DISTRIBUTED ONLINE ANOMALY DETECTION IN HIGH-CONTENT SCREENING

Adam Goode[†]*, Rahul Sukthankar*[‡]*, Lily Mummert*[‡]*, Mei Chen*[‡]*, Jeffrey Saltzman*[•]*,*
David Ross[•]*, Stacey Szymanski*[•]*, Anil Tarachandani*[•]*, M. Satyanarayanan*[†]

[†]Carnegie Mellon University, [‡]Intel Research Pittsburgh, [•]Merck & Co., Inc.

ABSTRACT

This paper presents an automated, online approach to anomaly detection in high-content screening assays for pharmaceutical research. Online detection of anomalies is attractive because it offers the possibility of immediate corrective action, early termination, and redesign of assays that may require many hours or days to execute. The proposed approach employs assay-specific image processing within an assay-independent framework for distributed control, machine learning, and anomaly reporting. Specifically, we exploit coarse-grained parallelism to distribute image processing over several computing nodes while efficiently aggregating sufficient statistics across nodes. This architecture also allows us to easily handle geographically-distributed data sources. Our results from two applications, adipocyte quantitation and neurite growth estimation, confirm that this online approach to anomaly detection is feasible, efficient, and accurate.

Index Terms— Anomaly detection, biomedical image processing, distributed database searching, high-content microscopy, OpenDiamond® search platform

1. INTRODUCTION

The science of anomaly detection plays an increasingly important role in pharmaceutical research organizations, both as a research tool and as a process control tool. In research, experiments are designed to systematically explore a large space of parameters and to detect rare outcomes that merit deeper investigation. In process control, anomaly detection is used to explore and discover metrics and methods that lead to more formal quality-control measures.

High-content screening (HCS) refers to those biological assays that run with a high degree of automation, contain large numbers of parallel experiments (typically 10^4–10^6), and primarily generate image data for further analysis. For example, so-called silencing RNA (siRNA) experiments may simultaneously use up to 30,000 RNAs to investigate the knock-down of every known gene [1, 2]. Anomalies, in this instance, may be those genes that cause unusual or important phenotypes that are characteristic of a specific disease. Large chemical libraries may substitute for siRNA-induced changes in pathway fluxes in treated cells, leading to anomalies in cell morphology or more deliberate fluorescence readouts. The same readouts used for finding differences in cell functions may also hint about the quality of the experiments themselves. For example, a loss of reagent potency may lead to patterns in the cell expression that are anomalous in a different and systematic manner.

This research was partly supported by the National Science Foundation (NSF) under grant number CNS-0614679. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, Carnegie Mellon University, Intel, or Merck. OpenDiamond is a registered trademark of Carnegie Mellon University.

In this paper, we describe an automated, online approach to anomaly detection in HCS assays. The term "online" refers to an approach that naturally lends itself to data processing and anomaly reporting in a continuous manner, while an HCS assay is still in progress. This is in contrast to approaches that defer anomaly detection until the completion of the assay. Online anomaly detection is attractive because it offers the possibility of early corrective action or early termination and redesign of assays that may run continuously for many hours or days. Our approach uses assay-specific image processing within an assay-independent framework for distributed control, machine learning and anomaly reporting.

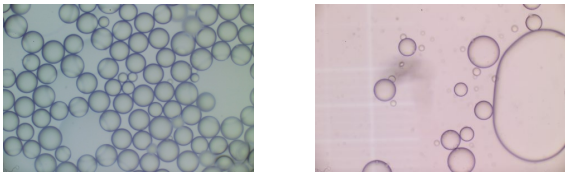
2. A FRAMEWORK FOR ANOMALY DETECTION

Anomaly screening at interactive speeds requires the adaptive learning algorithm to be embedded in an efficient infrastructure for compute-intensive image processing. This infrastructure is the OpenDiamond® platform for distributed search [3,4], which cleanly separates assay-specific and assay-independent aspects of image processing. The assay-specific image processing for an application defines a set of descriptors that determines the types of anomalies that are detectable. An anomaly is statistical outlier with respect to the descriptor set.

For each descriptor, OpenDiamond maintains a compact set of statistics, namely, the mean and standard deviation accumulated in the form of the count, sum, and sum of squares. A compact data representation is needed for performance: the size of the descriptor data must be constant with respect to the number of images processed. For online anomaly detection, an initial estimate of each descriptor is created by processing a number of images determined by a configurable *priming count*. These initial images are not subject to anomaly detection, but may be revisited and reprocessed later in the session. Descriptor statistics accumulated during a session can be saved and reused for future examinations of the data set.

Our framework is well suited for parallel computation since OpenDiamond supports distribution of data and processing over many servers. Each server processes a subset of the data independently of other servers. The client coordinates the sharing of descriptor statistics across servers by periodically collecting, aggregating, and distributing these statistics. Since the time to perform sharing is typically less than the time to process a single image, there is little skew across servers and hence no loss of statistical accuracy. The sharing period is configurable, with a default of 5 seconds.

Image processing is performed on servers through code components called *searchlets*. The searchlet is logically part of an application, the remainder of which runs on a client for user interaction. Descriptor statistics are calculated by the searchlet as part of image processing. The searchlet examines the existing descriptor statistics to determine if an image is anomalous, and if so, it writes additional data called *attributes* that indicate the nature and



(a) Normal (b) Anomalous

Fig. 1. Example Adipocyte Images

extent of the anomaly. Only anomalous images are transmitted to the client. OpenDiamond accomodates a variety of implementation methods for image processing. For example, the searchlet code for adipocyte images, described in Section 3.1, is implemented in C++. In contrast, the searchlet code for neurite images, described in Section 3.2, is implemented as a collection of ImageJ macros [5].

3. APPLICATIONS

We have validated our framework by applying it to two different problems: adipocyte quantitation and neurite growth estimation.

3.1. Detecting Anomalies in Adipocyte Images

Adipocytes, or fat cells, serve as reservoirs of energy in humans and are tightly regulated both in size and number. Significant alteration in body mass involves changes in both adipocyte size and number. In the field of lipid research, techniques are needed to locate and quantitate adipocytes in large repositories of cell microscopy images.

3.1.1. Data Collection

This work is based on high-resolution images of unfixed live adipocytes in suspension. Example images are shown in Figure 1. A live adipocyte suspension was prepared using collegenase to separate the cells from adipose tissue. A small drop of the suspension was placed on a slide with a circular ridge of silicone grease. The cells typically floated to the top of the drop, where they could be viewed on a Nikon Diaphot microscope and photographed with a 14-megapixel Kodak DCS Pro14n digital camera.

3.1.2. Image Processing

Because adipocytes in suspension are typically circular, they are located in the images by searching for elliptical objects. Quantitation is semi-automated; an investigator defines a reference adipocyte that takes into account variations in cell size, shape, and focus. Adipocytes are located as follows:

1. An image pyramid is built by scaling down the high-resolution images. The pyramid is necessary because the ellipse extraction algorithm is overly sensitive to scale. The next two steps are applied independently to each level of the pyramid.
2. A Canny-style edge detector is applied that uses color contrast gradients rather than grayscale contrast. A standard grayscale Canny detector was evaluated, but the color contrast variant provided better results.
3. The resulting binary edge images are used as input to an ellipse extraction algorithm that can locate overlapping and partially occluded cells [6]. This algorithm works well with noisy and incomplete data and is much faster than a Hough transform.

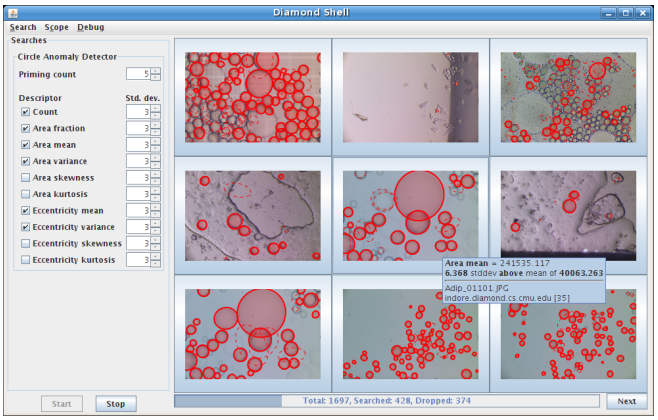


Fig. 2. Screenshot of Anomaly Detection Application

4. The results from all pyramid levels are merged. Ellipses found in scaled-down parts of the pyramid are scaled back up to match the original image size. A non-maxima suppression heuristic is employed to identify and eliminate idential ellipses that were detected in multiple pyramid levels.
5. Statistics such as the cell count and cell size distribution are tabulated.

Further details on adipocyte detection appear in Goode *et al.* [7].

Anomalies in the adipocyte images are detected based on the cell count, the fraction of the image covered by cells, and the first four statistical moments of cell size and shape (eccentricity). Figure 2 shows an application for detecting anomalies in adipocyte images based on the framework described in Section 2. The user selects descriptors on the left panel, configures the priming count, and starts the search. Anomalous images are shown as the search progresses. In the example shown, approximately 3% of the 1697 images searched were declared anomalous.

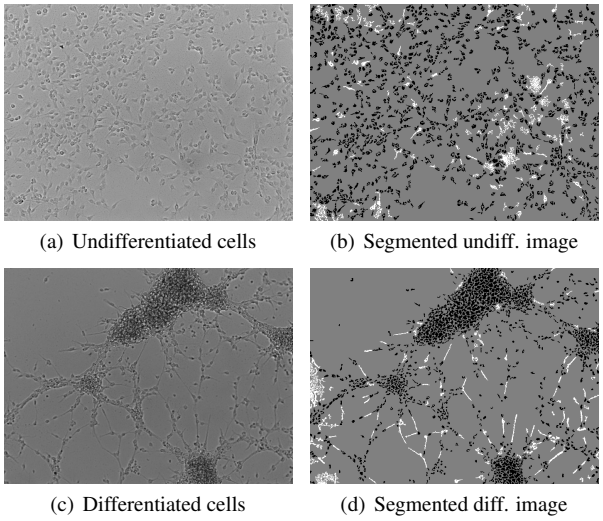
3.2. Detecting Anomalies in Neurite Images

Cells of the central nervous system, such as neurons and oligodendrocytes, have neurite processes that are involved in the synaptic function of nerves. Many human cognitive diseases cause or result in the degradation of neuronal cell health. *In vitro* imaging assays using cell culture models can utilize the status of neurites as a surrogate measure of cell health. The ability to measure neurite outgrowth enables identification of compounds and/or siRNAs that influence cell health or survival; increased neurite number and length correspond to increased cell health. In typical cell microscopy images (Figure 3), neurites appear as low-contrast linear features branching from high-contrast neuron bodies.

3.2.1. Data Collection

This work uses neuronal stem cells, which have the ability to differentiate into cells of the central nervous system (neurons, astrocytes and oligodendrocytes). As these undifferentiated neuronal stem cells undergo the differentiation process, they extend neurites. The length of the neurites is a measure of the differentiation state.

Neuronal stem cells were routinely cultured in the undifferentiated state using a defined growth media containing RHB-A media (Stem Cell Sciences; Cambridge, UK), supplemented with FGF2 and EGF (Peprotech; Rocky Hill, NJ). To image undifferentiated



(a) Undifferentiated cells (b) Segmented undiff. image

(c) Differentiated cells (d) Segmented diff. image

Fig. 3. Examples from the neurite growth domain. Two conditions from the same well are shown (temporally spaced): undifferentiated cells (a) and differentiated cells (c). The segmented versions of those images are shown in (b) and (d), respectively; detected neurites are evident as bright linear features, cell bodies as dark spots.

neuronal stem cells, cells were seeded on uncoated 384 well plates in growth media. After 24 hours, cells were fixed in a final solution of 4% paraformaldehyde. To image differentiated neuronal lineages, cells were seeded on Laminin coated 384 well plates in growth media for the first 24 hours. After 24 hours, media was changed to differentiation media which was similar to that of growth media, but only supplemented with low amounts of FGF2. Media was changed every 2 days for the entire differentiation period. Differentiation periods took place over 1–3 weeks, followed by fixation with 4% paraformaldehyde. Brightfield images were captured on an ImageXpress Micro (Molecular Devices, Sunnyvale, CA) using a 10x Nikon Plan Fluor DL objective.

Since neurites can exhibit significant 3D structure, it is possible that a single focal plane may fail to image all of the cells. Thus, images were obtained for 5 focal planes. The first image was captured at the focal plane that was optimal for the majority of cells. Additional images were captured on either side of this optimal plane, two on each side at equidistant intervals. The exposure settings for all of images were kept constant.

3.2.2. Image Processing

Anomalies in neurite images are evidenced by differences between expected and observed attributes, such as numbers, shapes, or density of neurites. Specifically, we characterize anomalies according to the following criteria: total number of neurites observed in the image; the aggregate lengths of these neurites; the number of cells (identified by cell bodies) detected in the image; the average size (area in pixels) occupied by such cells; the ratio of neurites to cell bodies; the total area of the image occupied by neurites; and ratio neurite area to neural cell body area.

The image processing required to extract these attributes is summarized as follows: first we find cell bodies, then we find neurites. Specifically:

1. The 5 focal plane images are merged into a single image by computing the median value at each pixel location.
2. Non-uniform background illumination is corrected by fitting a second-order polynomial to the image and subtracting.
3. A straightforward adaptive thresholding procedure is applied that exploits the fact that cell bodies correspond to high-intensity regions in the image.
4. Once the image has been thresholded, a watershed procedure teases apart clumped cell bodies.
5. Cell bodies within a specified size and eccentricity parametrs are extracted using a standard connected components operation.
6. Once the cell bodies have been identified, we measure their appropriate attributes (as discussed above).
7. Using the segmented cell bodies as a mask, the relatively low contrast between neurite pixels and the background becomes sufficiently distinct to enable segmentation.
8. A series of classical image processing steps (morphological filtering followed by connected components analysis) then produces a usable set of neurites. Neurites can still occasionally be oversegmented into multiple components, but our experiments indicate that this bias is not sufficiently severe as to impair the detection of anomalies.
9. We compute statistics from the extracted neurites.

An example of the first steps of multi-focal plane processing is illustrated in Figure 4.

4. EVALUATION

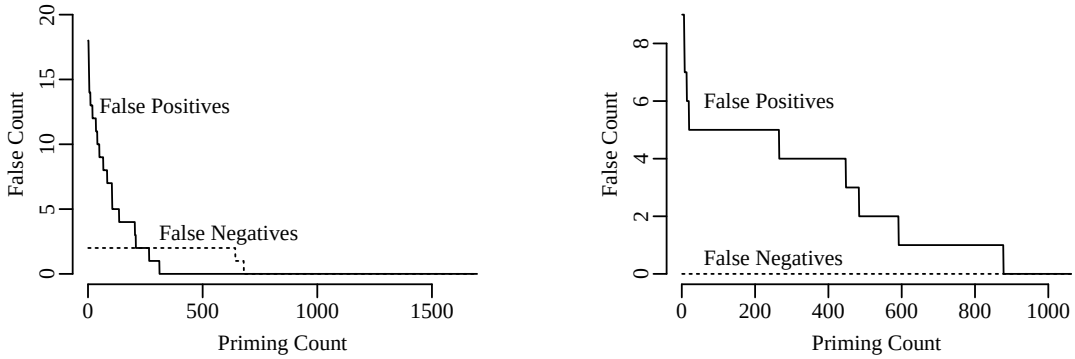
We compare our online distributed approach against traditional anomaly detection, both in terms of accuracy and speed, on each of our two application domains.

Our first set of experiments (Figure 5) explores how the size of the priming set affects accuracy. We characterize accuracy both in terms of false positives (normal images incorrectly flagged as anomalous) and false negatives (anomalous images that were missed). We define ground truth to be the output of a two-pass offline anomaly detection system that gathers statistics over the entire data set in the first pass and identifies anomalies in the second pass. The priming set in our approach consists of those images that are used to seed the initial parameter estimates (the priming set is distributed across servers). The reported accuracy is measured on the remaining images in the dataset. The adipocyte dataset contains 1697 images and the neurite dataset contains 1062 images. Consistent with our expectations, the accuracy of the system improves quickly with the size of the priming set; this is important since in practice the priming set should be as small as possible.

The second set of experiments confirms that anomaly detection can be distributed over multiple compute/storage nodes without loss of accuracy. Although no single node can access the entire dataset, the sharing of descriptor statistics enables each node to build sufficiently accurate models for anomaly detection. We see no loss of accuracy in our 8-node distributed system; this is also true with greater numbers of nodes. Our approach is very amenable to parallelism and we observe near-linear scaling of performance with the number of nodes in the system (results not shown due to space limitations).

5. CONCLUSION

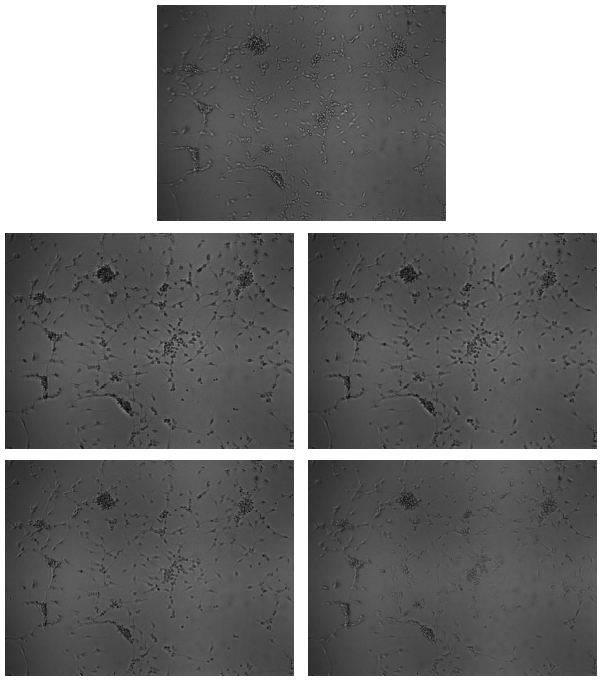
In the future, we plan to extend our work to larger image repositories and to other types of HCS images. We also plan to relax the assumption that the distribution of non-anomalous data is Gaussian in each



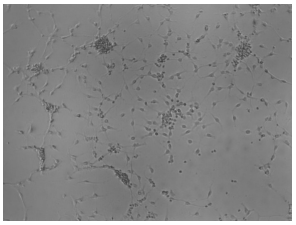
(a) Adipocytes

(b) Neurites

Fig. 5. Accuracy of online anomaly detection improves quickly with the size of the priming set



(a) Five focal views from a single neurite well



(b) Composite neurite image with focal planes merged and background illumination corrected

Fig. 4. Multi-focal plane brightfield neurite images.

feature dimension. This will enable our framework to operate with more sophisticated distributions such as mixtures-of-Gaussians and non-parametric representations such as histograms.

In closing, this work has presented an automated, online approach to anomaly detection in high-content screening assays for pharmaceutical research. This approach employs assay-specific image processing within an assay-independent framework for distributed control, machine learning, and anomaly reporting. Our results confirm that this online approach to anomaly detection is feasible, efficient, and accurate.

6. REFERENCES

[1] A. Fire, S. Xu, M. Montgomery, S. Kostas, S. Driver, and C. Mello, "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans," *Nature*, vol. 391, 1998.

[2] S. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl, "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells," *Nature*, vol. 411, 2001.

[3] L. Huston, R. Sukthankar, R. Wickremesinghe, M. Satyanarayanan, G. Ganger, E. Riedel, and A. Ailamaki, "Diamond: A Storage Architecture for Early Discard in Interactive Search," in *Proceedings of File and Storage Technologies*, 2004.

[4] "Diamond: Interactive Search of Non-Indexed Data," `http://diamond.cs.cmu.edu`.

[5] "ImageJ: Image Processing and Analysis in Java," `http://rsb.info.nih.gov/ij/`, National Institutes of Health.

[6] E. Kim, M. Haseyama, and H. Kitajima, "Fast and Robust Ellipse Extraction from Complicated Images," in *Proceedings of IEEE Information Technology and Applications*, 2002.

[7] A. Goode, M. Chen, A. Tarachandani, L. Mummert, R. Sukthankar, C. Helfrich, A. Stefanni, L. Fix, J. Saltzmann, and M. Satyanarayanan, "Interactive Search of Adipocytes in Large Collections of Digital Cellular Images," in *Proceedings of the International Conference on Multimedia and Expo*, 2007.