Visualization in Medicine and Life Sciences (2013)
L. Linsen, H. -C. Hege, and B. Hamann (Editors)

Short Papers

Towards a Structured Analysis of Quantitative Descriptors from Segmented Biological Image Data

H. Leitte[1], J. Portl[1], I. V. Röder[2], R. R. Schröder[2], and I. Wacker[3]

[1]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany
[2]CellNetworks, BioQuant, Heidelberg University, Germany
[3]Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract

Topological and morphological descriptions of (sub-)cellular structures play a central role in the understanding of biological processes. Deriving such descriptions from image data, however, is a challenging task that has so far only been addressed for individual objects at a coarse resolution with small numbers of samples. For larger samples, the structured analysis is highly challenging as little a priori knowledge exists. In this paper, we address the design of a generic parameter space for segmented objects that forms the basis for subsequent structural analysis. We detail theoretical considerations, discuss the proposed model using examples from electron microscopy, and summarize lessons learned for subsequent implementation and analysis.

Categories and Subject Descriptors (according to ACM CCS):

1. Introduction

Visual analysis of microscopy images has played a dominant role in the history of biology. For small samples and a limited number of images this proved to be a very powerful analysis strategy. With the rise of high-resolution high-throughput microscopy, however, this is no longer feasible. Modern electron microscopes, for example, can automatically scan large tissue blocks resulting in image stacks consisting of several hundred images, each with a resolution of several giga pixels. Manual analysis in this setting is no longer practicable and the need for automatic support has been widely expressed in the past years [WSB*10].

Excellent progress has been achieved in the last years for images capturing cell data [CJL*06,JKW*08], gene expression patterns [RWK*06,TC05,SND05], or cellular pathways [SHF*03, DBD*02]. In these applications, there already exists a quite good understanding of how to quantify relevant parameters in the data using, for example, size, shape, or texture statistics which can be automatically computed using computer vision and machine learning algorithms.

For segmented image structures and their analysis, however, the required parameters are often not clear as there is yet too little understanding of what is in the data, how structures look like, and how much natural variability there is. A precise quantification has not yet been addressed as (a) the necessary image modalities to record such data have not been available and (b) tools that are currently available for the analysis of segmented image data are not yet powerful enough and often require a large amount of manual tweaking and implementation experience from the biologist.

The data acquisition part has lately been addressed with the new generation of electron microscopes that allow for the rapid acquisition of huge amounts of image data each at a high level of detail. Figure 1(left) shows an electron microscopy image of a mouse's muscle tissue along with the segmented structures. Such tissue scans hold the promise of telling what the "standard" cell looks like and how much natural variability there exists. The big remaining challenge is to derive relevant description parameters and to build an appropriate model organism with inherent uncertainty.

In this paper, we address the first issue –parameter space analysis– and make the following contributions:

- (i) We detail a systematic description of the feature space for segmented biological image data.
- (ii) We describe lessons learned on how to model these feature spaces in software.
- (iii) We discuss all findings using examples from electron microscopy data of neuromuscular junctions.

DOI: 10.2312/PE.VMLS.VMLS2013.025-029

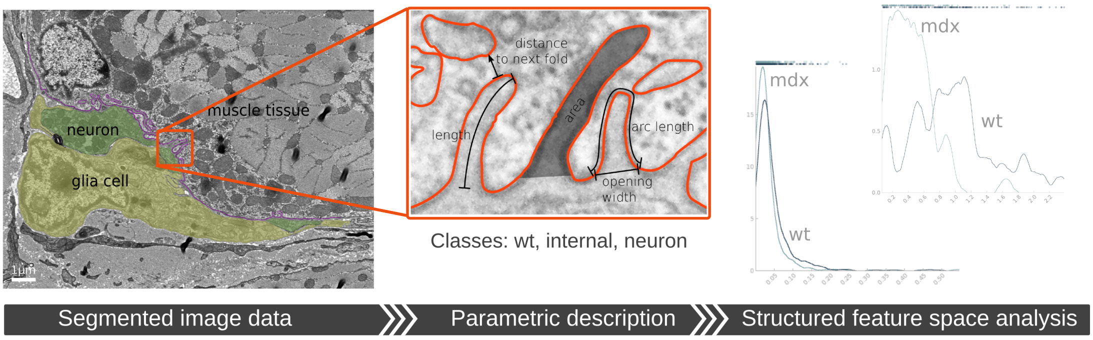


Figure 1: Analysis pipeline (left to right): (i) Classification of segmented data objects. (ii) Parameter space design. (iii) Final analysis of parametric features data.

2. Related Work

In visualization there has been much effort to visualize multi-dimensional parameter spaces. Starting with static methods like parallel coordinates [Ins85] and scatterplot matrices, increasingly sophisticated algorithms have been designed using, for example, multivariate projection (PCA, multidimensional scaling [BG05], worlds within worlds [FB90] or hyperslice [vWvL93]). Additional support is provided using animation or animated interaction to guide the user on their tour through the parameter space. Examples in this area are the grand tour [Asi85], rolling the dice [EDF08], or the TripAdvisor$^{N\text{-}D}$ [NM13].

An alternative approach analysis the influence of parameter settings on an underlying model. This domain can be divided into two groups: The first one examines the impact of parameters on the model and wants to understand how changing the settings affects the model (e.g., for simulation models [PBK10,BPFG11,BMPM12] or image segmentation [TWSM*11, PBCR11, CJL*06]). The second group supports parameter selection by result, i.e., the user is presented a number of outcomes and explores the parameter space based on the outcome. Applications in the area of computer graphics model selection are [MAB*97, BM10].

The methods discussed so far assume that the parameter space is known a priori. In our applications, we found that the scientific question was not very precise. The goal is to derive a metric description of segmented structures. Which aspects of the data are relevant and need to be integrated into the parameter space is often not clear. Additionally, we usually do not deal with continuous parameters but with power sets of nominal parameters where many of the above presented methods will fail. Hence, we concentrate in this paper on the modeling of a flexible parameter space design for the analysis of segmented objects from biological image data.

3. The Parameter Space

The first step in the comprehensive quantitative description of biological structures is the definition of a parameter space that comprises all potentially relevant information concerning the segmented structures. Often relevant parameters comprise structural properties (e.g. length, width, volume, shape), textural properties (e.g. intensity, granularity, patterns), or neighborhood information (e.g. nearest object of same type, number of surrounding objects of different type). Commonly there is also secondary information such as the age of the subject the sample was taken from, the type of subject (e.g. healthy vs. pathological), or the location of the sample within a larger context (e.g. muscle tissue sample from the leg, heart, or face). The parameter space has to encompass all this information to allow for a versatile subsequent analysis.

3.1. Biological Questions and Model Implications

Before going into detail about the precise definition of the parameter space, we want to look at some typical questions that arise when analyzing segmented image data:

- In which features do healthy and pathological subj. differ?
- What is the mean shape and how does it vary?
- Does spatial location influence the segmented structures?
- How do structures change with increasing age?

These questions illustrate some of the fundamental difficulties in visualizing biological segmentation data, which strongly influence the theoretical data basis and the software design: (i) **Problem specification:** The scientific question is commonly rather fuzzy. The transition from the general “I want to understand my data.” to precise questions that can be answered using visualization or mathematical analysis is a central aspect in the “visualization” process. (ii) **Feature selection and definition:** Features necessary to answer the extracted questions are commonly not clear a priori. A variety of potentially helpful features has to be identified, implemented, and tested. (iii) **Interactive subset selection:** The software has to support database-like functionalities to select and/or group subsets of the data. (iv) **Coupling of multiple algorithms interactively:** The users commonly require a coupled set of algorithms for a holistic analysis. Visualizations of the raw data with highlighting functionality have to be connected to parameter space visualization and exploration algorithms via linking-and-brushing. (v) **Ensemble visualization:** The software needs to be able to visualize ensembles to compare multiple subjects across different groups.

3.2. Features and Classes

An important part of the parametric data description is the distinction of two types of information: we will divide object parameters into features and classes. *Features* are quantifiable properties of segmented objects that are commonly directly measured using the input image/segmentation data. Examples for features are size, shape, or texture. *Classes* are properties of the segmented objects that are used to cluster data. Examples for classes are healthy vs. pathological, age group, species, or location of sample. During our research we found some cases, where classification was more difficult. These are commonly derived attributes such as spatial location within the data set, which might be used for data aggregation (→ class) as well as statistical analysis (→ feature). The implications for the software design will be discussed later. Overall, we found this distinction very helpful to provide a first structuring of the data, which is used to add levels of detail to the analysis process, and to make the biologists render their research questions more precisely.

Class information is commonly used to separate the segmented data into groups that are to be compared. The easiest setting is to chose one primary class, e.g., a certain age group, and compare it to the rest of the data. Often a more detailed analysis is necessary to derive meaningful information. For example, subjects of a certain age group may only differ in healthy subjects but not in pathological ones.

Feature information is the one that is subsequently analyzed and used to derive model information. In manual analysis of the feature space, biologists commonly rely on low-dimensional standard techniques such as histograms, scatterplots or heat maps [WSB*10]. For many data attributes, the distinction between feature and class is readily given.

In summary, classes are for data selection and clustering and features are used to quantify commonalities and differences between subsets of the data.

3.3. Theoretical Model

Taking the previous considerations into account the theoretical model for the parameter space for quantitative descriptors of segmented biological image data consist of two major parts for data classes and features respectively. The first n dimensions of the parameter space are dedicated to the n classes. In many cases classes take nominal values that have no implicit ordering, such as type of species, gender, or spatial location. There also exist classes with ordinal and even continuous properties, such as age or data acquisition parameters. The next m dimensions represent the derived object features which are often continuous scalar quantities but may also have more complex structure such as shape information or graph structures to represent neighborhood properties.

Data points are stored in the common form: each segmented object is assigned an n+m-dimensional vector containing class and feature information. In our collaborations, we often revised and extended the feature space, which does not affect the theoretical model, but has strong implications on the resulting implementation.

Operations on the data directly arise from the feature and class discussions. For classes we require functionalities for interactive selection and grouping of data. For features we need data analysis routines from statistics and machine learning to extract data characteristics and clusters and structures that emerge in the parameter space.

For the visual inspection of the parameter space, information visualization techniques and statistical graphics are necessary to render the parameter space. Augmented image visualizations including highlighting and interactive selection are required for the image and segmentation data. As discussed in section 3.1, all methods have to be linked via linking-and-brushing and have to be interactive. They also have to support ensemble visualization.

4. Example: Folds in the Neuromuscular Junction

In this example, we illustrate the design of the parameter space for junctional folds. Junctional folds are subcellular structures at neuromuscular junctions (NMJs) located at the junction between an axon terminal of a motoneuron and a muscle fiber (see fig. 1). Samples were taken from muscle tissue of mice.

First we collected the relevant object classes:

$$\begin{aligned}
\texttt{Synapse Type} &\in \{wt, mdx\}\\
\texttt{Location Type} &\in \{neuron, interspace, glia\}\\
\texttt{Fold Type} &\in \{primary, secondary, internal\}
\end{aligned}$$

where synapse type describes the animal's state of health. *wt* encodes healthy wild type mice and *mdx* pathological ones. The location type indicates the cell organelle that lies opposite the junctional fold. The fold type characterizes the structure of the fold. Regular ones on the membrane are called *primary*, those having a protruding extra bulge are called *secondary* and folds without a direct connection to the membrane section in 2D are called internal. At a later stage, we found that we have to extend the model to account for an additional class: spatial location within the entire synapse.

For the features we started with a small initial set of widely used features. As large scale quantitative descriptions have so far not been made for NMJs and junctional folds the set of relevant features was not clear a priori and we will extend it on demand. So far we measure for each junctional fold (see figure 1(center) for an illustration): area, length, arc length, opening width, distance to next fold.

This initial distinction forced the biologists to formulate the research questions more precisely. On a very coarse level, they had to decide which groups they want to compare and which features they want to analyze statistically. This added valuable user knowledge to the system, which intrinsically structured the resulting info graphics (compare fig. 1(right)) in a hierarchical fashion.

Additionally, we used this information to provide a more structured interface to the data that readily summarizes all classes and features (fig. 2). Classes and their values are presented on the left hand-side, features on the right hand-side. We allow the user to select a primary class that is used to partition the data into different groups. Data selection is performed by selecting or deselecting presented classes. Two sample configurations are presented in fig. 2b. In an additional pixel-based visualization, we record the configurations that the user has already investigated, which gives feedback about potentially interesting missing configurations. Upon selection the user is presented with statistical graphics of the selected parameter (compare fig. 1(right)). The new interface helped them to quickly update their selection and see the implications on the resulting statistical graphics.

Using the described parameter space and the visual interface, biologists had for the first time an easy and clearly structured access to their data. During their trial they made the following observations:

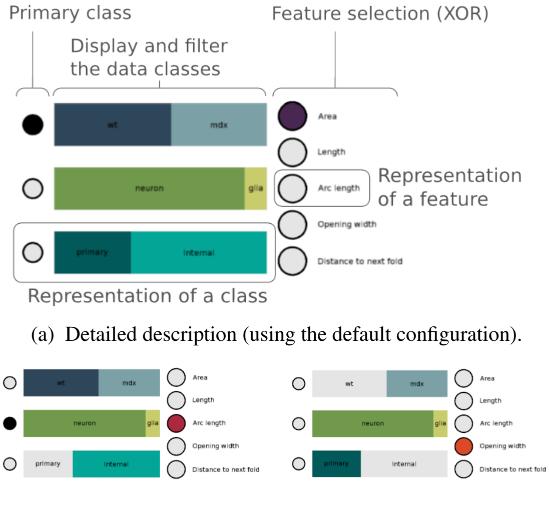
- Differences could be observed in all features when comparing folds of mdx and wt mice.
- Those differences became clearer and more pronounced for subgroups. In particular, they found that internal and primary folds should not be analyzed jointly.
- In literature it is postulated that folds are usually opposite to neurons. With our interface biologists rapidly observed that a substantial number of folds are located outside these regions, what they want to investigate in more detail.

5. Lessons Learned for Software Development

The parameter space design resulted from several iterations of software development and monthly discussions with biologists over the last two years. From this work we learned:

The structuring helped in the analysis of the parameter spaces with dozens of attributes. In a first implementation we summarized classes and features in a common data model which constantly resulted in confusions about the allowed operations and the configurations that already had been analyzed. A clear distinction helps in the analysis process and in the formulation of biological questions.

The structured navigation panel, a visual interface holding all attributes compared to simple drop-down menus, helped us select valid configurations while presenting at the same time information about the amount of data records to be analyzed, which is relevant to ensure statistical significance.



(a) Detailed description (using the default configuration).

(b) Alternative valid configurations.

Figure 2: The navigation panel for the parameter space: The left column of circular items selects the class that is used for partitioning. The rectangular boxes in the center name for each class its entities and visualize the proportion of data in this class (width). Features are listed in the right column.

Databases provide good means to store and structure the data on demand. A simple mechanism has to be implemented to distinguish between features and classes as they allow for different operations.

The data attribute hierarchy (attribute → class/feature → name) has to be stored in an easily accessible and modifiable way. Our system is implemented in C++. In the first stage we employed enumerates that were edited on demand. We faced problems when trying to derive for a given name the respective class/feature and when trying to iterate over all values within a class/feature. In the current implementation, we employ a lookup-table with an accompanying operator class that provides both functionalities.

One unsolved problem is on how to store and pass the configuration of the selected data between algorithms efficiently. As classes, features, and their assignment still change, hard coded identifiers are not applicable.

6. Conclusion

In this paper we described a theoretical model for the design of parameter spaces for the analysis of segmented structures in biological image data. We exemplified our considerations using examples from electron microscopy images of neuromuscular junctions in mice and detailed lessons learned regarding software development.

References

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput. 6*, 1 (Jan. 1985), 128–143. 2

[BG05] BORG I., GROENEN P.: *Modern Multidimensional Scaling: theory and applications (2nd ed.)*. Springer-Verlag, 2005. 2

[BM10] BRUCKNER S., MÖLLER T.: Result-driven exploration of simulation parameter spaces for visual effects design. *IEEE Trans. Vis. Comput. Graph. 16*, 6 (2010), 1468–1476. 2

[BMPM12] BOOSHEHRIAN M., MÖLLER T., PETERMAN R. M., MUNZNER T.: Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. *Comp. Graph. Forum 31*, 3 (June 2012), 1235–1244. 2

[BPFG11] BERGER W., PIRINGER H., FILZMOSER P., GRÖLLER M. E.: Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. *Comp. Graph. Forum 30*, 3 (June 2011), 911–920. Best Paper Award. 2

[CJL*06] CARPENTER A. E., JONES T. R., LAMPRECHT M. R., CLARKE C., KANG I. H., FRIMAN O., GUERTIN D. A., CHANG J. H., LINDQUIST R. A., MOFFAT J., GOLLAND P., SABATINI D. M.: Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology 7*, 10 (2006), R100. 1, 2

[DBD*02] DEMIR E., BABUR O., DOGRUSOZ U., GURSOY A., NISANCI G., CETIN-ATALAY R., OZTURK M.: Patika: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics 18*, 7 (2002), 996–1003.

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Trans. Vis. Comput. Graph. 14*, 6 (2008), 1539–1148. 2

[FB90] FEINER S., BESHERS C.: Worlds within worlds : Metaphors for exploring n-dimensional virtual worlds. In *Proc. of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology* (1990), ACM Press, pp. 76–83. 2

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1* (1985), 69–91. 2

[JKW*08] JONES T. R., KANG I. H., WHEELER D. B., LINDQUIST R. A., PAPALLO A., SABATINI D. M., GOLLAND P., CARPENTER A. E.: Cellprofiler analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics 9* (2008). 1

[MAB*97] MARKS J., ANDALMAN B., BEARDSLEY P. A., FREEMAN W., GIBSON S., HODGINS J., KANG T., MIRTICH B., PFISTER H., RUML W., RYALL K., SEIMS J., SHIEBER S.: Design galleries: A general approach to setting parameters for computer graphics and animation. In *SIGGRAPH '97: Proc. of the 24th annual conference on Comp. graph. and interactive techniques* (08/1997 1997), pp. 389–400. 2

[NM13] NAM J. E., MUELLER K.: Tripadvisor$^{\text{N-D}}$: A tourism-inspired high-dimensional space exploration framework with overview and detail. *IEEE Trans. Vis. Comput. Graph. 19*, 2 (2013), 291–305. 2

[PBCR11] PRETORIUS A. J., BRAY M.-A., CARPENTER A. E., RUDDLE R. A.: Visualization of parameter space for image analysis. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 2402–2411. 2

[PBK10] PIRINGER H., BERGER W., KRASSER J.: Hypermoval: Interactive visual validation of regression models for real-time simulation. *Comp. Graph. Forum 29*, 3 (2010), 983–992. 2

[RWK*06] RÜBEL O., WEBER G. H., KERÄNEN S. V. E., FOWLKES C. C., HENDRIKS C. L. L., SIMIRENKO L., SHAH N. Y., EISEN M. B., BIGGIN M. D., HAGEN H., SUDAR D., MALIK J., KNOWLES D. W., HAMANN B.: Pointcloudxplore: Visual analysis of 3d gene expression data using physical views and parallel coordinates. In *Eurographics/IEEE-VGTC Symposium on Visualization Proceedings* (2006), pp. 203–210. 1

[SHF*03] SAURO H. M., HUCKA M., FINNEY A., WELLOCK C., BOLOURI H., DOYLE J., KITANO H.: Next generation simulation tools: the systems biology workbench and biospice integration. *OMICS 7* (2003), 355–372. 1

[SND05] SARAIYA P., NORTH C., DUCA K.: An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans. Vis. Comput. Graph. 11* (2005), 443–456. 1

[TC05] THOMAS J. J., COOK K. A.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. 1

[TWSM*11] TORSNEY-WEIR T., SAAD A., MÖLLER T., HEGE H.-C., WEBER B., VERBAVATZ J.-M.: Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (Dec. 2011), 1892–1901. 2

[vWvL93] VAN WIJK J. J., VAN LIERE R.: Hyperslice: visualization of scalar functions of many variables. In *In Proc. of the 4th conference on Visualization '93* (Washington, DC, USA, 1993), VIS '93, IEEE Computer Society, pp. 119–125. 2

[WSB*10] WALTER T., SHATTUCK D. W., BALDOCK R., BASTIN M. E., CARPENTER A. E., DUCE S., ELLENBERG J., FRASER A., HAMILTON N., PIEPER S., RAGAN M. A., SCHNEIDER J. E., TOMANCAK P., HÉRICHÉ J.-K.: Visualization of image data from cells to organisms. *Nat Methods 7*, 3 Suppl (2010), S26–41. 1, 3