International Journal of Semantic Computing
Vol. 1, No. 1 (2007) 67–85

World Scientific
www.worldscientific.com

SEMANTIC ANALYSIS OF BIOLOGICAL IMAGING DATA: CHALLENGES AND OPPORTUNITIES

WAMIQ MANZOOR AHMED*, MUHAMMAD NAEEM AYYAZ*,
BARTEK RAJWA†, FARRUKH KHAN‡, ARIF GHAFOOR*
and J. PAUL ROBINSON†,§

*School of Electrical and Computer Engineering,
Purdue University, 465 Northwestern Avenue, West Lafayette,
IN 47907, USA
ghafoor@ecn.purdue.edu

†Bindley Bioscience Center, Purdue University, 1203 West State Street,
West Lafayette, IN 47907, USA

‡ Department of Computer Science, Texas Southern University,
Houston, TX 77004, USA

§Weldon School of Biomedical Engineering, Purdue University,
206 South Intramural Drive, West Lafayette,
IN 47907, USA

Microscopic imaging is one of the most common techniques for investigating biological systems. In recent years there has been a tremendous growth in the volume of biological imaging data owing to rapid advances in optical instrumentation, high-speed cameras and fluorescent probes. Powerful semantic analysis tools are required to exploit the full potential of the information content of these data. Semantic analysis of multi-modality imaging data, however, poses unique challenges. In this paper we outline the state-of-the-art in this area along with the challenges facing this domain. Information extraction from biological imaging data requires modeling at multiple levels of detail. While some applications require only quantitative analysis at the level of cells and subcellular objects, others require modeling of spatial and temporal changes associated with dynamic biological processes. Modeling of biological data at different levels of detail allows not only quantitative analysis but also the extraction of high-level semantics. Development of powerful image interpretation and semantic analysis tools has the potential to significantly help in understanding biological processes, which in turn will result in improvements in drug development and healthcare.

Keywords: Biological imaging; semantic analysis; spatio-temporal modeling.

1. Introduction

Understanding complex biological systems requires thorough information about their basic building blocks, like molecules, genes, proteins, and cells [1, 2]. Microscopic imaging is one of the most common techniques for these multi-level investigations. Various microscopy techniques are used to image the response of cell populations under different experimental conditions [3]. Traditionally, biologists

image biological samples and visually examine them to understand biological phenomena. This approach is slow, prone to human error, and unsuitable for quantitative comparison of results across experiments [4]. With the recent proliferation of fast cameras, sophisticated microscope optics, and fluorescent dyes, there has been a tremendous growth in the volume of imaging data. High-content screening (HCS) approaches have been developed that combine sophisticated optics with automation techniques for imaging large populations of cells under different experimental conditions [5]. These technologies generate unprecedented amounts of biological imaging data.

Most biological processes have spatio-temporal semantics associated with them [6]. Extracting these semantics from large sets of images in an automated and quantitative manner is the key to unraveling complex biological phenomena. Semantic analysis in the context of biological imaging refers to the development and use of image-processing and knowledge-extraction tools and algorithms for automated image interpretation. This interpretation may require different levels of processing, including low-level image processing like segmentation and feature extraction and high-level processing for extracting, representing, and modeling spatio-temporal semantics of biological objects. Semantic analysis of these spatio-temporal data has the potential to significantly improve the process of biological discovery and drug development. The volume of these data, however, gives rise to daunting computational and data management challenges. These include the development of powerful and robust tools for extraction of low-level features as well as high-level spatio-temporal semantics, mining such data to identify hidden patterns, and database management systems for efficient storage and retrieval. In this paper we outline the challenges that need to be addressed for semantic analysis and management of such data, in order to exploit the full potential of high-throughput imaging technologies.

Different biological applications have different requirements in terms of semantic analysis [7]. While some require only fluorescence intensity quantification, others require cell tracking and identification of spatio-temporal events like cell division and motility [6]. This heterogeneity in the semantic computing requirements of biological applications calls for multiple levels of modeling. Such modeling approaches can be highly beneficial in many ways. Firstly, they help in extracting from biological data objective and quantitative information that can be compared across experiments. Secondly, they improve the searching and retrieval of such data by allowing conceptual and semantic queries. Thirdly, they make the extracted knowledge available for design of future experiments.

The rest of this paper is organized as follows. Section 2 introduces different optical microscopy modalities that are used to generate biological images. In Sec. 3, we discuss data modeling and knowledge representation issues for biological imaging data. Section 4 describes some successful applications of semantic analysis in the domain of biology. Section 5 presents a generic component-based system for knowledge-based semantic analysis of biological imaging data, and Sec. 6 summarizes the paper.

2. Optical Imaging Modalities

Various optical imaging modalities are used for biological microscopy [3, 8]. These can be divided into two-dimensional and three-dimensional techniques. 2-D techniques include all forms of traditional wide-field microcopy, with multiple ways of creating contrast, such as bright-field, dark-field, phase-contrast, differential interference contrast, and fluorescence. Bright-field microscopy is the simplest mode of imaging, in which the contrast is created through the absorptive properties of a naturally colored or stained specimen. The dyes used in bright-field imaging selectively absorb a portion of the light passing through. However, most thin biological specimens if unstained do not absorb enough light to yield useful contrast. Dark-field and phase-contrast microscopy aim at increasing contrast in unstained samples. In dark-field microscopy, oblique light rays produced by a specialized condenser illuminate the sample in a way which allows only the light diffracted by the specimen to enter the objective. In consequence, the sample appears white against a dark background. The phase-contrast technique takes advantage of the fact that unstained biological specimens act as phase objects, i.e. they retard the wave of light which passes through. Although human eyes or traditional cameras cannot detect these phase differences, a modification in microscopy optics is capable of converting the phase change into visible differences in amplitude. Differential interference contrast (DIC) is another technique relying on optical path differences and phase gradients experienced by light passing through a transparent specimen. Phase-contrast and DIC microcopy are mostly used to study living cells and tissue.

Fluorescence microscopy is probably the most important imaging technique employed by cell biologists. It relies on a huge number of natural and synthetic fluorescent labels capable of staining various regions of cells and tissue. These luminescent probes can be linked with monoclonal antibodies or other biological carriers, providing very high specificity of staining. Apart from specificity, fluorescence microscopy offers unparalleled sensitivity, being able to detect the presence of a single fluorescent molecule!

The majority of 3-D microscopy techniques used in cell biology are based on modified versions of the fluorescence detection modalities. Confocal fluorescence microscopy — arguably the most significant advance in optical microscopy in the 20th century — uses a pinhole in the light path to reject most of the out-of-focus light. The most commonly used variation of confocal microscopy is based on point scanning. In this approach, a laser beam, guided by galvanometer-driven mirrors, is deflected into a microscope objective and focused into the specimen. The fluorescence emitted from the specimen passes through the microscope objective, is focused onto a confocal pinhole, and is collected by a photon detector. Light that is emitted from locations either in front of or behind the focal point in the object is focused either in front of or behind the detector pinhole, and does not contribute to the measured signal. This way a thin optical slice (section) of the sample can be obtained. A 3-D representation of the specimen is generated by collecting a large number of

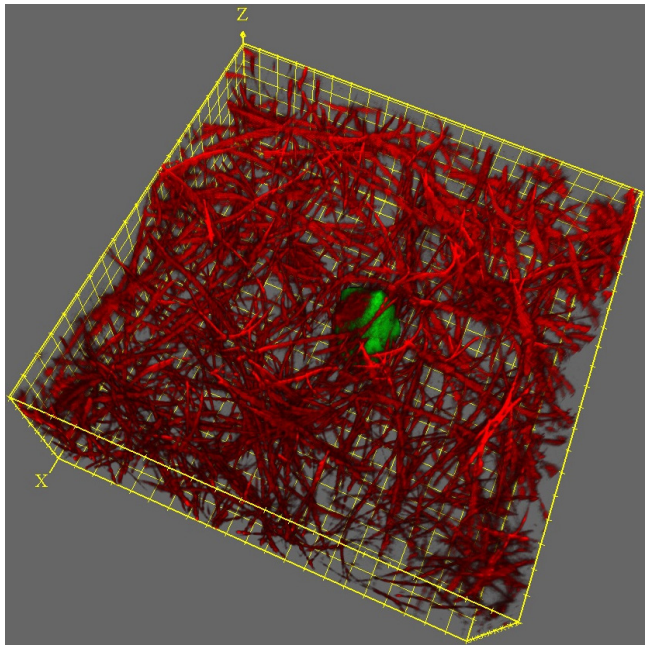


Fig. 1. 3-D confocal image of a liver cancer cell embedded in an extracellular matrix of collagen fibers.

such slices. Coupled with digital reconstruction techniques, confocal microscopy can help in better visualization of data, as shown in Fig. 1.

Multiphoton microscopy (MPM) represents another 3-D application of fluorescence imaging. MPM uses a laser to trigger a localized nonlinear fluorescence excitation process in which multiple low-energy photons can cause the same excitation typically produced by the absorption of a single high-energy photon. By focusing laser and raster scanning across a sample, MPM builds a 3-D map of intensities, collecting the fluorescence signal from a single voxel at a time. MPM has become the technique of choice for fluorescence microscopy of thick, highly scattering biological samples.

3. Multi-Layered Semantic Analysis

In recent years, quantitative imaging has been extensively used for observing biological phenomena [9]. The focus has traditionally been on extracting simple image and object features like size, intensity, etc. For example, the cell cycle is studied using a DNA-specific dye like a Hoechst dye [9]. Dividing cells generally have higher amounts of DNA than do non-dividing cells. A histogram of the amount of DNA is used to estimate the number of dividing and non-dividing cells. A representative image of a population of bovine aortal endothelial cells is shown in Fig. 2 and the corresponding histogram of intensity in the Hoechst channel is shown in Fig. 3.

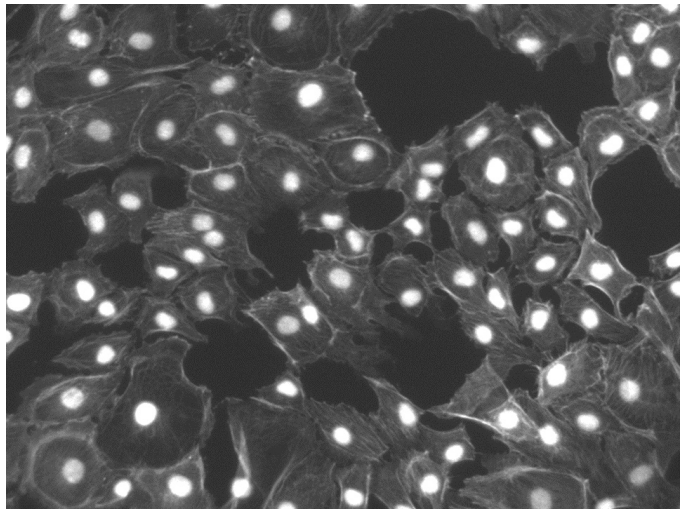


Fig. 2. A representative image of bovine aortal endothelial cells showing nuclear dye in white and cytoplasm dye in gray.

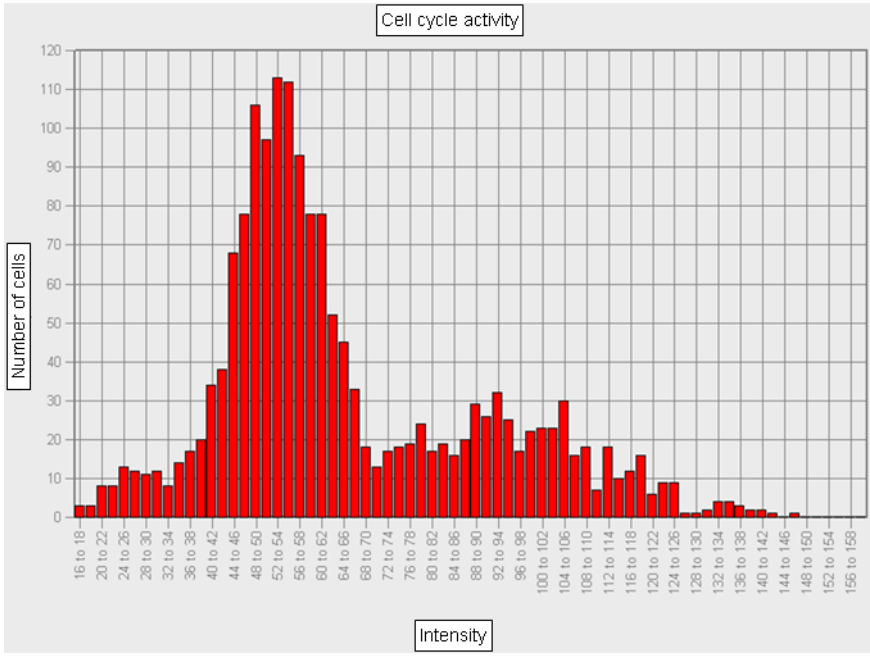


Fig. 3. Histogram of intensity in Hoechst channel for bovine aortal endothelial cells.

Such simple analysis, though useful for some cases, is not adequate for many applications where the dynamics of large populations of cells are to be studied. Such applications require extraction of high-level spatio-temporal semantics. Information content in biological images includes cell shape, texture, size, and location, as well as changes in these parameters as time progresses. Automatic image interpretation tools and bio-image database management systems are needed for efficient analysis and management of such data. Extraction of high-level semantics from biological images requires modeling of images at different levels. A multi-layered architecture for modeling of biological images is shown in Fig. 4. We use this figure to elaborate current approaches and challenges at each layer. The lower two layers of this architecture deal with object detection, tracking, and recognition, which are the most computationally intensive tasks. The upper three layers deal with high-level semantic interpretation of images and representation of spatio-temporal knowledge.

3.1. *Object detection and tracking layer*

This layer has two major functions. Firstly, it preprocesses the data to remove undesirable artifacts, and separates objects of interest from the background using segmentation algorithms. Secondly, it tracks the movements of biological objects in time-lapse images.

3.1.1. *Preprocessing and segmentation*

Imaging instruments are composed of different components. These include microscopes, cameras, filters, and lenses. All of these components add noise to biological images. Illumination may also vary from one image to another when a large population of cells is imaged [9]. Images need to be processed to remove this noise. The variable illumination problem for microscopic imaging has been addressed by providing methods for illumination normalization [9]. Microscope optics blur the image as a result of the limited aperture of the microscope objective [10]. Deconvolution

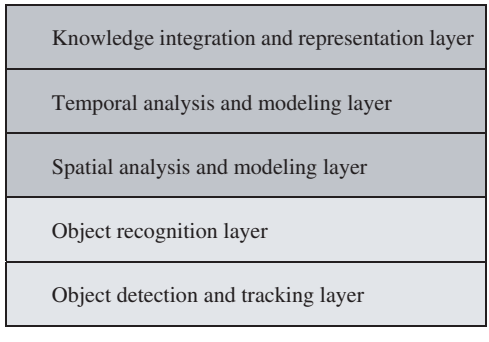


Fig. 4. Multi-layered architecture abstraction.

algorithms attempt to remove this blur and improve the contrast in microscopic images. Numerous deconvolution algorithms have been proposed in the literature [10, 11]. For time-lapse or multi-channel imaging, image registration issues also need to be addressed. Automated as well as interactive image registration algorithms have been proposed for this purpose [12].

The next step after preprocessing is the separation of objects from the background. Segmentation of biological images is a hard problem because of many complicating factors [9, 13]. There are large inter- and intra-cell variations. Pixel intensities can vary widely within cells. Different types of cells have very different shapes: for example, red blood cells are somewhat round whereas neurons and microtubules are long and branching. Even cells of the same type can look very different when imaged using different microscopic modalities. Poor contrast between objects and background is another complicating factor for many imaging modalities. In fluorescence imaging, different staining protocols can significantly change the appearance of cells. Touching and overlapping cells also complicate the problem. Different approaches have been taken for biological image segmentation in the literature [14–20]. Approaches based on deformable templates and active contours have been proposed in [15, 16]. Such techniques generally require an interactive initialization step and have poor performance for overlapping cells. For fluorescence imaging, it is generally easier to segment nuclei compared to cytoplasm because of the relatively uniform appearance and separation of nuclei from each other. Many algorithms segment cells by first segmenting the nuclei and then using them as seed points for segmenting cytoplasm. Jones *et al.* present a segmentation algorithm that uses nuclei as seed regions and locates the boundaries of cells by finding Voronoi regions around seed regions, guided by local image properties like edges [17]. Cell and nuclei segmentation algorithms based on the watershed transform have also been proposed [18–20]. Watershed algorithms generally suffer from the over segmentation problem and require a post processing step for improving segmentation accuracy.

Despite many efforts, accurate and fully automated segmentation still remains a challenge. As fluorescence imaging provides better contrast between objects and background, segmentation algorithms for fluorescence imaging generally perform better than those for bright-field and other imaging modalities. However bright-field imaging is more appropriate for many applications that require cells to be imaged for extended periods of time, as staining can have cytotoxic effects. Automated analysis of such data requires robust segmentation algorithms for all imaging modalities. As the shape and appearance of cells vary widely from one cell type to another and from one imaging modality to another, it is not likely that a single algorithm will perform well for all types of cells imaged using different modalities. It is therefore essential to develop segmentation algorithms tailored for specific applications. Such algorithms need to be robust and fully automated so that large sets of images can be quickly analyzed. Moreover, test data sets and performance measures need to be developed for objective comparison of different segmentation algorithms. Segmentation is the

first step for quantitative analysis and knowledge extraction. Defects at this level seriously affect the whole knowledge extraction process. Developments in this area would therefore have a significant impact on the overall semantic analysis process.

3.1.2. *Object tracking*

Extraction of temporal information requires tracking objects of interest in time-lapse images. Tracking of biological objects is a challenging problem and has attracted significant attention because of its use in many biological applications [13, 21–25]. The challenges arise from the fact that biological objects are not uniform, have large inter- and intra-object variability, and keep changing their shape with the passage of time. Tracking in the perspective of biological systems generally focuses on two themes, namely the tracking of particles like quantum dots and intracellular molecules [21], and the tracking of individual cells [13]. Different algorithms have been developed for these tracking applications. Cheezum *et al.* present a comparison of four particle-tracking algorithms in [22]. These include algorithms based on Gaussian fit, sum-absolute difference, centroid, and cross-correlation. They conclude that all of these algorithms perform poorly at very low signal to noise, although Gaussian fit performs relatively better than the others. Algorithms for cellular tracking include deformable models, mean shift, image level sets, and correlation-based algorithms. Deformable models have received relatively more attention than other methods as cells keep changing their shape over time [23–25]. Zimmer *et al.* present a comparison of some cell-tracking algorithms in [13]. They note that most of these algorithms require some level of user interaction for initialization. This interaction slows down the tracking process. Moreover, these algorithms are prone to serious tracking errors and thus require validation by an expert.

While it may be relatively easier to segment and track cells in fluorescent microscopy, it is very difficult to do so satisfactorily in bright-field microscopy because of poor contrast between cells and background. The problem of reliable unsupervised tracking of a large number of cells imaged using different imaging modalities is still unresolved. Extraction of spatio-temporal information from time-lapse image sequences will require the development of fully automated and robust tracking algorithms. Standardized test sets are also needed for objective comparison of different tracking algorithms. As in the case of segmentation, application-specific tracking algorithms are needed that meet the requirements of a given application. Moreover, integration of domain knowledge about the spatio-temporal dynamics of the objects with tracking algorithms can improve their performance.

3.2. *Object recognition layer*

This layer addresses the problem of identifying objects in the images. It extracts different features of objects and uses pattern recognition algorithms for identifying objects of interest. Templates of objects of interest are provided beforehand to train

the classifiers. The classifiers learn the object features and use them for classification of objects in image sets.

3.2.1. *Feature extraction*

An important step in image understating and interpretation is the extraction of image and object features. Different types of features can be used for this purpose. Simple features like size, area, and intensity, though useful for simple applications, have limited utility for more challenging applications like the ones described in Sec. 4. More complex descriptors such as texture, shape moments, and moment invariants can be useful for such applications. Image features can generally be divided into two categories, global and local. Global features, such as color histograms, transform the whole image into descriptors. Local features are generally based on objects inside the image. As most applications in cellular imaging and biological imaging in general are based on objects in the images, local features are more important. Most currently available cellular image analysis packages provide only simple features like size, intensity, roundness, etc. More complex features like boundary descriptors, region descriptors, and texture have been successfully used for protein localization and bacterial colony identification [26, 27]. Most boundary and region descriptors like moments and moment invariants have high computational cost; hence feature extraction becomes a bottleneck for high-throughput applications. It is therefore desirable to have fast implementations of these algorithms. Challenging biological applications will require development of new, fast, and more powerful features. This will require collaboration among biologists and image processing, computer vision and computational geometry experts.

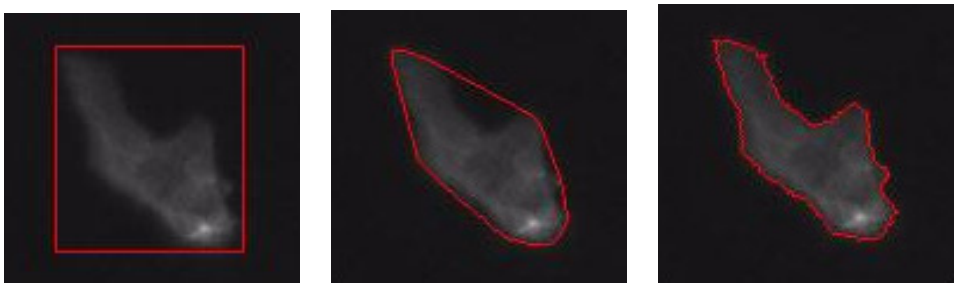
3.2.2. *Object recognition*

The objective of many biological applications is to identify cells, proteins, or other biological objects. These objects are usually defined in terms of their features. Pattern recognition algorithms are employed to identify these objects using their features. Pattern recognition is a mature field and numerous algorithms have been proposed for this purpose [28, 29]. These include decision trees, Bayesian classifiers, maximum likelihood classifiers, support vector machine (SVM) classifiers, and neural networks. In biological imaging, statistical pattern recognition techniques have been used for many successful applications. A neural network classifier capable of recognizing major sub-cellular proteins is reported in [26]. Classifiers for automated phenotyping have also been reported [30]. Our group has successfully used an SVM classifier for bacterial colony classification [27]. As noted by many experts in pattern recognition, no single algorithm performs best for all different scenarios [28]. It is therefore important to test different classifiers for a given application and to choose the one that meets the requirements of the application in terms of classification accuracy and speed.

3.3. *Spatial analysis and modeling layer*

Spatial events provide useful information about biological processes. The spatial analysis and modeling layer deals with two main issues. Firstly, it identifies patterns in the spatial distribution of cells, for example, neurons in brain tissue. Such patterns provide information about the behavior of cells. Changes in these patterns may indicate disease states [31]. Algorithms for spatial distribution try to find patterns in the spatial distribution of cells. These algorithms generally develop the neighborhood graph first and then analyze the spatial distribution using this graph [31]. Neighborhood graphs are developed using Voronoi diagrams and Delauny triangulation, or using methods based on k-nearest-neighbor graphs or other appropriate approaches. Spatial analysis is then carried out on these graphs using graph matching algorithms. Such analysis yields information about the patterns in spatial arrangement of cells in tissue and about the changes in their distribution in response to different diseases or abnormalities. Secondly, this layer deals with inter-object spatial relationships. Spatial relations among subcellular compartments and different particles and proteins provide information about the localization of these objects inside cells. For example, finding the cellular compartment in which specific drugs or nanoparticles are localized requires establishing spatial relationships between different biological objects in the image. Such spatial modeling also helps in content-based retrieval of biological imaging data. Minimum bounding rectangle (MBR) is the most commonly used spatial representation for objects. MBR, however, is not very accurate, as shown in Fig. 5. Convex hull- and exact outline- based representations along with polygon intersection tests can be used for more accurate spatial analysis of biological objects [32].

Although algorithms for spatial analysis exist, newer and faster approaches will be required for large biological image repositories to cope with the large volume of data. Moreover, biological applications vary in their demands for speed and accuracy. While accuracy may be the most important factor for small-scale biology, speed is also important for high-throughput imaging. It is therefore required that



(a) (b) (c)

Fig. 5. Different spatial representations. (a) MBR, (b) Convex hull, (c) Exact outline.

the analysis algorithms be able to make trade-offs between speed and accuracy so that they can be adapted to the requirements of particular biological applications.

3.4. *Temporal analysis and modeling layer*

Biological cells and organisms are dynamic systems that evolve over time. Monitoring changes in different observable cell parameters such as size, color, shape and texture can help detect and quantify cellular events. For example, time-lapse imaging has been used to study the cell cycle [33]. Similarly, temporal texture parameters have been used for protein localization [34]. Analysis of spatio-temporal information associated with cellular systems requires development of models for spatio-temporal knowledge representation. Models based on finite state machine (FSM) and Petri-nets have been proposed for such analyses [32]. The FSM approach models cells in terms of their attributes like size, shape, texture, and spectrum. Spatio-temporal events in this approach are modeled in terms of the participating objects along with specific values of their attributes and spatial relations between them. This model is able to capture a variety of spatio-temporal biological events. For example, changes in the morphology of cells can be modeled by the changes in their observable attributes. Similarly, events like cell division, apoptosis, and phagocytosis can be modeled in terms of attributes of cells and their spatial relations. For more complicated events, for example multi-threaded events involving multiple objects with temporal constraints between constituent events, Petri-nets are used. A Petri-net is a directed bipartite graph which is used for modeling distributed systems. A Petri-net contains place nodes, transition nodes, and directed arcs. Place nodes contain information about events and may be composed of simpler events. They may also contain delay nodes to express timing constraints among different events. This way, spatio-temporal constraints for a complex sequence of events can be represented.

Spatio-temporal modeling of biological images is crucial for extracting high-level semantics from such data, which in turn is essential for understanding biological processes. While significant work has been done on low-level image processing, development of models for extracting high-level semantics from biological images has lagged behind. There is an urgent need for developing powerful models that can capture the semantics of a wide variety of biological objects and events. Development of such models will require interdisciplinary collaborations between biologists and experts from computer vision, multimedia, and knowledge representation communities.

3.5. *Knowledge integration and representation layer*

This layer deals with the integration and representation of spatio-temporal knowledge extracted by lower layers. The extracted spatio-temporal knowledge can be analyzed by data mining algorithms to discover association rules and hidden patterns in this data. This conceptual information can then be stored in a database along with spatial and temporal information. This high-level knowledge can subsequently be used for developing simulation models for biological processes.

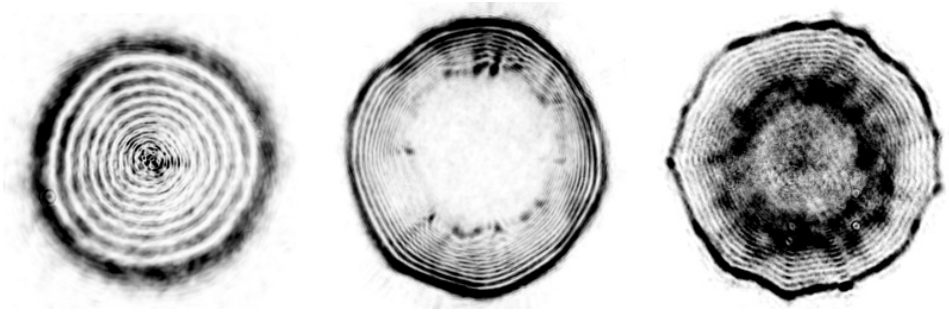
The biological data that are routinely collected contain much more information than currently available analysis tools can extract or analyze. Even though effective tools have been developed for multimedia databases in the past [35], powerful semantic data modeling techniques and content based retrieval databases still do not exist for biological imaging. Open microscopy environment (OME) is a step towards a standard data storage format that stores the context information and analysis results along with the data [36]. OME defines XML schemas for this purpose. These schemas define the context information associated with the experiment, such as who performed the imaging and under what conditions. OME also defines schemas for processing algorithms that were used to analyze the data, and for the results of such analyses. Similarly, protein subcellular location image database is an effort towards content based retrieval for subcellular protein images [37]. Yet many challenges remain to be addressed. These include the development of tools for automatic indexing, spatio-temporal knowledge extraction, fast content-based retrieval, and mining of biological imaging data. Development of these tools will significantly improve the analysis and management of such data.

4. Current Applications of Semantic Analysis

Semantic analysis of biological imaging data has been used for a variety of successful applications. In this section, we briefly survey some of these applications. These include bacterial colony identification, location proteomics, and drug discovery.

4.1. *Bacterial colony identification*

Bacterial contamination of food products puts the public at risk and generates a substantial cost for the food-processing industry. One of the greatest challenges in the response to incidents of bacterial contamination is rapid recognition of the bacterial agents involved. Only a few currently available technologies allow testing to be performed outside of specialized microbiological laboratories [38]. Most current systems are based on the use of expensive techniques like polymerase chain reaction (PCR) or antibody-based techniques, and require complicated sample preparation for reliable results. A system for automatic interpretation and classification of scatter patterns produced by bacterial colonies irradiated with red laser light is reported in [27]. The optical measurements are performed with a laser scattering system. A CCD image sensor is used to acquire scatter patterns like those shown in Fig. 6. The laser generates a collimated beam of light that is directed through the center of the bacterial colony and the substrate of agar medium. The forward-scattered light and the transmitted light form scatter patterns directly on the detector. Different types of features, including Zernike and Chebyshev moments and Haralick texture descriptors, are extracted from these scatter patterns. A subset of these features with the highest discriminative power is selected for classification with the help of an SVM algorithm. This approach offers a solution to the problem of rapid discrimination



(a) (b) (c)

Fig. 6. Examples of scatter patterns formed by (a) Listeria monocytogenes 19113, (b) Salmonella ser. enteritidis PT28, and (c) Vibrio fluvialis.

of cultured organisms in environments which do not allow biochemical processing of the samples, but require automation, robustness, and simplicity.

4.2. *Location proteomics*

Knowledge about the production, location, and spatio-temporal dynamics of various proteins expressed in a cell is crucial for understanding cell behavior. Acquiring and analyzing this information for all proteins expressed in different types of cells under different developmental and environmental conditions is extremely challenging because of the sheer volume of data [39]. Manual analysis of such high-volume data is not feasible. Recently, intelligent tools have been developed for automated interpretation of protein sub-cellular location patterns [26, 34, 40, 41]. Boland *et al.* propose to extract image features from fluorescence images of subcellular location patterns and employ a neural network to classify such patterns [26]. The features used for this analysis include shape, texture, and morphological features. As many proteins are in motion inside cells, this temporal information can also be useful for localizing protein location patterns. This analysis, however, is complicated by the fact that protein patterns are normally not fully connected, so simple tracking methods do not work very well. Hu *et al.* address this problem by using temporal texture features to classify location patterns of five very similar proteins, using the spatial and temporal information provided by 3-D time-lapse imaging [34]. Another important issue is the comparison and systematic analysis of location patterns. Chen *et al.* use a set of features to build a tree of proteins based on the similarity of their location patterns [40]. As the collection of data for a large number of proteins for different cell types under different conditions leads to a combinatorial explosion, high-throughput image acquisition also becomes a challenge and requires automated instrumentation for this purpose. Schubert *et al.* deal with this problem by using a robotic platform for finding the proteins expressed in a cell by repeated

staining [41]. The combined thrust of these efforts to improve acquisition and interpretation of protein localization information will significantly help in understanding and simulating cell behavior.

4.3. *Drug discovery*

HCS technologies have emerged as a promising solution for compound lead selection and drug target validation [42]. These technologies image large populations of cells under different experimental conditions for understanding cell behavior. For example, monitoring cell cycle progress yields information about the effects of drugs on cancer cells [33]. Chen *et al.* present an automated system for segmenting, tracking, and classifying cells during different phases of the cell cycle in [33]. They use automatic thresholding, a watershed algorithm, and shape and size information to segment touching and overlapping nuclei. A set of seven features and a k-nearest neighbor classifier are then used for cell phase identification. Some errors introduced by this classification approach are removed by using knowledge-based heuristic rules. DNA microarray imaging is another powerful technology for studying expression patterns of genes in a high-throughput manner [43]. These technologies can help in identifying targets for drugs by studying gene function. Image analysis is an important step for extracting information from microarray images. A comparison of different techniques for addressing, segmentation and background correction for microarray image analysis appears in [43]. A high-throughput drug profiling technique using automated microscopy and image analysis is reported in [44]. The authors consider the effect of drug concentration on cell phenotypes using a titration invariant similarity score. The information about cell states, genes, and protein regulatory networks and their changes can help in understanding biological systems and developing simulation models for their behavior. These models and detailed information about genes, proteins, cells, organs, and organisms have the potential to improve our understanding of disease mechanisms. This can not only result in improvements in drug discovery process but can also potentially change medical practice from reactive medicine to preventive and personalized medicine.

5. A Component-Based System for Semantic Analysis

Based on the discussion in the previous sections, we present a component-based system for knowledge-based semantic analysis of biological imaging data as shown in Fig. 7. XML schemas can be used to store domain knowledge at different layers of the system. We have presented schemas for representing biological objects and spatio-temporal events in [45]. The user interacts with the system through a graphical user interface (GUI). The GUI provides an XML editor that is used for describing the attributes of objects and events manually. Alternatively, a feature extraction module is provided that can be used for specifying objects and events. The user provides example images of objects and events and the feature extractor automatically

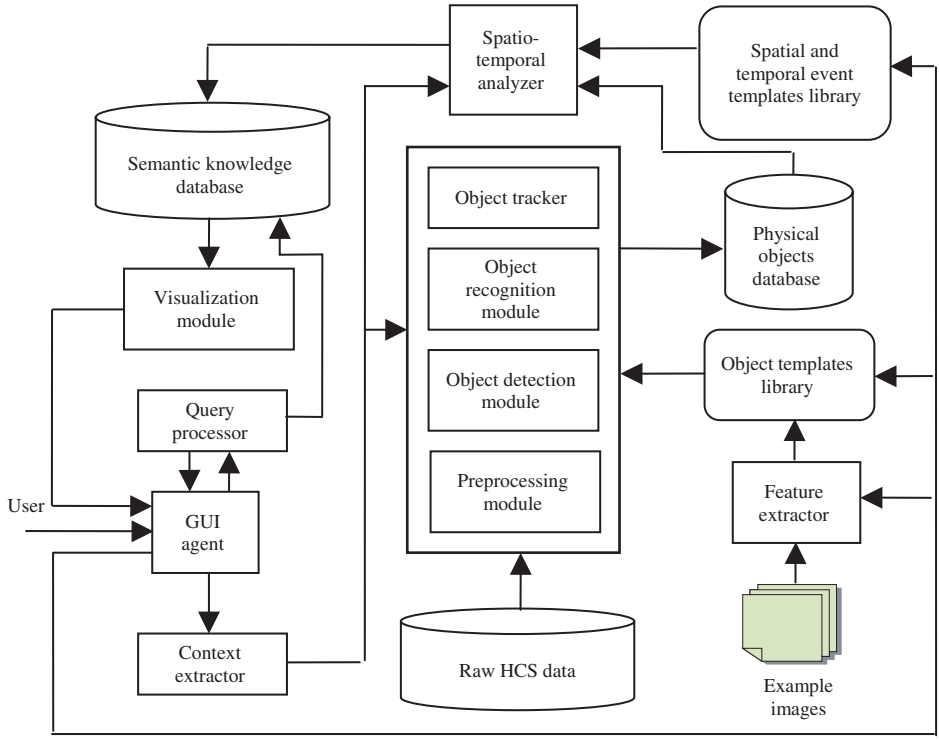


Fig. 7. Component-based system for semantic analysis of biological images.

extracts the features and generates XML representations that are stored in object and event templates libraries. The context extractor module collects the context information from the user. The user context information as well as the information in object and event templates libraries is used by object and event recognition tools for identifying their specific instances in image sets. Low-level object information is extracted using preprocessing, object detection, recognition, and tracking modules. These low-level object features for all the objects found in the image set are stored in the physical objects database. This object-level information is then analyzed by the spatio-temporal analyzer. The spatial and temporal event templates library holds information about the events of interest in the form of XML schemas. This library is populated either manually using the XML editor or by providing example images for events of interest. The high-level spatio-temporal knowledge extracted by the spatio-temporal analyzer is stored in the semantic knowledge database as XML schemas. The visualization module is used for summarization and visualization of the extracted semantic knowledge. The user queries the semantic knowledge database using the query processor. Textual as well as visual query languages can be used for this purpose.

6. Summary

In this paper we have outlined major challenges and current approaches for semantic analysis of biological imaging data. Modern biological imaging technologies produce huge volumes of imaging data. Powerful knowledge extraction and data management tools are required for analyzing these data. Intelligent semantic interpretation tools have already started to have an impact on biological data analysis. However, there are significant challenges to be overcome. We summarize these as challenges at the lower and the higher levels of processing. At the lower level, powerful and robust algorithms are needed for segmentation and tracking of biological objects. Owing to the heterogeneity of biological systems, specialized algorithms tailored to the needs of specific applications will be required. Standardized test sets and performance measures will also be needed for objective and quantitative comparison of different algorithms. Powerful and computationally efficient feature extraction algorithms will be essential for robust classification of biological objects. Since the low-level processing is the most computationally intensive, all the algorithms at this level must have low computational complexity so that large biological image sets can be processed in a reasonable amount of time. At the higher level, models and formalisms will be required for extracting and representing spatio-temporal semantics from biological images. These models must be powerful enough to capture the semantics of a wide variety of biological objects and events. Such models should also support interoperability. XML-based schemes will be useful to enable interoperability of such models among different research groups. Moreover, efficient database management systems will be needed for storage and fast content-based retrieval of these data. It is clear that overcoming these challenges will require collaborative efforts by biologists and computer vision, image processing, and knowledge representation experts. The benefits of this multidisciplinary approach will include a more detailed understanding of biological systems, significant improvements in the drug design process, and the resultant predictive role of information technology in preventive medicine.

References

[1] H. Kitano, Systems biology: A brief overview, *Science* **295**(5560) (2002) 1662–1664.
[2] E. C. Butcher, E. L. Berg and E. J. Kunkel, Systems biology in drug discovery, *Nature Biotechnology* **22**(10) (2004) 1253–1259.
[3] D. J. Stephens and V. J. Allan, Light microscopy techniques for live cell imaging, *Science* **300**(5616) (2003) 82–86.
[4] X. Zhou and S. T. C. Wong, Informatics challenges of high-throughput microscopy, *IEEE Signal Processing Magazine* **23**(3) (2006) 63–72.
[5] K. A. Giuliano, High-content screening: A new approach to easing key bottlenecks in the drug discovery process, *Journal of Biomolecular Screening* **2**(4) (1997) 249–259.
[6] A. Rodriguez, N. Guil, D. M. Shotton and O. Trelles, Automatic analysis of the content of cell biological videos and database organization of their metadata descriptors, *IEEE Trans. on Multimedia* **6**(1) (2004) 119–128.

[7] J. R. Swedlow, I. Goldberg, E. Brauner and P. K. Sorger, Informatics and quantitative analysis in biological imaging, *Science* **300**(5616) (2003) 100–102.
[8] C. Vonesch, F. Aguet, J. Vonesch and M. Unser, The colored revolution of bioimaging, *IEEE Signal Processing Magazine* **23**(3) (2006) 20–31.
[9] T. R. Jones, A. E Carpenter, P. Golland and D. M. Sabatini, Methods for high-content, high-throughput image-based cell screening, in *Proc. of the Workshop on Microscopic Image Analysis with Applications in Biology (MIAAB)*, Copenhagen, Denmark, 2006, pp. 65–72.
[10] W. Wallace, L. H. Schaefer and J. R. Swedlow, A workingperson's guide to deconvolution in light microscopy, *Biotechniques* **31** (2001) 1076–1097.
[11] J. G. McNally, T. Karpova, J. Cooper and J. A. Conchello, Three-dimensional imaging by deconvolution microscopy, *Methods* **19** (1999) 373–385.
[12] B. Zitova and J. Flusser, Image registration methods: a survey, *Image and Vision Computing* **21** (2003) 977–1000.
[13] C. Zimmer *et al.*, On the digital trail of mobile cells, *IEEE Signal Processing Magazine* **23**(3) (2006) 54-62.
[14] P. Bamford, Empirical comparison of cell segmentation algorithms using an annotated dataset, in *Proc. Int. Conf. on Image Processing*, 2003, pp. II-1073–1076.
[15] A. Garrido, N. Perez de la Blanca, Applying deformable templates for cell image segmentation, *Pattern Recognition* **33** (2000) 821–832.
[16] M. Hu, X. Ping and Y. Ding, Automated cell nucleus segmentation using improved snake, in *Proc. International Conference on Image Processing*, 2004, pp. 2737–2740.
[17] T. R. Jones, A. E. Carpenter and P. Golland, Voronoi-based segmentation of cells on image manifolds, in Proc. of the ICCV Workshop on Computer Vision for Biomedical Image Applications (CVBIA), Lecture Notes in Computer Science 3765 (Springer-Verlag, Berlin), pp. 535–543.
[18] C. Wahlby, I. -M. Sintorn, F. Erlandsson, G. Borgefors and E. Bengtsson, Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections, *Journal of Microscopy* **215** (2004) 67–76.
[19] Y. Wang, Y. Sun, C. K. Lin and M. Ju, Nerve cell segmentation via multi-scale gradient watershed hierarchies, in *Proc. of the 28th IEEE EMBS Annual Internationa Conference*, Aug.–Sept. 2006.
[20] N. Malpica, C. O. 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** (1997) 289–297.
[21] E. Meijering, I. Smal and G. Danuser, Tracking in molecular bioimaging, *IEEE Signal Processing Magazine* **23**(3) (2006) 46–53.
[22] M. K. Cheezum, W. F. Walker and W. H. Guilford, Quantitative comparison of algorithms for tracking single fluorescent particles, *Biophysical Journal* **81**(4) (2001) 2378–2388.
[23] C. Zimmer, E. Labruyere, V. Meas-Yedid, N. Guillen and J. Olivo-Marin, Segmentation and tracking of migrating cells in video microscopy with parametric active contours: A tool for cell-based drug testing, *IEEE Trans. on Medical Imaging* **21**(10) (2002) 1212–1221.
[24] A. Dufour, V. Shinin, S. Tajbakhsh, N. Guillen, J. Olivo-Marin and C. Zimmer, Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces, *IEEE Trans. on Image Processing* **14**(9) (2005)1396–1410.
[25] N. Ray, S. Acton and K. Ley, Tracking leukocytes in vivo with shape and size constrained active contours, *IEEE Trans. on Medical Imaging* **21**(10) (2002) 1222–1235.

[26] M. V. Boland and R. F. Murphy, A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells, *Bioinformatics* **17**(12) (2001) 1213–1223.
[27] B. Bayraktar, P. P. Banada,, E. D. Hirleman, A. K. Bhunia , J. P. Robinson and B. Rajwa, Feature extraction from light-scatter patterns of Listeria colonies for identification and classification, *Journal of Biomedical Optics* **11**(3) (2006) 34006.
[28] A. K. Jain and R. P. W. Duin, J. Mao, Statistical pattern recognition: a review, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22**(1) (2000) 4–37.
[29] A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, *ACM Computing Surveys* **31**(3) (1999) 264–323.
[30] B. Neumann, M. Held, U. Liebel, H. Erfle, P. Rogers, R. Pepperkok and J. Ellenberg, High-throughput RNAi screening by time-lapse imaging of live human cell, *Nature Methods* **3**(5) (2006) 385–390.
[31] R. Fernandez-Gonzalez, M. H. Barcellos-Hoff and C. Oritz-de-Solorzano, A tool for the quantitative spatial analysis of complex cellular systems, *IEEE Trans. on Image Processing* **14**(9) (2005).
[32] W. M. Ahmed, A. Ghafoor and J. P. Robinson, FSM-based model for spatio-temporal event recognition for HCS, Technical Report TR-ECE 07-04, Purdue University, February 2007.
[33] X. Chen, X. Zhou and S. T. C. Wong, Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy, *IEEE Trans. on Biomedical Engineering* **53**(4) (2006) 762–766.
[34] Y. Hu, J. Carmona and R. F. Murphy, Application of temporal texture features to automated analysis of protein subcellular locations in time series fluorescence microscope images, in *Proc. of the 2006 IEEE International Symposium on Biomedical Imaging (ISBI 2006)*, pp. 1028–1031.
[35] W. Al-Khatib, Y. F. Day and A. Ghafoor, Semantic modeling and knowledge representation in multimedia databases, *IEEE Trans. on Knowledge and Data Engineering* **11**(1) (1999) 64–80.
[36] I. G. Goldberg, C. Allan, J. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P. K. Sorger and J. R. Swedlow, The open microscopy environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging, *Genome Biology* **6**(5) (2005) R47.
[37] K. Huang, J. Lin, J. A. Gajnak and R. F. Murphy, Image content-based retrieval and automated interpretation of fluorescence microscope images via the protein subcellular location image database, in *Proc. 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002)*, pp. 325–328.
[38] D. R. Call, M. K. Borucki and F. J. Loge, Detection of bacterial pathogens in environmental samples using DNA microarrays, *Journal of Microbiological Methods* **53**(2) (2003) 235–243.
[39] R. Murphy, Putting proteins on the map, *Nature Biotechnology* **24** (2006) 1223–1224.
[40] X. Chen, M. Velliste, S. Weinstein, J. W. Jarvik and R. F. Murphy, Location proteomics — Building subcellular location trees from high resolution 3D fluorescence microscope image of randomly-tagged proteins, in *Proceedings of SPIE* **4962** (2003) 298–306.
[41] W. Schubert, B. Bonnekoh, A. J. Pommer, L. Philipsen, R. Bockelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode and A. W. M. Dress, Analyzing proteome topology and function by automated multidimensional fluorescence microscopy, *Nature Biotechnology* **24** (2006) 1270–1278.

[42] X. Zhou and S. T. C. Wong, High content cellular imaging for drug development, *IEEE Signal Processing Magazine* **23**(2) (2006) 170–174.

[43] Y. H. Yang, M. J. Buckley, S. Dudiot, and T. P. Speed, Comparison of methods for image analysis on cDNA microarray data, *Journal of Computational and Graphical Statistics* **11**(1) (2002) 108–136.

[44] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu and S. J. Altschuler, Multidimensional drug profiling by automated microscopy, *Science* **306**(5699) (2004) 1194–1198.

[45] W. M. Ahmed, A. Ghafoor and J. P. Robinson, An XML-based system for providing knowledge-based grid services for high-throughput biological imaging, Technical Report TR-ECE 07-03, Purdue University, February 2007.