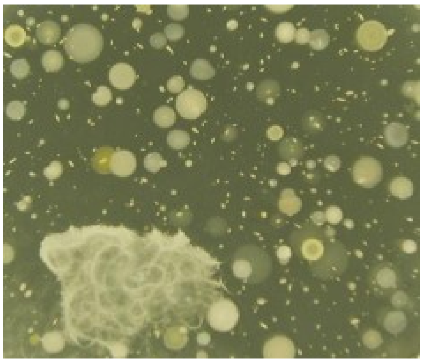


Fig. 1. Colony image with few noise

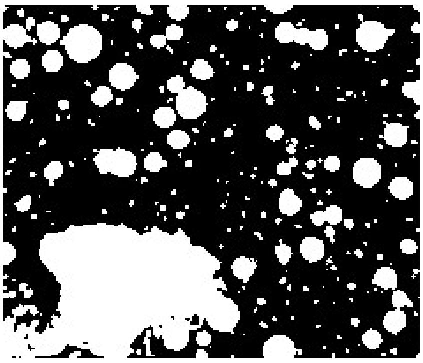


Fig. 2. Segmented result by KFCM_S

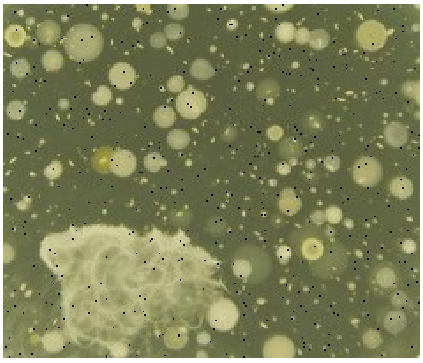


Fig. 3. Colony image with some salt and pepper noise

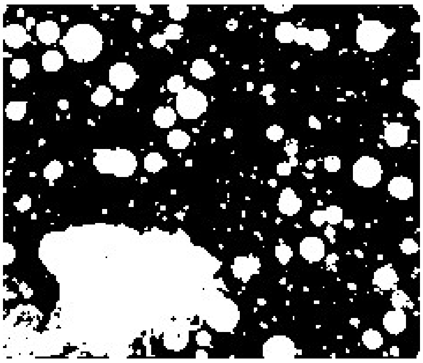


Fig. 4. Segmented result by KFCM_S

5 Fast Fuzzy c-means Clustering with Spatial Constraints

Even though KFCM_S performs well on normal images or images with noise, it costs too much time. Fig.1 and Fig.3 are images with 236×201 pixels. The running time of KFCM_S is about 68.5 seconds. In order to decrease the running time at the same time spatial information is considered, we present a new objective function $JS_m = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m} \| \bar{x}_k - v_i \|^2$, $\bar{x}_k = \dfrac{w_k \cdot x_k + \sum_{r \in N_k} w_{kr} \cdot x_{kr}}{w_k + \sum_{r \in N_k} w_{kr}}$ is a weighted average value. Where w_k and w_{kr} are weight values, N_k denotes the set of neighbors that exit in a window around x_k , x_{kr} denotes the *r*th pixel in the neighbor set. Through the weighted average value, spatial information can be considered. If there are a few of noises in the image, w_k needs to be much larger and w_{kr} need to be much smaller. In this case, the pixel x_k belongs to which class mainly depends on x_k . While there are much noises in the image, w_k needs to become smaller and w_{kr} need to become larger. So the pixel x_k belongs to which class depends on the integrative information of x_k and its neighbors.

Similarly, $$u_{ik} = \frac{\left(\| \bar{x}_k - v_i \|^2 \right)^{-\frac{1}{m-1}}}{\sum_{j=1}^{c} \left(\| \bar{x}_k - v_j \|^2 \right)^{-\frac{1}{m-1}}}, i = 1, 2 \cdots c, k = 1, 2 \cdots N,$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^m \bar{x}_k}{\sum_{k=1}^{N} u_{ik}^m}, i = 1, 2, \cdots c.$$

Thus, the kernel form of objective function is $JS_m^\Phi = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^m \left(1 - K\left(\bar{x}_k, v_i \right) \right)$,

and $$u_{ik} = \frac{\left(1 - K(\bar{x}_k, v_i) \right)^{-\frac{1}{m-1}}}{\sum_{j=1}^{c} \left(1 - K(\bar{x}_k, v_j) \right)^{-\frac{1}{m-1}}}, i = 1, 2 \cdots c, k = 1, 2 \cdots N \quad ,$$

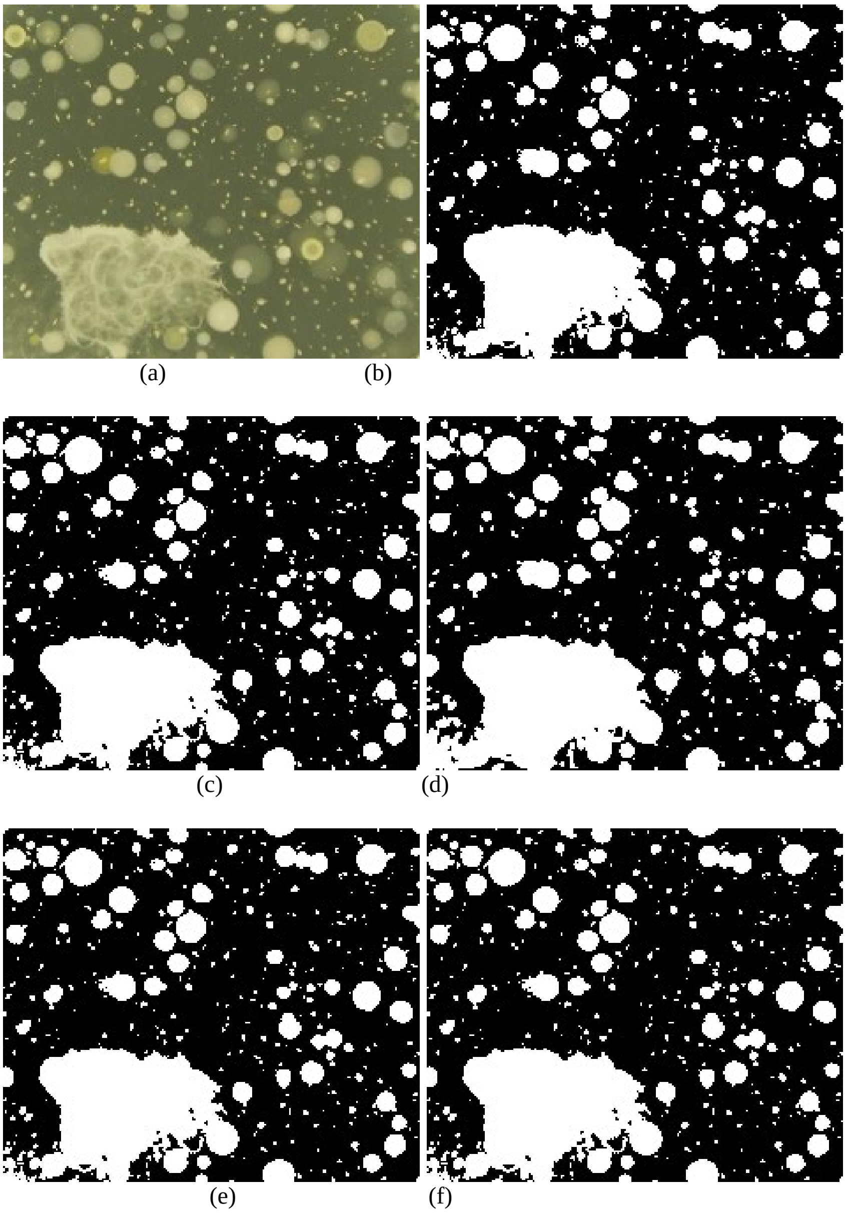
$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^m K\left(\bar{x}_k, v_i \right) \bar{x}_k}{\sum_{k=1}^{N} u_{ik}^m K\left(\bar{x}_k, v_i \right)}, i = 1, 2, \cdots c.$$

6 Results

There are two groups of images. Images of group 1 are the colony image with few noise and its segmented images by BCV、FCM、KFCM_S、FFCM_S、FKFCM_S, while images of group 2 are the colony image with salt and pepper noise and its segmented images.

Here the colony images have 236×201 pixels. When the colony image with few noise is processed, parameters c =2; ε =0.001; m=2.0; α =2.0; β =0.00002; N_R =4; w_k =4; $w_{kr} = 1, r = 1, 2, 3, 4$. While the colony image with salt and pepper noise is processed, c=2; ε =0.001;

m=2.0; α =2.0; β =0.00002; N_R =4; w_k =1; $w_{kr} = 1, r = 1, 2, 3, 4$.


(a) (b)

(c) (d)

(e) (f)

Fig. 5. Colony image segmentation example. (a) Original colony image. (b) BCV result. (c) FCM result. (d) KFCM_S result. (e) FFCM_S result. (f) FKFCM_S result.

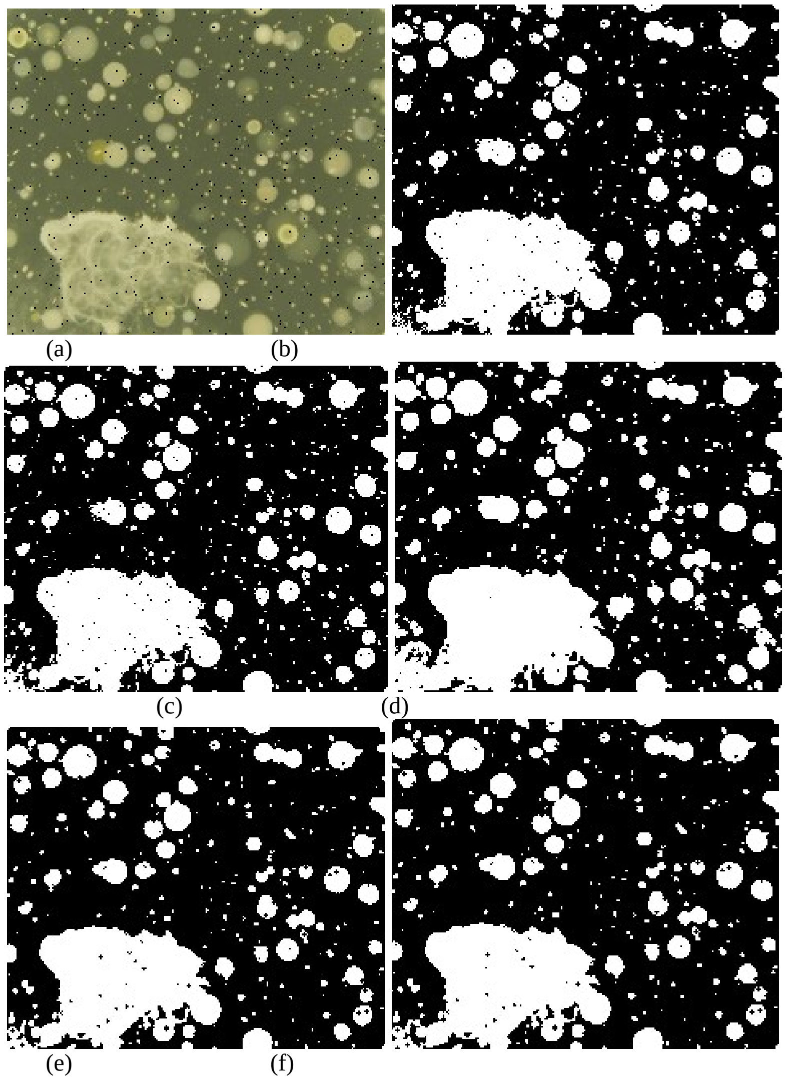


Fig. 6. Colony image with salt and pepper noise segmentation example. (a) Original colony image with salt and pepper noise. (b) BCV result. (c) FCM result. (d) KFCM_S result. (e) FFCM_S result. (f) FKFCM_S result.

Table 1. Running time of algorithm FCM、FCM_S、KFCM_S、FFCM_S、FKFCM_S on two colony images(*ms*).

	FCM	FCM_S	KFCM_S	FFCM_S	FKFCM_S
Original colony image	1000	3543	69086	2633	3894
Original image with salt and pepper noise	1036	3684	68680	2781	3883

7 Conclusion

FCM does well in segmenting images with few noise and outliers. But if there are some noises in an image, spatial information needs to be considered. FCM_S realizes the goal by adding spatial information to the objective function. But the Euclidean distance is not robust. Kernel functions can perform transformation from an input space with low dimension into a feature space with low dimension. KFCM_S is performing FCM in the kernelized feature space. And KFCM_S performs well on images with noises, but its computational complexity is very high. In this paper, we improved the objective function of original KFCM_S. In the process of images with many noises are segmented, though the segmenting result of FKFCM_S can't be better than the result of KFCM_S, the running time decreases sharply. And FKFCM_S can adapt different noise condition by adjusting the weight values of currently processed pixel and its neighbor pixels

References

1. D. L. Pham，C.Y. Xu，J. L. Prince: A survey of current methods in medical image segmentation. Annu. Rev., vol. 2, Biomed. Eng. (2000)315-337.
2. Xinfeng Zhang，Lansun Zhang: Image segmentation research. Journal of Circuits and Systems，vol.9，(2004-04).
3. Zhaoqi Bian，Xuegong Zhang: Pattern recognition（The second edition）, Tsinghua University Press，BeiJing (2000-01).
4. A. W. C. Liew，S. H. Leung，W. H. Lau: Fuzzy image clustering incorporating spatial continuity. Image Signal Process，vol. 147, (2000)185-192.
5. Jianwu Liu，Zhiqian Ye，Jinfang Lu: Application of FCM segmentation algorithm of MRI brain tumor images. Medicine & Engineering，vol. 4, (2002)33-35.
6. R. J. Hathaway，J.C Bezdek: Generalized fuzzy c-means clustering strategies using L_p norm distance. IEEE Trans. Fuzzy Syst，vol. 8，(2000)572-576.
7. D. L. Pham: Fuzzy clustering with spatial constraints. Image Processing. IEEE Int. Conf, New York，(2002) Ⅱ-65-Ⅱ-68.
8. M. N. Ahmed，S. M. Yamany，N. Mohamed: A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. IEEE Trans. Med. Imaging，vol. 21，(2002)193-199.
9. K. R. Muller，S. Mika et al.: An introduction to kernel-based learning algorithms. IEEE Trans. Neural Networks, vol. 12, (2001)181-202.
10. Zhang Li，Weida Zhou，Licheng Jiao: Kernel clustering algorithm. Chinese J. computers，vol. 25, (2002)587-590.
11. Saunders C，Stitson M O，Weston J et al.: Support vector machine-reference manual. Technical Report CSD-TR-98-03，London，Egham，UK (1998).
12. K. L. Wu，M . S. Yang: Alternative c-means clustering algorithms. Pattern Recognition，vol. 35, (2002)2267-2278.

An object delineation algorithm for rock particles under uneven lightning

Weixing Wang

School of Electronic Engineering, University of Electronic Science and Technology of China, Post code: 610054, Chengdu, China
wxwang@ee.uestc.edu.cn or znn525d@yahoo.com

Abstract. Uneven illumination creates difficulty for image processing and segmentation in general. This paper shows that an algorithm technique involving image classification and valley-edge based fragment delineation is a highly efficient way of delineating densely packed rock particles for the images of uneven illumination. The paper shows the usefulness of this technique for complicated rock particles. The reason for the powerfulness of the technique is that image classification and fragment delineation are highly cooperative processes. Moreover, valley-edge detection is a nonlinear filter picking up evidence of valley-edges by only considering the strongest response for a number of directions. The algorithm has been compared to the other existing algorithms. The result shows that it is not affected much by fragment surface noise and image uneven illumination. It is robust for densely packed rock particles.

1 Introduction

In most applications, the quality of rock particle images varies too much, which make image segmentation hard. Therefore, this research subject becomes a hot topic in the world during last twenty years. Today, a number of image systems have been developed for measuring fragments in different application environments such as fragments on/in gravitational flows, conveyor belts, rockpiles, and laboratories [1-5].

In a rock particle image of size 768x576 pixels (e.g. ordinary CCD camera), the number of fragments may reach up to 2000. Moreover, if there is no clear void space (background) between fragments, the fragments often overlap and touch each other. If the illumination on the fragment surface is uneven, the light intensities of fragments are different; and if in some cases, rock types are varying, the edges between fragments are weak. All the mentioned characteristics of rock particle images make segmentation algorithm development hard. It is not practical to have the same segmentation procedure for images irrespective of quality and size distribution. Hence, it is crucial to extract qualitative information about a rock particle image to characterize images before starting segmentation. Characterization of rock particle images have been thoroughly investigated by extensive tests on hundreds of images, using several packages of commercial software for image segmentation, and some previous image segmentation algorithms[1-3] coded by the authors.

The paper stresses that our general approach is that of using two building blocks for algorithms, which is called “image classification” and “image segmentation”. It is the cooperation between image classifications and “image segmentation” which creates good delineation of rock particles.

Image classification was recognized to be essential in segmentation of rock particles. Therefore, it was started by developing procedures [6], for crude determination of number of rock particles in an image, the basic idea being that “edge density” is a rough measure of average size in images of densely packed fragments. It stresses: our segmentation algorithm is based on grey value valleys which is a grey-value structure occurring more frequently than traditional step edges. However, without knowledge of scale (approximate size of rock particles in the image) such an approach would be hard to realize. In fact, this goes for any segmentation technique which normally “handles” the problem by adjustment of various smoothing parameters, thresholds etc. Since it needs an automatic image segmentation procedure the algorithm performs “image classification” first, to avoid making “smoothing parameters” crucial for good results.

2 Rock particle image classification algorithm

Because of the large variation of rock particle patterns and quality, the image classification algorithm produces five different labels for the classes: Class 1: images in which most of the fragments are of small size; Class 2: images in which most of the fragments are of medium size; Class 3: images in which most of the fragments are of relative large size; Class 4: images with mixed fragments of different sizes; and Class 5: images with many void spaces.

If most fragments in an image are very small, the fine-detail information in the image is very important for image segmentation, and the segmentation algorithm must avoid destroying the information. On the contrary, if fragments are large, it is necessary to remove the detailed information on the rock particle surface, because it may cause image over-segmentation. If most fragments are of relative large size (e.g. 200 pixels for each fragment), the segmentation algorithm should include a special image enhancement routine that can eliminate noise of rock particle surface, while keeping real edges from being destroyed.

There is also a special class of images, Class 5. This class refers to any of Classes 1 to 4 on a clear background, hence only partially dense. In this special case, Canny edge detection [7] is a good tool for delineating background boundaries for clusters of rock particles. To classify images, the edge density is used. Edge density may reflect average size and surface texture patterns. The underlying question here is not only whether or not edge density is used for estimating average fragment size, but rather, more generally: Is this number useful for classifying images of densely packed rock particles into size categories, and, can this classification facilitate segmentation of rock particles?

A direct method for estimating edge density may be based on, e.g., a Canny edge detector [5]. An indirect method for estimating edge density may also be based on *moment-preserving thresholding* [2], applied to a smoothed gradient magnitude image (instead of the original image), whereby a crude estimate of edge density is based on information in a gradient magnitude histogram. The method described below is more accurate, but both methods are useful.

By size of fragments, it means diameter or approximate diameter of fragments. Under the assumption of closely packed fragments of –roughly- elliptical-like shape, it shows that there is a relation between on one hand average size and on the other edge density, an average shape factor, and variance of size. If the variation of size is not too large, edge density and average shape factor are sufficient for estimating average size, it maybe within an accuracy of 5 to 10 percent (as investigated), if the direct method is used.

Edge-density based average size estimation is a relatively computationally inexpensive way of obtaining average size. If using some specially tailored hardware card for quick convolutions for image smoothing, the rest of the operations only require a few operations per pixel. For the direct approach, Canny edges only require a few comparisons and multiplications per pixel, and the same holds for moment calculations in the indirect approach. Thus, edge-density based approaches should be possible to use in real-time applications in processing of image sequences. Author has performed a series of experiments with edge density-based average size estimation.

Edge density δ will be measured. It denotes estimated edge density by $\hat{\delta}$. One possible way of calculating $\hat{\delta}$ is to divide the total number of edge pixels in an image by the total number of pixels n_{tot} .

$$\hat{\delta} = \frac{n_e}{n_{tot}} \tag{3}$$

where n_e = number of edge pixels.

Let the original image be $f(x,y)$ and its gradient magnitude image $M_f(x,y)$, and * be convolution. A smoothed image

$$g(x,y) = f(x,y) * h(x,y) \tag{4}$$

where h is some smoothing filter. The Canny edge detector uses a Gaussian filter:

$$h(x,y;\sigma_{gauss}) = \exp\left(-\frac{x^2+y^2}{2\sigma_{gauss}^2}\right) \tag{5}$$

$\nabla g(x,y)$ is the vector field of gradients of g, which in practice is stored as a pair of images: $\nabla g(x,y) = (g_x(x,y), g_y(x,y))$. $M_g = |\nabla g|$ is the gradient magnitude image of the smoothed image g . The smoothing parameter c_{gauss} is the so-called scale parameter.

A Canny edge image is defined as maxima of gradient magnitude in the gradient direction. In a discrete grid, then, pixels corresponding to such maxima will be edge pixels and the binary image where 0 means edge pixel and nonzero value non-edge pixel, is an edge image denoted $\varepsilon_g(x,y)$, or, more precisely, $\varepsilon_g(x,y;c_{gauss})$.

Our edge density $\hat{\delta}$ will always be based on a n_e value calculated in an ε_g edge image.

The Canny edge detector compares three M_g -values locally in the gradient direction, $\nabla g(x, y)$. For instance, if the edge runs along the y-direction, then and $M_g(x, y) \geq M_g(x + 1, y)$ is required for assigning the label edge pixel to a pixel at (x, y). Normally, a threshold, denoted by t_M here, is used to eliminate edge pixel candidates of low contrast by requiring $M_g(x, y) > t_M$. This gradient magnitude threshold could be adaptive to each image, which normally yields a t_M -value in the range 5% to 10% of maximal M_g -value in the image, the exact value of t_M epending on whether there is a peak in the M_g -histogram in that interval and its position.

Consider the case of an image containing closely packed rock particles, which can be approximated by ellipses in the image plane. The approximation is not done for the purpose of describing individual fragment shape, but for setting up a model for relating edge density to average size. The concept size is defined below.

The ellipses are indexed $i = 1,2,\cdots,n$. Let minor and major axes be W_i and L_i , with $W_i < L_i$, $r_i = W_i / L_i$. We use L_i as a measure of size, and call it length. Denote area and perimeter by A_i and P_i respectively. Assume that there are no boundaries in the interior of the ellipses. Define the following edge density concept δ_* :

$$\delta_* = \frac{P_1 + P_2 + \cdots + P_n}{A_1 + A_2 + \cdots + A_n} \qquad (6)$$

And relate to size L_i :

$$\frac{\sum_i P_i}{\sum_i A_i} = \frac{\sum_i 2L_i E(\sqrt{1 - r_i^{\,2}})}{\sum_i \pi r_i L_i^{\,2} / 4} \approx \frac{4}{\sqrt{2}} \frac{\pi \sum_i \sqrt{1 + r_i^{\,2}} L_i}{\pi \sum_i r_i L_i^{\,2}} = \frac{\frac{4}{\sqrt{2}} \sqrt{1 + (r(\xi_1))^2}}{r(\xi_2)} \cdot \frac{\sum_i L_i}{\sum_i L_i^{\,2}}, \qquad (7)$$

where E( ) is the complete elliptic integral.

The last equality is due to the mean value theorem of integrals: $\int G(x)H(x)dx = G(\xi)\int H(x)dx$, which also applies to sums if we replace discrete functions by continuous step functions $\xi_1, \xi_2 \in [1, n]$.

The value $r(\xi_2)$ is by definition a weighted average of $r_1, r_2, \cdots, r_n$ with $L_i^{\,2} / \sum L_i^{\,2}, i = 1,2,\cdots,n$ as weights, and $\sqrt{1 + (r(\xi 1))^2}$ is another weighted average, but with $L_i / \sum L_i, i = 1,2,\cdots,n$ as weights. Note that $r(\xi_1) \neq r(\xi_2)$, in general.

$$\frac{\sum L_i}{\sum L_i^{\,2}} = \frac{\frac{1}{n}\sum L_i}{\frac{1}{n}\sum L_i^{\,2}} = \frac{\bar{L}}{\bar{L}^2 + \sigma_L^{\,2}} = \frac{1}{\bar{L} + \sigma_L^{\,2} / \bar{L}} \qquad (8)$$

$\bar{L}$ is average length ($\bar{L} = n^{-1} \sum L$), and σ_L^2 the sample variance of L defined as $\sigma_L^2 = n^{-1} \sum (L_i - \bar{L})^2$. We call s$_i$ = $(4/\sqrt{2})\sqrt{1 + r_i^2} / r_i$ the shape factor and call

$$\bar{s} = \frac{\frac{4}{\sqrt{2}}\sqrt{1 + (r(\xi_1))^2}}{r(\xi_2)} \,, \; \bar{s}_{exact} = \frac{\frac{8}{\pi} E(\sqrt{1 - (r(\xi_1))^2})}{r(\xi_2)} \qquad (9)$$

the “average shape factor”.(When all ellipses are of the same form r$_i$ = r, ∀ i, it is easily seen that $s_i = \bar{s}$.) One may note that the shape factor is closely related to compactness P^2/A. The approximation $E(\sqrt{1 - r^2}) \approx 0.5\pi\sqrt{(1 + x^2)/2}$ comes from Spiegel (1992, p7), [8], and is fairly well known.

With known average shape factor=$\bar{s}$, average size $\bar{L}$ in a single frame can be solved from Eq.7, using Eqs.8-9:

$$\bar{L} + \frac{\sigma^2{}_L}{\bar{L}} = \frac{\bar{s}}{\sum P \big/ \sum A} = \frac{\bar{s}}{\hat{\delta}_*} \qquad (10)$$

We now have a relation between average length $\bar{L}$ and a kind of edge density $\hat{\delta}^*$. The measured edge density in our experiments $\hat{\delta}$ is related to $\hat{\delta}^*$ by $\hat{\delta}^*$=β·$\hat{\delta}$ where β ≈ 1.2 accounts for the empty space between fragments (not included in $\sum A$), as discussed earlier. Now, introduce the quantity $\tilde{\sigma}_L = \sigma_L / \bar{L}$, which is a kind of normalized standard deviation. Then, $\bar{L} + \sigma_L^2 / \bar{L} = \bar{L} + \tilde{\sigma}_L^2 \cdot \bar{L}$, leading to

$$\bar{L} = \frac{\bar{s}}{\beta\hat{\delta} \cdot (1 + \tilde{\sigma}_L^2)} \qquad (11)$$

Of course, we should not expect to be able to calculate the average $r(\xi_1)$) and $r(\xi_2)$ exactly. An approximation $r_m \approx r(\xi_1)$, $r_m \approx r(\xi_2)$ may be calculated from crudely split-merge segmented data by using a kind of “equivalent ellipse” concept, yielding an estimate

$$\bar{s} = (4 \big/ \sqrt{2}) \cdot \sqrt{1 + r^2{}_m} \big/ r_m \qquad (12)$$

which is the shape factor we use in the experiments.

If there are clear dark void spaces in an image, the algorithm can also be used. Before the classification, the void spaces can be detected by a simple thresholding algorithm or by a Canny edge detector along the between-class boundaries. Based on estimates of average number of fragments, images are labeled automatically into four classes.

3 Rock particle delineation algorithm

An algorithm has been proved to be useful for rock particles. This is not just a fortuitous coincidence. Rather, densely packed rock particles, typically, are separated by dark valleys (of variable width), if defining "valley" not too locally. A valley needs not to be symmetric. It could not be steeply sloping on one side and gently sloping on the other. Classic edge detection register any change in grey value above a threshold, and works either locally (bad for our data), or, more globally after isotropic smoothing.

For the details on the algorithm we refer to [9]. Here we emphases on the experimental results using that algorithm on rock particles instead, and emphasize which parts of the valley edge segmentation are gaining from cooperation with the image classification algorithm and its five classes. We suggest that image classification is a reasonable way of handling scale or multiple scales, in a segmentation procedure.

In the example, a valley point P is surrounded by strong negative and positive differences in the diagonal directions:

$\nabla_{45} < 0$, and $\Delta_{45} > 0$, $\nabla_{135} < 0$, and $\Delta_{135} > 0$, whereas, $\nabla_0 \approx 0$, and $\Delta_0 \geq 0$, $\nabla_{90} \approx 0$, and $\Delta_{90} \approx 0$

where Δ are forward differences: $\Delta_{45} = f(i+1, j+1) - f(i, j)$, and ∇ are backward differences: $\nabla_{45} = f(i, j) - f(i-1, j-1)$, etc. for other directions. We use $\max(\Delta_\alpha - \nabla_\alpha)$ as a measure of the strength of a valley point candidate. It should be noted that we use sampled grid coordinates, which are much more sparse than the pixel grid $0 \leq x \leq n$, $0 \leq y \leq m$. f is the original grey value image after weak smoothing. What should be stressed about the valley edge detector is:

(a) It uses four instead of two directions;

(b) It studies value differences of well separated points: the sparse $i \pm 1$ corresponds to $x \pm L$ and $j \pm 1$ corresponds to $y \pm L$, where $L >> 1$;

(c) It is nonlinear: only the most valley-like directional response $(\Delta_\alpha - \nabla_\alpha)$ is used. By valley-like, we mean $(\Delta_\alpha - \nabla_\alpha)$ value. To manage valley detection in cases of broader valleys, there is a slight modification whereby weighted averages of $(\Delta_\alpha - \nabla_\alpha)$- expressions are used.

$$w_1\Delta_\alpha(P_B) + w_2\Delta_\alpha(P_A) - w_2\nabla_\alpha(P_B) - w_1\nabla_\alpha(P_A),$$

where P_A, P_B are neighbors of detecting point P, opposite. For example, $w_1 = 2$ and $w_2 = 3$ are in our experiments.

Without image classification, there is a substantial difficulty choosing an appropriate L, the spacing between sampled points. Let L refer to spacing in an image of given resolution, where the given resolution may be a down sampling of the original image resolution.

Since the image classification described earlier leads to an automatic down-sampling (see Eqs. (1) and (2)) of images belonging to Class 2 or Class 3, large or medium-sized objects, the choice of L is not critical.

For Class 4 images – mixed size images – multiple values of L are used. Within each detected coarse-scale fragment, edge density variations in the un-shrunk image (may) trigger a new valley-detection search. The L-value has not changed, numerically, but we are operating in higher resolution variations of the image (Class 2 and Class 1), hence L in terms of the original image decreases.

After valley edge point detection, we have pieces of valley edges, and a valley edge tracing subroutine, filling gaps is needed (Some thinning is also needed.). We refer to [9] for details.

After the above procedure, the remaining thing is to close the boundaries or contours for each of the fragments. If valley edge image is of the same size of the original image, it firstly detects end points and junction points of edges on the valley edge image, then calculates the orientations for each of the points, subsequently, detects flatness, weakness of the curves between the two corresponding points which are oriented in the similar directions and have a relative short distance, choose the best fitting curve for the connection of the two corresponding points, which is repeated until no end points exists. If valley edge image is of the size less than the one of the original image that means that the valley edge image is resulted from a shrunk fragment image. In this case, it remaps valley edges to the original image; it does not just simply enlarge the edges, and it uses the valley edges as cues to search possible edge locations in the large image based on grey level information. After edge remapping, it carries out boundary closing operation in the same way as before described.

As a background process, there is a simple grey value thresholding subroutine which before classification creates a binary image with quite dark regions as the bellow-threshold class. If this dark space covers more than a certain percentage of the image, and has few holes, background is separated from fragments by a Canny edge detector [7] along the between-class boundaries.

In that case, the image is then classified into Class 1 to 4, only after separation of background. This special case is not unusual in rock particle data. This is reasonable cooperative process. If background is easily separable from brighter rock particles this is done, and dense sub-clusters are handled by the image classification and valley-edge segmentation. This part of the segmentation process is specific for rock particle images where part of a homogeneous (dark) background is discernible.

To test the segmentation algorithm, we have taken a number of different fragment images from a laboratory, a rockpile, and a moving conveyor belt. It is often that there is a lot of noise on the surface of fragments, which gives problems for image segmentation, over-segmentation and under-segmentation. Since surface noise and 3D geometry of rock particles create step edges in most cases, and our new algorithm is studied based valley edge detection, it disregards step edges. Therefore it works not only for less surface noise image, and also works for the images of serious surface noise. The other existing algorithms [10-12] have difficulties to process this kind of images.

When one acquires (or takes) rock particle images in the field, the lightning is uncontrolled; therefore, it cannot be avoided having uneven illumination images. Uneven illumination is a serious problem for image processing and image segmentation not only for rock particles and also for other object. Uneven illumination correction is a hot topic in the research of image processing. In general, the regular shadows can be removed by using some standard filters, but for the random shadows, there is no standard filter or algorithm can be used for uneven illumination correction.

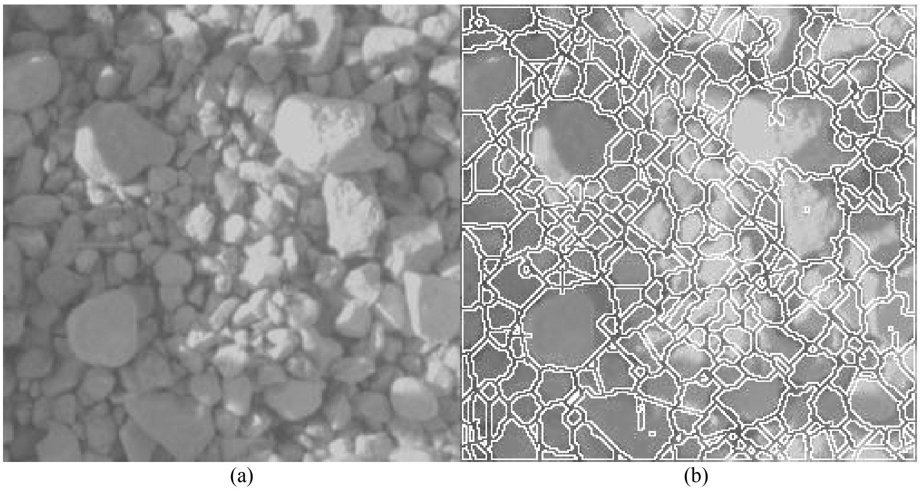


Fig. 1 Fragment delineation for a typical image under uneven illumination: (a) original image; and (b) fragment delineation result.

Rock particles are in field, lightning is from the natural sun (light strength varies from time to time) , some natural objects (e.g. clouds, forest, mountains) and large man-made objects (e.g. trucks, trans) maybe nearby the area one wants to take images, which may create uneven illumination (i.e. shadows) on the images. Some times, in a fragment image, it includes high lightning area and dark shadows, which make image segmentation extremely difficult. It is not possible to use the segmentation algorithms based on grey level similarity. In the newly studied fragment delineation algorithm, since it uses valley edges as cues for object delineation, it is not affected by uneven illumination much. As examples, we show two uneven illumination images in Figs. 1-2. The image in Fig. 1(a) has random shadows. By using the new algorithm, the fragment delineation results are satisfactory too.

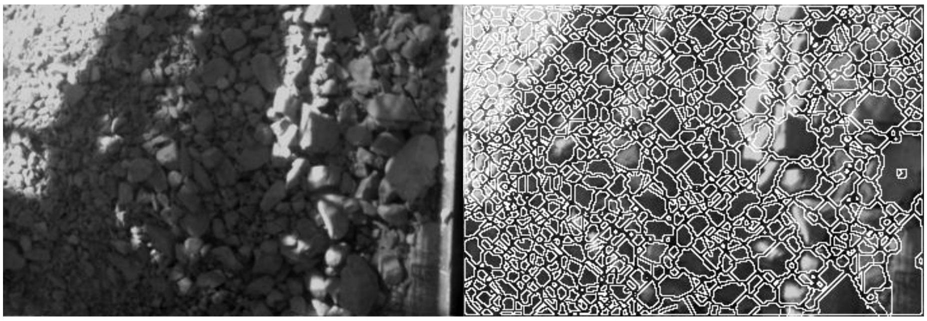


Fig. 2 Fragment delineation for the image of random shadows: (a) original image; and (b) fragment delineation result.

4 Conclusions

In this paper, a new image algorithm for delineating densely packed rock particles has been studied; it is a combination of image classification algorithm and fragment delineation algorithm. The rock particle image classification algorithm was developed based on edge density and average size estimation. The edge density is obtained by using Canny edge detection, and size estimation is from a deduction of series of mathematics formulae. For general-purpose rock particle image segmentation by valley-edge detection, the classification algorithm produces image class labels (5 classes), useful in subsequent image segmentation. Without image classification, image down–scale cannot be carried out correctly, and valley edge could not detect edges clearly. The fragment delineation algorithm was studied based on both valley-edge detection and valley-edge tracing; it differs from traditional edge based segmentation algorithms, as compared to the other edge detectors, the studied edge detection algorithm is more suitable for densely packed rock particles. The delineation algorithm uses valley edges as cues for boundary tracing and closing, it also uses multiple scale technique to remap fragment boundaries.

The presented rock particle delineation algorithm has been tested for a number of rock particle images where fragments packed densely. The algorithm has been compared to the other widely used fragment image segmentation algorithms, the result shows that it is much robust than the other algorithms for densely packed rock particles under the condition of uneven illumination, it is not affected much by the surface noise of rock particles and image uneven illumination which affect the other existing algorithms seriously. Therefore, it is powerful and suitable for rock particle images.

Acknowledgement

This contribution is part of the SCST (Science and Technology Bureau of Sichuan Province) project, Project No.05JY029-070.

References

[1] Wang WX. Computer vision for rock aggregates. Ph.D. thesis, Division of Engineering Geology, Department of Civil and Environmental Engineering, Royal Institute of Technology, Stockholm, Sweden, 1997.

[2] Wang WX, Bergholm F. On Moment - Based Edge Density for Automatic Size Inspection. In: Proceedings of the 9th Scandinavian Conference on Image Analysis, in Uppsala, Sweden, on June 6-9, 1995, pp. 895 – 904.

[3] Wang WX. Image analysis of aggregates. J Computers & Geosciences, 1999, No. 25, pp. 71-81.

[4] Kemeny J, Mofya E, Kaunda R, Lever P. Improvements in Blast Fragmentation Models Using Digital Image Processing, Fragblast, Volume 6, Numbers 3-4 / December 2002, pp. 311 - 320, Publisher: Taylor & Francis.

[5] Norbert H Maerz, Tom W, Palangio. Post-Muckpile, Pre-Primary Crusher, Automated Optical Blast Fragmentation Sizing, Fragblast, Volume 8, Number 2 / June, 2004, pp. 119 – 136, Publisher: Taylor & Francis.

[6] Wang WX, Bergholm F, Yang F. Froth delineation based on image classification. J Mineral Engineering, Volume 16, Issue 11, November 2003, p. 1183-1192.

[7] Canny JF. A computational approach to edge detection, J PAMI-8, No.6, 1986.

[8] Spiegel MR. Schaum's outline series. Mathematical Handbook of Formulas and Tables, 28th printing, U.S.A. 1992.

[9] Bergholm F, Wang WX. Image characterization for segmentation. In: Proceedings of the First International Conference on Image and Graphics Technology toward 21 Century and Beyond, Tianjin, China, August 16-18, 2000, pp.320-323.

[10] Otsu N. A threshold selection method from gray-level histogram. IEEE Trans. Systems Man Cybernet, 1979, SMC-9, 62-66.

[11] Suk M，Chung SM. A new image segmentation technique based on partition mode test. J Pattern Recognition Vol. 16, No. 5, 1983, 469-480.

[12] Gonzalez R, Wintz P. Digital image processing. Second edition, Addison-Wesley Publ. Comp., printed in the USA, 1987, p.354-374.

ibai Institute of Computer Vision
and Applied Computer Sciences
Dr. Petra Perner

Institute of Computer Vision and Applied Computer Sciences IBaI

Director: Dr. Petra Perner

Address: Körnerstr. 10
04107 Leipzig
Germany

Phone: +49 341 8612273
FAX: +49 341 8612275
E-Mail: info@ibai-institut.de

Personal Homepage:
www.ibai-research.de

Institute`s Homepage:
www.ibai-institut.de

International Conference on Data Mining and Machine Learning MLDM
www.mldm.de

Industrial Conference on Data Mining ICDM
www.data-mining-forum.de

BioMedVision Center
www.biomedvision.de

Data Minng Tutorial
www.data-mining-tutorial.de