

2-21-2007

An XML-based System for Providing Knowledge-based Grid Services for High-throughput Biological Imaging

Wamiq Manzoor Ahmed
Purdue University, wahmed@purdue.edu

Dominik Lenz
Purdue University

Jia Liu
Purdue University

J. Paul Robinson
Purdue University

Arif Ghafoor
Purdue University, ghafoor@purdue.edu

Follow this and additional works at: <http://docs.lib.purdue.edu/ecetr>

Ahmed, Wamiq Manzoor; Lenz, Dominik; Liu, Jia; Robinson, J. Paul; and Ghafoor, Arif, "An XML-based System for Providing Knowledge-based Grid Services for High-throughput Biological Imaging" (2007). *ECE Technical Reports*. Paper 344.
<http://docs.lib.purdue.edu/ecetr/344>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

An XML-based system for providing knowledge-based grid
services for high-throughput biological imaging

Wamiq Manzoor Ahmed
Dominik Lenz
Jia Liu
J. Paul Robinson
Arif Ghafoor

TR-ECE-07-03

School of Electrical and Computer Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN, 47907-1285

Table of contents

| | |
|---------------------------------------------------------------|----|
| 1. Introduction..... | 5 |
| 1.1 High throughput fluorescence microscopic imaging | 7 |
| 1.2 Examples of spatio-temporal cellular events | 8 |
| 1.3 Related work | 9 |
| 2. Knowledge-based grid for HCS..... | 11 |
| 3. Spatio-temporal event recognition and representation | 15 |
| 4. Architecture of the XML-based grid for HCS | 17 |
| 4.1 Spatio-temporal Data model | 18 |
| 4.2 Representation of objects..... | 20 |
| 4.3 Representation of spatial and temporal events..... | 21 |
| 4.4 Representation of analysis results..... | 24 |
| 5. An example of event representation | 25 |
| 6. Prototype and discussion..... | 28 |
| 7. Conclusion | 32 |
| References..... | 33 |

List of Figures

| | |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1. Cell nuclei and cytoplasm stained with different dyes | 9 |
| Figure 2. Hierarchy of grid services..... | 12 |
| Figure 3. HCS workflow utilizing grid services..... | 13 |
| Figure 4. Services provided by the proposed architecture | 14 |
| Figure 5. Graphical representation of three different spatial relations (a) Disjoint, (b) Overlap, (c) Contain..... | 16 |
| Figure 6. Architecture of the grid-enabled component-based system for HCS | 17 |
| Figure 7. Spatio-temporal data model..... | 19 |
| Figure 8. Representation of a cancer cell..... | 20 |
| Figure 9. Representation of a cell division event | 22 |
| Figure 10. Representation of the expansion phase in a cell division event | 22 |
| Figure 11. Representation of the nuclear division phase in a cell division event. | 23 |
| Figure 12. Graphical representation of sub events for cell division..... | 23 |
| Figure 13. Temporal relations | 24 |
| Figure 14. Representation of events involved in an apoptosis screen | 27 |
| Figure 15. Representation of analysis results for the apoptosis screen..... | 28 |
| Figure 16. Graphical interface for objects and events specificaiton..... | 30 |
| Figure 17. A representative set of images for the apoptosis screen showing HL60 cells in different states. (Upper left) Hoechst33342, (Upper right) Annexin V FITC, (Lower left) PI, (Lower right) Merged image. Images were aquired using an E1000 fluorescence microscope (Nikon, Tokoyo, Japan) using a 60x 1.4 NA lens. Fluorescence images of the PI, Annexin V FITC, and Hoechst 33342 were sequentially aquired with a RETIGA EXi Cooled monochrome 12-bit camera (QImaging, Burnaby, BC Canada). Image aquisition and processing was performed using Image- Pro Plus (MediaCybernetics, Silver Spring, MD). The three images were pseudo- colored and merged to make the composite image. Blue, green and red colors correspond to Hoechst33342, Annexin V FITC and PI respectively. . | 31 |

An XML-based system for providing knowledge-based grid services for high-throughput biological imaging

Wamiq Manzoor Ahmed^{1,2}, Dominik Lenz², Jia Liu², J. Paul Robinson², Arif Ghafoor¹

¹School of Electrical and Computer Engineering, 465 Northwestern Avenue,

²Purdue University Cytometry Laboratories, 1203 West State Street,
West Lafayette, IN 47907

Abstract

High-throughput biological imaging uses automated imaging devices to collect a large number of microscopic images for analysis of biological systems and validation of scientific hypotheses. Efficient manipulation of these data sets for knowledge discovery requires high performance computational resources, efficient storage, and automated tools for extracting and sharing such knowledge among different research sites. Newly emerging grid technologies provide powerful means for exploiting the full potential of these imaging techniques. Efficient utilization of grid resources requires the development of knowledge-based tools and services that combine domain knowledge with analysis algorithms. In this paper we first investigate how grid infrastructure can facilitate high-throughput biological imaging research, and present an architecture for providing knowledge-based grid services for this field. We identify two levels of knowledge-based services. The first level services provide tools for extracting spatio-temporal knowledge from image sets and the second level provides high-level knowledge management and reasoning services. We then present cellular imaging markup language (CIML), an XML-based language for modeling of biological images and representation of spatio-temporal knowledge. This scheme can be used for spatio-temporal event composition, matching, and automated knowledge extraction and representation for large biological imaging data sets. We demonstrate the expressive power of this formalism by means of different examples and experimental results.

1. Introduction

Microscopic imaging is one of the most popular means for studying biological processes. Recent advances in microscopic and optical instrumentation and automation technologies have made sophisticated imaging tools available for the biologist [1]. The rapidly evolving field of high content/high throughput screening (HCS/HTS) uses automated imaging instruments for image-based analysis of large populations of cells, to monitor their spatio-temporal behavior under different experimental conditions [30]. These technologies provide valuable information about biological processes and can help in generating rich information bases about different cellular systems [31]. These information bases can in turn help in developing better systems-level understanding of biological organisms, and can also aid in developing new drugs. In order to exploit the full potential of these technologies, many challenges have to be overcome. Analysis of a large number of cells for validating scientific hypotheses requires high-throughput computing resources, efficient management of datasets and knowledge-based image understanding tools that can rapidly identify relevant biological objects and events and extract high-level semantic information [2, 4]. Significant progress has been made in recent years in the areas of distributed computing and distributed data management using grid technologies [33]. The availability of basic grid services provides a great opportunity for developing knowledge-based applications and services for high throughput biological imaging. This requires the development of tools and services for spatio-temporal knowledge representation and extraction. Development of such tools and knowledge representation schemes not only would expedite the analysis of high-throughput imaging research but would also make the new knowledge generated in the process available for the design of other experiments. It is increasingly clear that availability of such knowledge would significantly impact future biological research [34].

Grid technologies are a promising solution for the challenging demands of high-throughput biological imaging research. Computational grids have provided high-throughput computing resources for compute-intensive applications in different scientific domains [20]. Data grids have provided elegant solutions for the management of distributed data sources for applications that involve large volumes of data [20]. The combined potential of enormous computational and data resources can be best utilized by knowledge-based tools and services that can extract high-level knowledge from huge data sets and can help researchers in making informed decisions and thus aid in the process of biological discovery. These tools and services can allow automated extraction of biological knowledge and high-level information-rich metadata from large image sets and make them available to other researchers. The maturation of computational and data grids provides an opportunity for applying the powerful services provided by these technologies for solving challenging problems in the domain of HCS. A key challenge for data mining and knowledge discovery in large HCS image sets is the extraction of semantic information about cell states in terms of their spatio-temporal dynamics. Once this information is extracted, advanced data mining techniques can be applied for discovery of new knowledge from diverse repositories of these data. These issues are addressed in this paper. The specific contributions of this paper are as follows,

- (i) We investigate how grid infrastructure can facilitate high-throughput biological imaging research, and present an architecture for providing knowledge-based grid services for this field that builds on basic computing, communication and data grid services. The proposed architecture provides two layers, a knowledge extraction (KE) and a knowledge management (KM) layer. The KE layer deals with algorithms for extraction of spatio-temporal knowledge from images whereas the KM layer deals with higher-level knowledge representation and reasoning.

(ii) We introduce a data model and an XML-based language for specifying objects, and spatial and temporal events. This specification of spatio-temporal knowledge can be used by object and event identification tools to extract spatio-temporal knowledge from image sets. Schemas are also provided for the representation of this automatically extracted knowledge which makes this information-rich metadata available to other researchers in a machine-readable format.

There are a number of benefits to this approach. Firstly, it can speed up the analysis of large image data sets by providing grid-enabled automated tools for extraction of spatio-temporal knowledge that can help researchers in understanding biological systems. Secondly, these tools can also be used for automatic extraction of spatio-temporal metadata and representation of these metadata at varying levels of detail in a machine-readable format so that this knowledge becomes available to other researchers. Thirdly, it enhances the opportunity for applying advanced data mining tools to large image sets for identifying hidden patterns in such data.

1.1 High throughput fluorescence microscopic imaging

Fluorescence imaging is one of the most useful tools in the arsenal of biologists, as it enables the visual examination of biological systems using fluorescent markers [1, 3]. As many biological samples are relatively transparent, contrast agents called markers or probes are used to highlight cell compartments of interest. These markers are frequently fluorescent molecules which have specific absorption and emission spectra and they emit light in their emission spectrum when excited within their appropriate absorption wavelengths. Optical filters are used to collect light in particular spectral bands. Figure 1 shows cell nuclei stained with a Hoechst dye and cell cytoplasm stained with Alexa488 dye. Similarly, different cell locations can be stained with a variety of markers and this information is collectively used for understanding biological

phenomena. Recent advances in automation technologies combined with microscopic imaging techniques have given rise to the highly promising field of HCS. HCS technologies image large populations of cells in an automated fashion [30]. They allow the examination of cell populations during and subsequent to different experimental perturbations. A key feature of biological systems is their spatial and temporal dynamics in response to variations in experimental conditions. For example, location of different proteins inside cells reveals important information about their function, and changes in the phenotype of cells are important for understanding different cell states. Monitoring this spatio-temporal dynamics is crucial to understanding biological systems.

1.2 Examples of spatio-temporal cellular events

Many biological events involve spatial and temporal changes in different properties of cells and sub-cellular objects. Apoptosis is one such event [5]. Apoptosis is defined as programmed cell death. Every living cell spends its life cycle according to a set of instructions and at the end it initiates the process of apoptosis. It will be used as a detailed example in section 5 and is therefore explained more fully here. Studies of apoptosis are important for cancer cell research. Cancerous cells often avoid apoptosis and instead continue to divide without control. Many anti-cancer drugs are designed to induce apoptosis in cancer cells to stop the expansion of a tumor for example. Apoptosis can be detected by using a specific surface marker, for example Annexin V fluorescein isothiocyanate (FITC), as well as other probes for staining the nucleus and cytoplasm. Another fluorescent nuclear stain called propidium iodide (PI), is used to verify cell viability. This dye cannot penetrate the intact cell membrane of healthy cells, so it is not found in nuclei. As the cell undergoes apoptosis, its cell membrane becomes more permeable and PI stains the nucleus. Cell division is another example of such events. Study of the effect of different perturbations on cell division is also useful for monitoring the progression of cancer cells. Cell

division involves enlargement of cell followed by nuclear division and the formation of daughter cells.

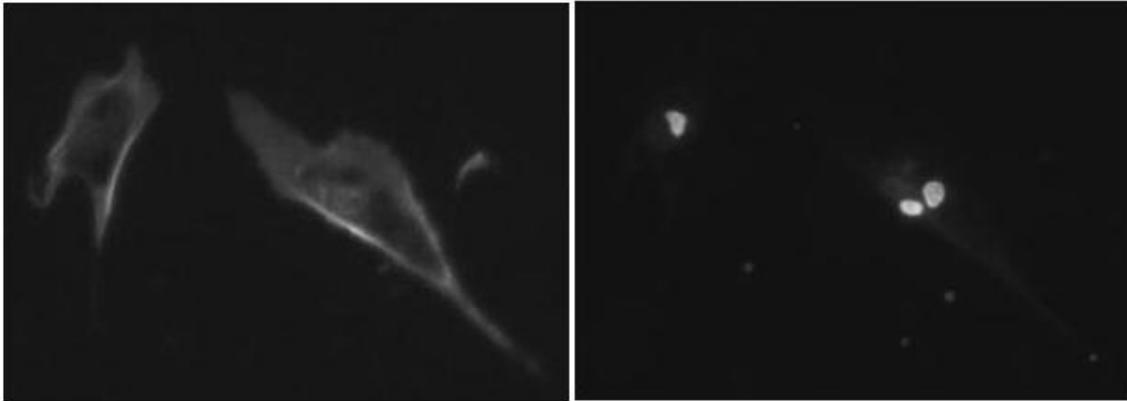


Figure 1. Cell nuclei and cytoplasm stained with different dyes

In addition to observing physical changes in cells, more subtle alterations within cellular metabolic states can be observed by monitoring the production and location of various proteins. Proteomics, which is the study of the structure, function and interactions of proteins produced by the genes of a particular cell, relies heavily on the location of different proteins inside cells. The function of proteins depends on the location where they are expressed [26]. Automated tools for localization and monitoring the dynamics of proteins can be very helpful for the field of proteomics [27].

1.3 Related work

Grid technologies have been applied to many problems in biological and medical sciences [20, 36]. Biomedical informatics research network (BIRN) focuses on applying grid services for neurological problems [36]. A grid-enabled image processing toolkit has been proposed in [25]. OpenMolGRID project aims at providing grid-enabled applications for molecular science and engineering [36]. All of these projects build on basic grid services and apply grid technologies to particular domains. Like these projects our aim is to apply the computing, communication and

data services provided by the grid to a specific domain (HCS). Additionally, we focus on providing knowledge-based services that can be used for representing and extracting spatio-temporal knowledge. These tools can be used not only for extraction of quantitative knowledge from image sets but also for automated metadata generation. We propose CIML for this purpose. This work also draws on related work in computer vision and knowledge representation communities. A language-based representation of video events and a markup language for video annotation based on XML have been proposed in [12]. The authors also discuss domain ontologies for physical security and meeting domains. A visual concept ontology-based approach for semantic image interpretation is proposed in [13]. The idea of embedding semantic information into multimedia documents has been proposed in [14]. Open Microscopy Environment (OME) has been proposed as a solution for the management of multidimensional biological image sets [2]. It uses XML schemas for saving image data along with experiment context and analysis results. OME focuses not on providing image analysis or knowledge extraction tools but on providing a unified data format for storage and sharing of multi-dimensional microscopic images. The challenging problem of knowledge-based semantic interpretation of biological images requires incorporation of domain-specific knowledge into intelligent image understanding tools. This requires modeling of images and mechanisms for representing spatio-temporal biological knowledge. At the same time, the high volume of image sets generated by HCS technologies requires powerful computational resources and efficient data storage and management systems. High-throughput biological imaging can greatly benefit from developments in grid, computer vision and knowledge representation communities. Based on these observations, we propose an architecture that builds upon the basic grid services and provides knowledge extraction and management services for HCS. We also propose an XML-based language for event representation and event markup with a specific focus on biological imaging. This approach enables the extraction of quantitative knowledge from large biological image sets for hypothesis-driven image analysis. This extracted knowledge can also be embedded

into an image format like OME. This approach can be useful in solving the interoperability issues of biological imaging data produced by heterogeneous sources. This makes it possible to use advanced data mining techniques for rule induction.

The rest of the paper is organized as follows. Section 2 describes how grid technologies can be used to provide knowledge-based services for HCS. Section 3 discusses the issues of spatio-temporal event recognition and representation. The architecture of our grid-enabled system for HCS along with our spatio-temporal model and CIML are presented in section 4. An illustrative example is described in section 5. Section 6 describes the implementation status and presents results of an apoptosis screen, and the paper is concluded in section 7.

2. Knowledge-based grid for HCS

Grid technologies have been proposed for applications that require large computational resources or deal with high volumes of data. Modeling and simulation of complex systems and research in particle physics are some of the examples [20]. Grids provide middleware and tools for harnessing diverse and distributed computational and storage resources in a seamless fashion and for providing the end user with a uniform front end [33]. Grid technologies were initially motivated by computationally intensive applications that required tremendous computing power and computational grids provided the required resources [33]. Development of data grids was motivated by the need for systems to analyze and process data stored at geographically diverse data storage facilities using the computational power provided by the computational grids [32]. Great progress has been made in recent years in providing powerful middleware and services for computational and data grids [33]. Recently there has been tremendous interest in knowledge grids as enabling technologies for better exploiting the underlying computational and data services provided by computational and data grids [21-25]. Knowledge grid technologies aim at

providing knowledge services that are tailored for particular domains and thus facilitate efficient utilization of available resources. Figure 2 shows the hierarchy of grid services. In this section we investigate how grid infrastructure can help high-throughput biological imaging, and present a knowledge grid architecture for this rapidly expanding field.

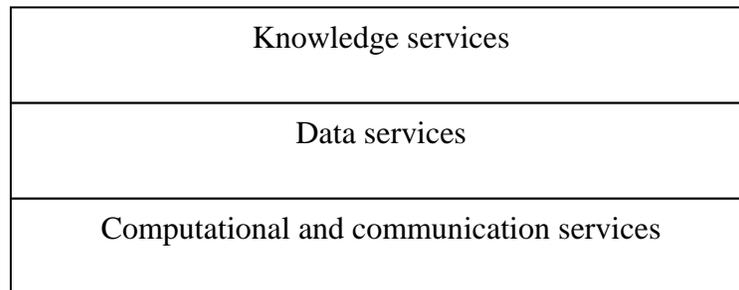


Figure 2. Hierarchy of grid services

Biological imaging investigations are normally aimed at validating scientific hypotheses. The first step therefore is to define the hypothesis followed by an experiment design. The experiment is then conducted and results are analyzed to understand the biological processes. Owing to the complexity of biological organisms, it is evident that future research in biological sciences and drug development will require more and more collaboration among research groups. This collaboration effort will be facilitated by automated knowledge extraction and metadata generation tools. Grid technologies, such as described in Figure 3, provide a promising solution to meet the computational, data processing, and collaboration needs of such tools to facilitate the HCS pipeline. Hypothesis development requires intensive searches for related knowledge available in diverse locations. As the data in this case happen to be in the form of large collections of images, a key requirement is to have automated knowledge extraction and metadata generation tools and formalisms for representation of this extracted knowledge so that software agents can search for relevant knowledge. The design of imaging experiments would also benefit from collaboration services. For example, if a few researchers are studying the proteome of the

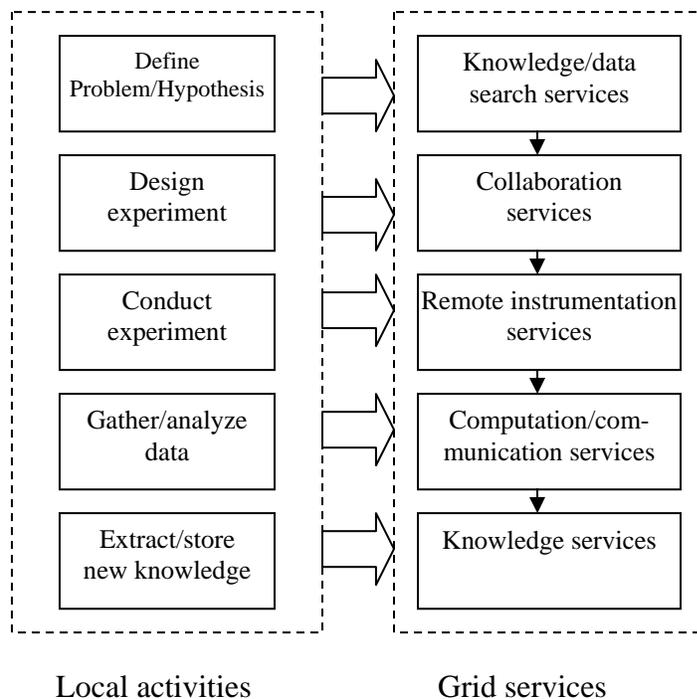


Figure 3. HCS workflow utilizing grid services

same organism, they may want to discuss the design of experiments and may even want to divide the whole proteome among themselves so as to avoid repetition of the same work. The human genome project is an example of a recent project of this nature [35]. Grid services can also make remote instruments available to scientists for conducting experiments. Remote microscopy has been used to provide access to expensive instruments like electron microscopes or for obtaining a second opinion from a pathologist [28]. A large drug screen may require expensive instruments like imaging cytometers at multiple geographic sites, requiring a high degree of coordination. Efficient grid middleware can provide these services. Once the experiments are conducted, the results are gathered and analyzed. In the case of HCS, transfer of large volumes of imaging data may require high speed communication infrastructure. These data are to be transferred to appropriate locations where they are analyzed using high performance computational resources. Two levels of analysis can be identified here. The first level is concerned with identifying the

basic information about objects and events depicted in image sets. This level essentially extracts “numbers” out of images. The second level deals with data mining and reasoning based on the basic information extracted by the first-level processing. Both these types of analyses require high performance computational resources. Finally, the high-level knowledge thus extracted should be represented in a machine-readable format so that it is available for new studies.

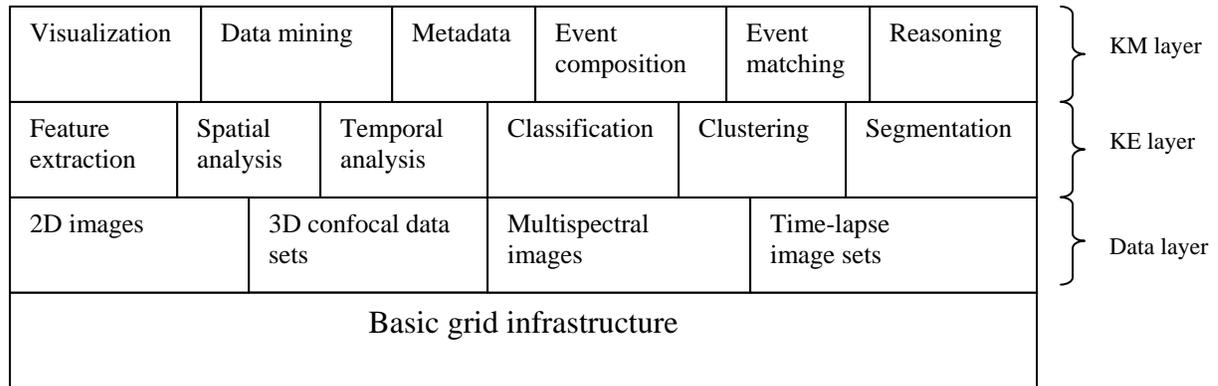


Figure 4. Services provided by the proposed architecture

Based on the discussion in previous paragraphs we envision two levels of knowledge services for HCS as shown in Figure 4. The KE layer provides algorithms for extraction of spatio-temporal knowledge from images, while the KM layer deals with higher-level knowledge representation and reasoning. Imaging data may consist of two-dimensional microscope images, three-dimensional confocal datasets, time-lapse images, or multispectral images [2]. The KE layer provides algorithms for identifying objects and events. Object identification by classification algorithms is based on object features like size, area, shape and spectrum to cite a few examples. This layer also provides algorithms for preprocessing of images like noise removal, registration and segmentation. Algorithms are also provided for identifying spatial and temporal events as explained in section 3. The KM layer provides services for specifying objects and events so that they can be identified by the algorithms provided by the KE layer. This layer also provides data

mining and reasoning services for discovering new information from the spatio-temporal information about the images. Other services provided by this layer include automatic metadata generation and visualization. These knowledge services are based on the basic computational and data grid services. The lower layers manage issues like resource discovery, scheduling, data movement, security, and collaboration. The focus of this paper is on spatio-temporal knowledge extraction and representation services.

3. Spatio-temporal event recognition and representation

Most of the events of interest in biological imaging arise either because of changes in the attributes of individual objects or because of interactions among multiple objects. For example, a cell undergoing division changes size and shape, and then divides into two daughter cells. In order to recognize particular events of interest, different objects in the scene need to be identified and then monitored over a period of time. Many approaches have been proposed in the literature for recognizing events in videos [17,18]. These can be adapted to the requirements of event detection in biological images as well. In our system we define predicates for spatial analysis based on bounding box, convex hull, or exact outline, depending on the speed and accuracy requirements of the application. Temporal event identification is based on a state machine-based approach where each state is defined by the objects participating in that sub-event, the specific values of their attributes, and the spatial relations between them. One of the objectives is to decouple event recognition from event representation. This way, newer events can be defined in a flexible manner and this representation can be used by event-recognition logic to identify events of interest independent of the lower-level algorithms used.

Objects of interest can be identified using different parameters, for example size, shape, color, texture, or spectral information in the case of multispectral imaging [6, 7]. Spatial location of

different objects relative to each other can be analyzed using bounding boxes around objects or other suitable spatial representation. The bounding box is a simple and efficient approach for capturing spatial information about objects in an image. Even though it is imprecise compared to the convex hull or the articulate boundary-based analysis, its speed and simplicity make it an attractive choice for high throughput biological imaging analysis where a large number of cells are to be analyzed. Inter-object spatial analysis can then be performed using projections on coordinate axes. An extension of temporal logic [8] to two and three dimensions is then used to describe the inter-object relationships as described in [9]. Figure 5 shows graphically some of the predicates used in our system. More predicates and functions can be defined, but only those used in later sections to explain our event representation scheme are introduced here. Qualifiers can also be specified for these predicates for different situations, for example disjoint along with a distance between the centroids of objects, or overlap with the area of the overlapping region.

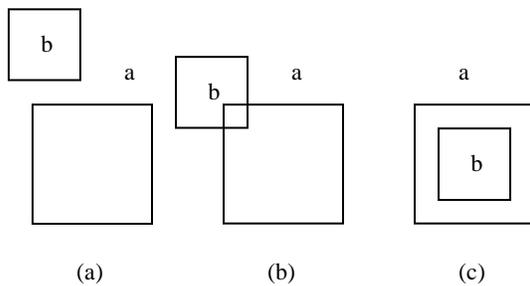


Figure 5. Graphical representation of three different spatial relations (a) Disjoint, (b) Overlap, (c) Contain

Many biological imaging experiments involve studying the kinetics of the diffusion of different proteins and chemicals among multiple sub-cellular compartments. Monitoring the variations of spatial relations between different objects over time is highly beneficial for such experiments. For example if a drug is designed to enter a certain compartment of the cell, it is very important to monitor the dynamics of this process. This will involve using markers that reveal the position of the drug and the location of the targeted compartment of the cell. If the drug penetrates from outside the cell into the particular region, the two markers will be initially at disjoint locations

and then would overlap as the time passes. Monitoring the evolution of objects and their spatial relations over time thus can be very helpful in understanding biological processes.

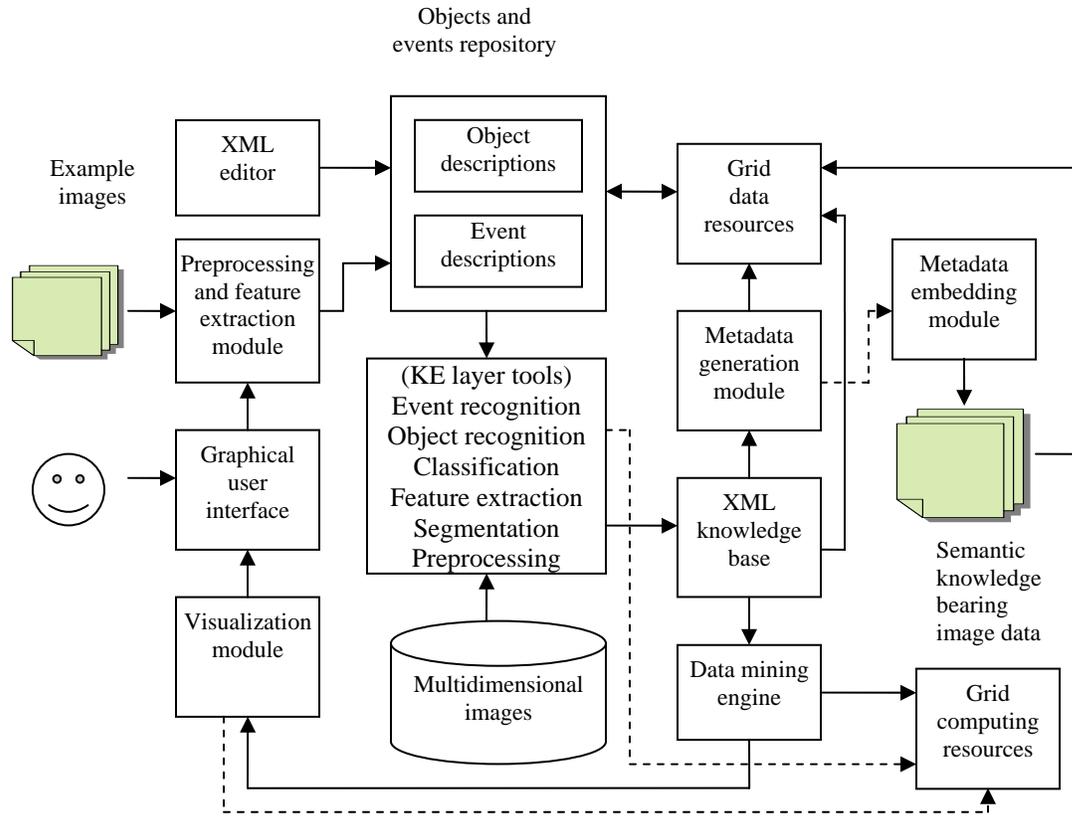


Figure 6. Architecture of the grid-enabled component-based system for HCS

4. Architecture of the XML-based grid for HCS

XML is a general-purpose markup language that provides a means for defining special-purpose markup languages [10]. It has been extensively used as a platform-independent mechanism for describing data in many different domains [11, 15, 16]. Its flexibility in defining different types of data makes it suitable for description of biological objects and events. The general architecture of our system is shown in Figure 6. This system uses a modular approach of describing objects in terms of their attributes and then spatial and temporal events based on the changes in attributes of the objects that constitute those events. Spatial events are described as the constituent objects

along with the relationships among their spatial representations (bounding boxes, convex hull, or exact outlines) depending upon the speed and accuracy requirements of the application. Temporal events are generally composed of a sequence of one or more spatial events occurring over a period of time. Some temporal events, such as, a monotonic increase in the size of a cell, can not be easily broken down into constituent spatial events and require special functions. Different components of the system are described next.

The XML editor is used for describing the attributes of objects and events manually. In fact, any text or XML editor can be used for this purpose. Alternatively, the feature extractor module can be used for specifying objects and events. The user provides example images of the objects and events and the feature extractor automatically extracts the attributes and generates XML representations that are stored in the objects and events repository. These representations are used by object and event recognition tools for identifying their specific instances in image sets. The information about the objects and events thus identified is kept in the XML knowledge base. This knowledge base represents the results of first-level processing as explained in section 2. Analysis results with different levels of detail can then be embedded into the image data using the XML embedding module.

4.1 Spatio-temporal Data model

Figure 7 shows the entity-relation diagram for the data model. We model biological images as a composition of biological objects that are specified by their features like size, shape, spectrum etc. The choice of the specific features to use for defining objects depends on the particular application and the choice of imaging modality. For many cases simple features like size and color may be adequate, whereas for others, more complicated features like Haralick texture and shape descriptors may be required [29]. Representation of objects in terms of their features allows

automated software tools to identify objects meeting those feature criteria in a large set of images. It also makes it possible to use the same analysis parameters for other image sets for comparison.

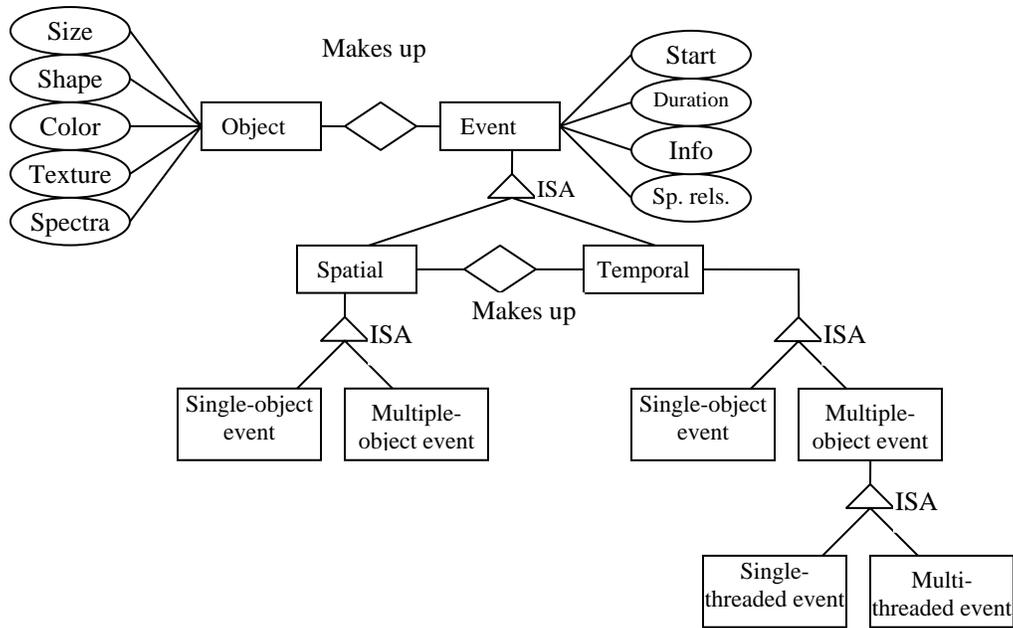


Figure 7. Spatio-temporal data model

Events are modeled as a composition of objects. Events are defined in terms of their attributes like a start time, duration, spatial relations between participating objects, and other conceptual information. As shown in Figure 6, spatial and temporal events are defined as the two types of events. Spatial events are defined solely by the participating objects and their spatial relations. For example two touching nuclei constitute a spatial event that is completely defined by the features of the objects and their spatial relation. The simplest spatial event is the single-object event that is defined by the participating object along with specific values of its features. Multiple-object spatial events are defined in terms of the participating objects and their spatial relations.

A spatial event persisting over a number of frames gives rise to a simple temporal event. Temporal events may also be composed of single or multiple events. Single-object temporal events involve evolution of an object over time, for example, expansion of a cell after undergoing

a certain treatment. Multiple-object events involve the evolution of the attributes and spatial relations of the participating objects, for example, the movement of a particular protein from nucleus to cytoplasm as the cell is undergoing expansion in response to a certain treatment. Temporal events may also be single-threaded or multi-threaded [12]. A single-threaded event is one in which all constituent events happen in a linear order, whereas in a multi-threaded event one or more constituent events happen in parallel.

4.2 Representation of objects

Biological cells, sub-cellular compartments, or other entities of interest constitute the objects. Objects are defined in terms of their parameters which may include color, size, shape, texture, spectrum, or other parameters depending on the application. Objects can be defined either manually using the XML editor or by providing example images of the objects to the feature extractor that extracts the relevant parameters of objects. These are then stored as XML schemas. An example schema is shown in Figure 8.

```
<object>
  <id>object1</id>
  <type>HeLa cancer cell</type>
  <compartments>
    <cytoplasm>
      <marker>FITC</marker>
      <size>
        <max>50</max>
        <min>10</min>
      </size>
      <shape_parameters> : </shape_parameters>
      <spectral_parameters> : </spectral_parameters>
    </cytoplasm>
    <nucleus>
      <marker>Hoechst</marker>
      <size>
        <max>50</max>
        <min>10</min>
      </size>
      <texture_parameters> : </texture_parameters>
    </nucleus>
  </compartments>
</object>
```

Figure 8. Representation of a cancer cell

4.3 Representation of spatial and temporal events

Spatial and temporal events can also be specified either manually or by using the feature extractor. Spatial events may involve one or more objects and may be characterized by certain values of an object's features or by one object's features, for example location, in relation to another. As an example of a spatial event one may be interested in finding all the cells that became compressed because of a certain treatment. This condition may be identified by the size or shape of the objects. Alternatively, one may wish to identify all cells that contain a specific number of sub-cellular objects. This task will require analyzing the location of different objects relative to each other. Temporal events arise because of changes in different parameters of objects over time. For example diffusion of a protein into a cell from the outside environment is a temporal event and in this situation the diffusion rate and the particular location to which the protein binds may be of interest. In order to perform such analysis tasks, sub-cellular objects need to be identified along with their positions relative to each other. This information then needs to be monitored over time to extract temporal information. XML specification of cell division as a temporal event is shown in Figure 9. Salient phases in a cell division would involve expansion of the cell followed by nuclear division and then the splitting of the cell into two daughter cells.

```
<TemporalEvent>
  <name> cell division </name>
  <concept> mitosis</concept>
  <start_frame> 10</start_frame>
  <duration>20</duration>
  <SpatialEvents>
    <SpatialEvent>
      <name> normal </name>
      <object_list> h1 </object_list>
      <start_frame> 10 </start_frame>
      <duration> 10 </duration>
    </SpatialEvent>
    <SpatialEvent>
```

```

    <name> elongation </name>
    <object_list> h1 </object_list>
    <start_frame> 20 </start_frame>
    <duration> 5 </duration>
  </SpatialEvent>
  <SpatialEvent>
    <name> nuclear_division </name>
    <object_list> h1 </object_list>
    <start_frame> 25 </start_frame>
    <duration> 1 </duration>
  </SpatialEvent>
  <SpatialEvent>
    <name> cell_splitting </name>
    <object_list> h1 h2 h3 </object_list>
    <start_frame> 26 </start_frame>
    <duration> 4 </duration>
  </SpatialEvent>
</SpatialEvents>
</TemporalEvent>

```

Figure 9. Representation of a cell division event

```

<SpatialEvent>
  <name>Expansion</name>
  <objects>
    <object>
      <id>h1</id>
      <type>Hela cancer cell</type>
      <compartments>
        <nucleus>
          <size>
            <min>size_exp_nuc_min</min>
            <max>size_exp_nuc_max</max>
          </size>
        </nucleus>
        <cytoplasm>
          <size>
            <min>size_exp_cyto_min</min>
            <max>size_exp_cyto_max</max>
          </size>
        </cytoplasm>
      </compartments>
    </object>
  </objects>
</SpatialEvent>

```

Figure 10. Representation of the expansion phase in a cell division event

```

<SpatialEvent>
  <name>Nuclear division</name>
  <objects>
    <object>
      <id>h1</id>
      <type>Hela cancer cell</type>
      <compartments>
        <nucleus>
          <size>
            <min>size_normal_nuc_min</min>
            <max>size_normal_nuc_max</max>
          </size>
        </nucleus>
        <nucleus>
          <size>
            <min>size_normal_nuc_min</min>
            <max>size_normal_nuc_max</max>
          </size>
        </nucleus>
        <cytoplasm>
          <size>
            <min>size_exp_cyto_min</min>
            <max>size_exp_cyto_max</max>
          </size>
        </cytoplasm>
      </compartments>
    </object>
  </objects>
</SpatialEvent>

```

Figure 11. Representation of the nuclear division phase in a cell division event

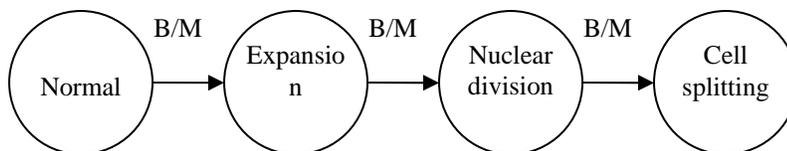


Figure 12. Graphical representation of sub events for cell division

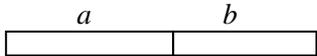
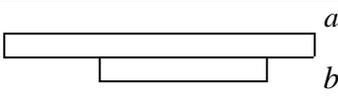
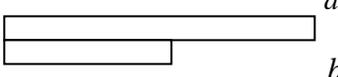
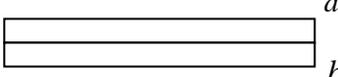
| Relation | Symbol | Graphical representation |
|----------------------|---------------|-------------------------------------------------------------------------------------|
| <i>a before b</i> | <i>B</i> |  |
| <i>a meets b</i> | <i>M</i> |  |
| <i>a overlaps b</i> | <i>O</i> |  |
| <i>a contains b</i> | <i>C</i> |  |
| <i>a starts b</i> | <i>S</i> |  |
| <i>a equals b</i> | <i>E</i> |  |
| <i>a completes b</i> | <i>CO</i> |  |

Figure 13. Temporal relations

4.4 Representation of analysis results

HCS experiments produce huge data sets. Hypothesis-driven biological imaging experiments require large cell populations in order for the results to have statistical significance. It is important to have a mechanism not only for automated knowledge extraction but also for sharing analysis results of cell screening experiments along with the imaging data. Most such experiments aim at identifying sample sub-populations that meet certain specific criteria. Different levels of detail may be required for the extracted knowledge. For example, in some cases it may be sufficient to identify only the different sub-populations. In others, locations of different objects may also be needed. These analysis results can then be embedded into image data using a data format like the one introduced in OME [2]. This approach facilitates searching such imaging data and also

reduces computational time if reanalysis is required. Also it makes it easy to share data and results among different research groups

Automatic identification of spatio-temporal events in biological imaging experiments opens new avenues for event mining and rule induction. A large population of cells can be monitored over time under different perturbations and the effects can be automatically analyzed. This approach can be very helpful in drug discovery for the identification of promising targets for potential drugs.

5. An example of event representation

In this section we explain the concepts introduced in earlier sections by means of a detailed example of an apoptosis screening. The knowledge representation schemas used for this example are also used for an actual apoptosis screening which is explained in section 6. A discussion of the process of apoptosis appears in section 1.2. The nuclei of cells are identified using a Hoechst marker. PI is used for identifying dead cells as it can penetrate the ruptured cell membranes of dead cells but can not penetrate normal live cells. Annexin V FITC is used as an apoptotic marker because it binds specifically to the membrane of apoptotic cells. The objective of such a study is normally to test the efficacy of drugs in inducing apoptosis.

```
<EventGroup>
  <name>apoptosis</name>
  <objects>
    <object> Hoechst </object>
    <object> PI </object>
    <object> Annexin V FITC </object>
  </objects>
  <Event>
    <name>lateapoptotic_necrotic </name>
    <objects>
      <object> Hoechst </object>
      <object> PI </object>
      <object> Annexin V FITC </object>
    </objects>
  </Event>
</EventGroup>
```

```

</objects>
<SpatialRelations>
  <Relation>
    <name>overlap</name>
    <arg>Hoechst</arg>
    <arg>PI</arg>
  </Relation>
  <Relation>
    <name>overlap</name>
    <arg>Hoechst</arg>
    <arg>Annexin V FITC</arg>
  </Relation>
</SpatialRelations>
</Event>
<Event>
  <name>live</name>
  <objects>
    <object> Hoechst </object>
  </objects>
  <SpatialRelations>
    <Relation>
      <name>nooverlap</name>
      <arg>Hoechst</arg>
      <arg>PI</arg>
    </Relation>
    <Relation>
      <name>nooverlap</name>
      <arg>Hoechst</arg>
      <arg>Annexin V FITC</arg>
    </Relation>
  </SpatialRelations>
</Event>
<Event>
  <name>dead</name>
  <objects>
    <object> Hoechst</object>
    <object> PI </object>
  </objects>
  <SpatialRelations>
    <Relation>
      <name>overlap</name>
      <arg>Hoechst</arg>
      <arg>PI</arg>
    </Relation>
    <Relation>
      <name>nooverlap</name>
      <arg>Hoechst</arg>
      <arg>Annexin V FITC</arg>
    </Relation>
  </SpatialRelations>

```

```

</Event>
<Event>
  <name>early apoptotic</name>
  <objects>
    <object> Hoechst</object>
    <object> Annexin</object>
  </objects>
  <SpatialRelations>
    <Relation>
      <name>overlap</name>
      <arg>Hoechst</arg>
      <arg>Annexin V FITC</arg>
    </Relation>
    <Relation>
      <name>nooverlap</name>
      <arg>Hoechst</arg>
      <arg>PI</arg>
    </Relation>
  </SpatialRelations>
</Event>
</EventGroup>

```

Figure 14. Representation of events involved in an apoptosis screen

Events of interests in this case are early apoptotic, late apoptotic, live, and dead cells. These conditions can be identified by observing the overlap between regions that contain different types of markers. Cells containing only Hoechst will be live, where as cells with both Hoechst and PI staining will be dead, cells with Hoechst and Annexin V FITC will be early apoptotic, and cells with all three markers will be late apoptotic and necrotic. XML representation for an apoptosis screen is given in Figure 14. This screen will provide the percentage of cells in live, dead, and apoptotic states. An example of a schema for the analysis results is shown in Figure 15.

```

<AnalysisResults>
  <Configuration>
    <dye>
      <name>Hoechst</name>
      <excitation>405</excitation>
      <emission>500</emission>
      <intensity>35</intensity>
    </dye>
  </Configuration>
</AnalysisResults>

```

```

    <object_size>100</object_size>
  </dye>
  <dye>
    <name>PI</name>
    <excitation>488</excitation>
    <emission>600</emission>
    <intensity>45</intensity>
    <object_size>100</object_size>
  </dye>
  <dye>
    <name>Annexin V FITC </name>
    <excitation>488</excitation>
    <emission>550</emission>
    <intensity>25</intensity>
    <object_size>100</object_size>
  </dye>
</Configuration>
<Population_distribution>
  <Population>
    <name>late_apoptotic</name>
    <percentage>70</percentage>
  </Population>
  <Population>
    <name>live</name>
    <percentage>20</percentage>
  </Population>
  <Population>
    <name>dead</name>
    <percentage>6</percentage>
  </Population>
  <Population>
    <name>early_apoptotic</name>
    <percentage>4</percentage>
  </Population>
</Population_distribution>
</AnalysisResults>

```

Figure 15. Representation of analysis results for the apoptosis screen

6. Prototype and discussion

We have developed a prototype of our system. The feature extraction, event specification and spatial event identification modules as well as XML parser are developed in Matlab. Objects can be either specified manually in terms of their attributes or some representative images can be provided to the feature extractor, which extracts object features and then stores them as XML object representation schemas. A screen shot of the graphical interface is shown in Figure 16. As

HCS screens typically generate a large number of images, the object representation schemas can be used to automatically identify objects within all images. The spatial analysis module provides algorithms based on bounding box, convex hull, and exact outline. Spatial events of interest can be specified manually or using the graphical interface shown in Figure 16. The event specification module then generates XML schemas corresponding to the spatial events of interest. This spatial event knowledge is then used by spatial analysis module to find all such events in the set of collected images.

An apoptosis screening, as described in the previous sections, was performed using the event specification and matching modules described above. Human myelogenous leukemia (HL60) cells were exposed to one of two apoptosis-inducing drugs, rotenone or camptothecin, for twenty four hours. A negative control consisting of cells without any treatment was also prepared. Subsequently, staining was performed. Three dyes were used: Hoechst 33342, PI and Annexin V FITC as described earlier. Four hundred sets of images were collected for each of the samples using an imaging cytometer (iCys research imaging cytometer, CompuCyte Corporation, Cambridge, MA). Each set consisted of three images, one for each of the dyes used. Figure 17 shows a representative set of images. As explained in section 5, events of interests in this case are early apoptotic, late apoptotic, live, and dead cells. These are defined in terms of spatial relationships between the three dyes used as specified in the schema shown in Figure 14. Spatial analysis module used this knowledge and identified the populations of live, dead, early and late apoptotic cells.

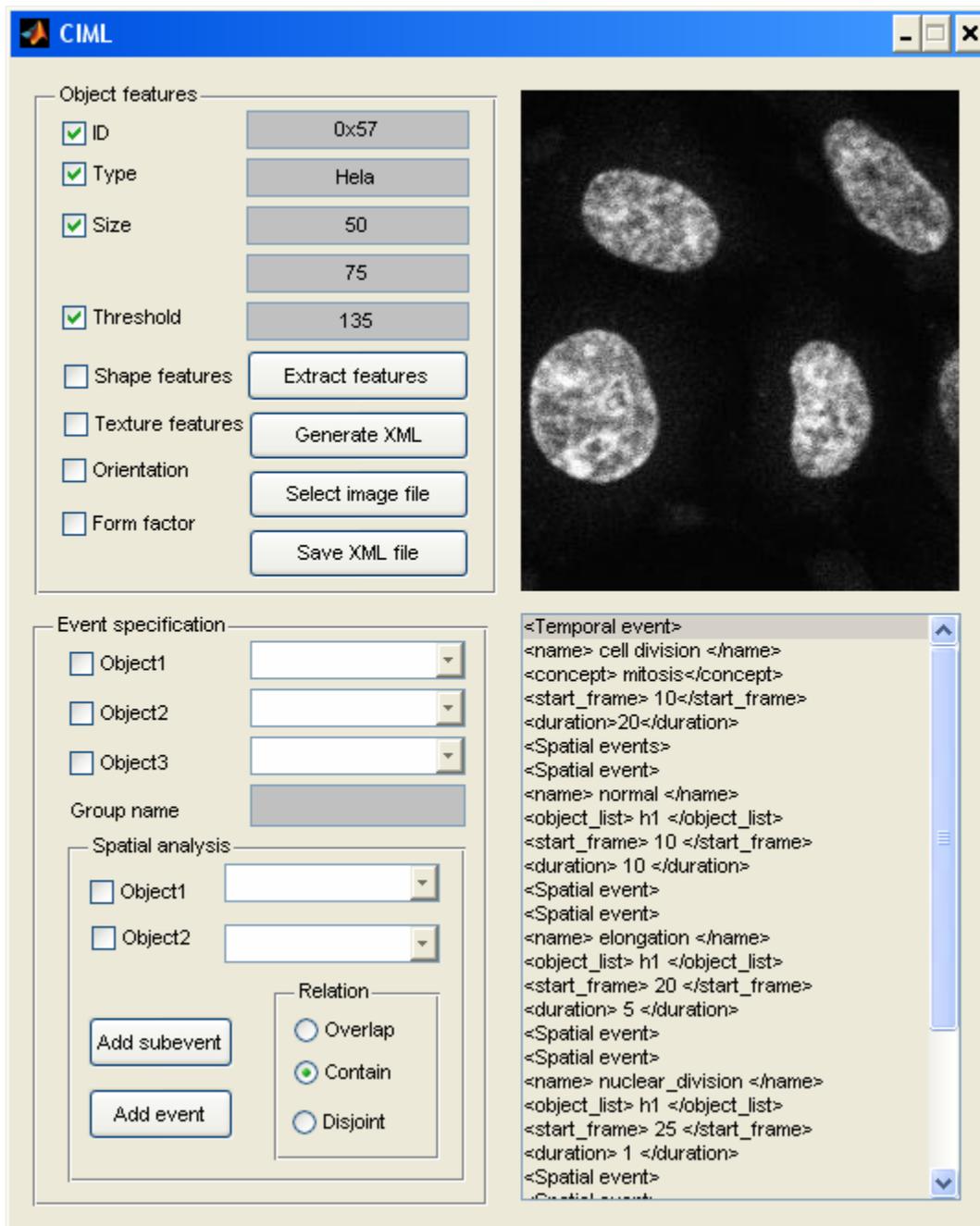


Figure 16. Graphical interface for objects and events specification

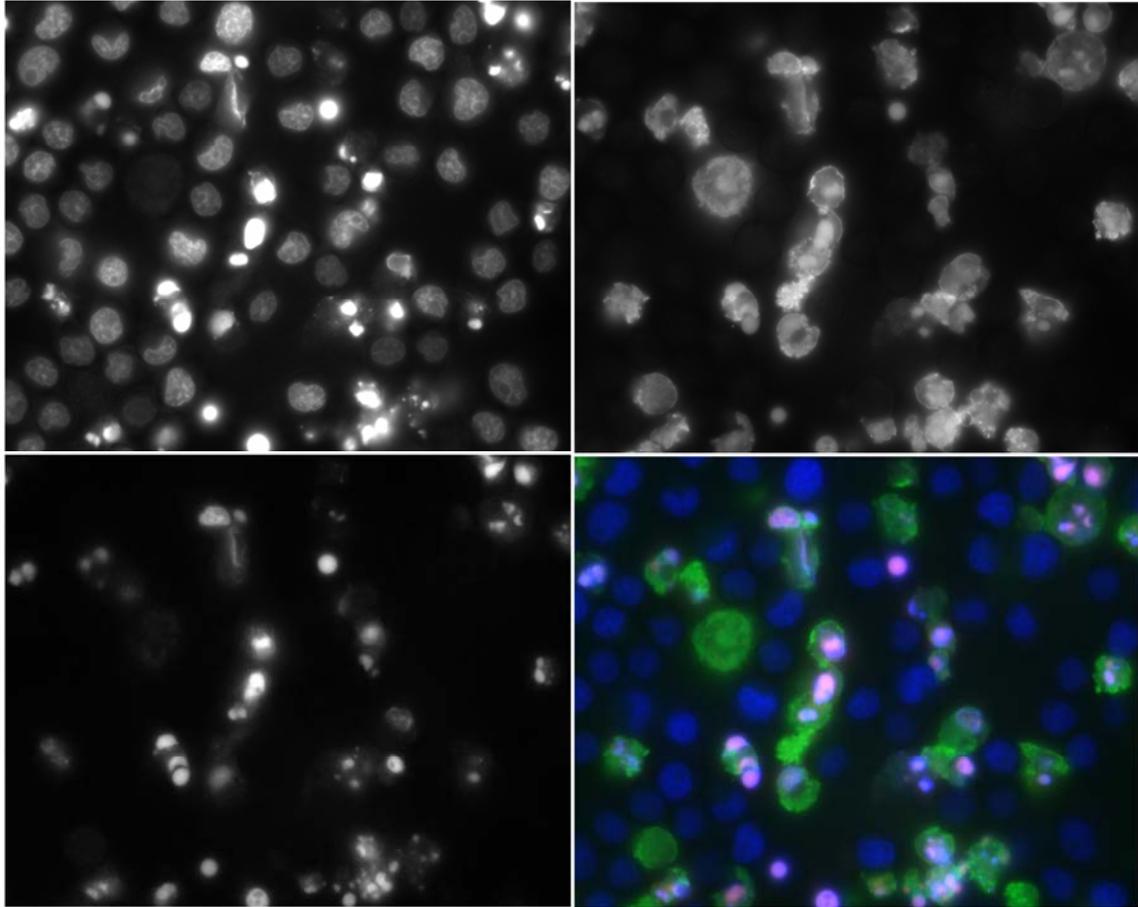


Figure 17. A representative set of images for the apoptosis screen showing HL60 cells in different states. (Upper left) Hoechst33342, (Upper right) Annexin V FITC, (Lower left) PI, (Lower right) Merged image. Images were acquired using an E1000 fluorescence microscope (Nikon, Tokyo, Japan) using a 60x 1.4 NA lens. Fluorescence images of the PI, Annexin V FITC, and Hoechst 33342 were sequentially acquired with a RETIGA EXi Cooled monochrome 12-bit camera (QImaging, Burnaby, BC Canada). Image acquisition and processing was performed using Image-Pro Plus (MediaCybernetics, Silver Spring, MD). The three images were pseudo-colored and merged to make the composite image. Blue, green and red colors correspond to Hoechst33342, Annexin V FITC and PI respectively.

We plan to compare the performance of different spatial analysis algorithms in terms of their speed and accuracy. We also plan to compare the results of spatio-temporal analysis with other cell analysis modalities like flow cytometry [19]. The goal of this paper is to demonstrate the expressive power of our spatio-temporal modeling-based language for representation of HCS events and the effectiveness of the proposed architecture for providing grid-enabled knowledge-

based services for HCS. HCS applications require powerful computing and distributed data management resources. At the same time these systems have to be affordable and economical. Grid technologies provide a promising solution. The realization of this solution requires powerful knowledge-based tools that combine domain knowledge with analysis algorithms to exploit the full potential of the underlying grid services. We believe that the grid-enabled HCS system and the spatio-temporal knowledge representation scheme proposed in this paper provide a powerful yet economic solution to the challenging requirements of HCS applications.

7. Conclusion

Automatic extraction and management of spatio-temporal semantic knowledge from large image sets produced by HCS screens requires high throughput computing resources, efficient management of data and intelligent knowledge-based tools. Grid technologies provide a basic infrastructure on which domain-specific services can be built. In this paper, we propose a grid-based architecture for providing knowledge-based services for HCS. Two main layers of this architecture are the KE layer and the KM layer. The KE layer provides tools for extracting information about biological objects and spatio-temporal events whereas the KM layer provides high-level knowledge representation and reasoning services. We also introduce an XML-based language for modeling biological images and for representing spatio-temporal knowledge. This mechanism provides a knowledge representation scheme that is used by event-recognition logic to identify events of interest. This approach combines domain knowledge with analysis tools to fully utilize the powerful computing and data management services provided by grid infrastructure. A prototype for spatial analysis has been developed. Implementation of temporal event-recognition algorithms for time-lapse imaging is currently underway. In the future we plan to integrate the spatio-temporal analysis-based knowledge services with the basic grid middleware like Globus [38]. We would also like to explore data mining techniques for automatic rule induction from spatio-temporal dynamics of biological systems.

References

- [1] D. J. Stephens, V. J. Allan, "Light microscopy techniques for live cell imaging", *Science*, 4 April 2003, pp. 82-86.
- [2] J. R. Swedlow, I. Goldberg, E. Brauner, P. K. Sorger, "Informatics and quantitative analysis in biological imaging", *Science*, 4 April 2003, pp. 100-102.
- [3] J. Paul Robinson, "Principles of confocal microscopy", *Methods Cell Biology*; 63:89-106.
- [4] X. Zhou, S. T. C. Wong, "High content cellular imaging for drug development", *IEEE Signal Processing Magazine*, March 2006, pp. 170-174.
- [5] J. D. Robertson, S. Orrenius, B. Zhivotovsky, "Review: Nuclear Event in Apoptosis", *Journal of Structural Biology*, Volume 129, Issues 2-3, April 2000, pp. 346-358.
- [6] M. V. Boland, 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*, Vol. 17, No. 12, 2001, pp. 1213-1223.
- [7] M. E. Dickinson, G. Bearman, S. Tille, R. Lansford, S. E Fraser, "Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy", *BioTechniques*, Vol. 31, No. 6, 2001, pp. 1272, 1274-1276, 1278.
- [8] J. F. Allen, "Maintaining knowledge about temporal intervals", *Communications of the ACM*, November 1983, pp. 832-843.
- [9] Y. F. Day, S. Dagtas, M. Iino, A. Khokhar, a. Ghafoor, "Object-oriented conceptual modeling of video data", 11th International Conference on Data Engineering(ICDE'95), 1995, pp. 401-408.
- [10] M. Klein, "XML, RDF, and relatives", *IEEE Intelligent Systems*, March/April 2001, pp. 26-28.
- [11] S. Decker et al., "The semantic web: the roles of XML and RDF", *IEEE Internet Computing*, September/October 2000, pp. 63-74.

- [12] R. Nevatia, J. Hobbs, B. Bolles, "An ontology for video event representation", IEEE Workshop on Event Detection and Recognition, June 2004.
- [13] N. Maillot, M. Thonnat, A. Boucher, "Towards ontology-based cognitive vision", Machine Vision and Applications, 2004, pp. 33-40.
- [14] K. Thirunarayan, "On embedding machine-processable semantics into documents", IEEE Transactions on Knowledge and Data Engineering, July 2005, pp. 1014-1018.
- [15] A. Yao, J. Jin, "The development of a video metadata authoring and browsing system in XML", ACM International Conference Proceeding Series, 2000, pp. 39-46.
- [16] S. Wrede, J. Fritsch, C. Bauckhage, G. Sagerer, "An XML based framework for cognitive vision architectures", 17th International Conference on Pattern Recognition, 2004, pp. 757-760.
- [17] D. Ayers, M. Shah, "Monitoring human behavior from video taken in an office environment", Image and Vision Computing, Vol. 19, No. 12, 2001, pp. 833-846.
- [18] A. F. Bobick, Y. A. Ivanov, "Action recognition using probabilistic parsing", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998, pp. 196-202.
- [19] J. Paul Robinson, "Flow Cytometry", Encyclopedia of Biomaterials and Biomedical Engineering, Ed. GE Wnek, GL Bowlin. Marcel Dekker Co., 2004, pp. 630-640.
- [20] W. E. Johnston, "Computational and data grids in large-scale science and engineering", Future Generation Computer Systems 18, 2002, pp. 1085-1100.
- [21] S. Hastings et al., "Grid-based management of biomedical data using an XML-based distributed data management system", 2005 ACM Symposium on Applied Computing, 2005, pp. 105-109.
- [22] D. D. Roure, N. R. Jennings, N. R. Shadbolt, "The semantic grid: past, present, and future", Proceedings of the IEEE, Vol. 93, No. 3, March 2005.
- [23] H. Zhuge, "China's e-science knowledge grid environment", IEEE Intelligent Systems, January/February 2004, pp. 13-17.

- [24] M. Cannataro, D. Talia, "Semantics and knowledge grids: building the next-generation grid", *IEEE Intelligent Systems*, January/February 2004, pp. 56-63.
- [25] S. Hastings et al., "Image processing for the grid: a toolkit for building grid-enabled image processing applications", *Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'03)*, 2003, pp. 36-43.
- [26] W. Huh et al., "Global analysis of protein localization in budding yeast", *Nature* 425, 16 October 2003, pp. 686-691.
- [27] K. Huang, R. F. Murphy, "From quantitative microscopy to automated image understanding", *Journal of Biomedical Optics*, 9(5), pp. 893-912.
- [28] K. Yoshida et al., "Design of a remote operation system for trans Pacific microscopy via international advanced networks", *Journal of Electron Microscopy*, Vol. 51, 2002, pp. S253-S257.
- [29] R. C. Gonzalez, R. E. Woods, S. L. Eddins, "Digital image processing using Matlab", Upper Saddle River, NJ, Pearson/Prentice Hall, 2004, Ch. 11, pp. 426-483.
- [30] K. A. Giuliano, "High-content screening: A new approach to easing key bottlenecks in the drug discovery process", *Journal of Biomolecular Screening*, Vol. 2, No. 4, 1997, pp. 249-259.
- [31] K. A. Giuliano et al., "Systems cell biology knowledge created from high content screening", *Assay and Drug Development Technologies*, Vol. 3, No. 5, 2005, pp. 501-514.
- [32] M. Cannataro, D. Talia, P. Trunfio, "Distributed data mining on the grid", *Future Generation Computer Systems* 18, 2002, pp. 1101-1112.
- [33] K. Krauter, R. Buyya, M. Maheswaran, "A taxonomy and survey of grid resource management systems for distributed computing", *Software-Practice and Experience*, 32, 2002, pp. 135-164.
- [34] F. S. Collins, M. Morgan, A. Patrinos, "The human genome project: lessons from large-scale biology", *Science*, Vol. 300, No. 5617, 11 April 2003, pp. 286-290.

- [35] J. C. Venter et al., "The sequence of the human genome", *Science*, Vol. 291, No. 5507, 16 February 2001, pp. 1304-1351.
- [36] M. Ellisman et al., "The emerging role of biogrids", *Communications of the ACM*, Vol. 47, No. 11, November 2004, pp. 52-57.
- [37] P. Mazzatorta et al., "OpenMolGRID: Molecular science and engineering in a grid context", *Proceedings of the 2004 International Conference on Parallel and distributed Processing Techniques and Applications*, 2004, pp. 775-779.
- [38] I. Foster, C. Kesselman, "Globus: a toolkit-based grid architecture", *The grid: blueprint for a new computing infrastructure*, I. Foster and C. Kesselman, eds., Morgan Kaufmann, 1999, pp. 259-278.