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[†]**

December 2007
CMU-CS-07-177

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[†]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.

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 Corp. or Merck, Inc.

Keywords: Anomaly detection, biomedical image processing, distributed database searching, high-content microscopy, 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 [6, 5]. 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.

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. This clean separation of assay-specific and assay-independent aspects of anomaly detection is made possible by our use of the OpenDiamond platform for distributed search [9, 1]. The image processing is performed by code components called *searchlets* that execute on servers close to the points of data collection. For each image of cells in a well, a searchlet generates a list of quantitative descriptors that are assay-specific. For example, in one assay, the total number of cells in a well, the average diameter of cells in that well, and the number of malformed cells in the well might all be descriptors. It is the values of these descriptors that define the data ranges over which anomaly detection is performed. One can thus view the descriptors as defining the *universe of discourse* over which anomaly detection is performed. Our approach is friendly to parallel processing, since the compute-intensive image analysis, generation of descriptors and statistics local to wells may be performed in parallel without coordination. It is only the summation and global analysis of this information, which is a computationally lightweight task, that needs to be serialized.

2 Background and Related Work

Anomaly detection is a broad concept, with applications ranging from network intrusion detection [3] to autonomous inspection of power plants [8]. While there has been work in anomaly detection for biomedical and pharmaceutical applications [4, 13], we are not aware of any work on detecting anomalies in an online setting across a large and growing collection of high-resolution images distributed across multiple computers.

Online learning is highly adaptive and highly scalable, and is implemented by an incremental algorithm that sweeps through a sequence of data items only once. On each data item, the learner makes a prediction and receives feedback so that training and testing take place at the same time. By its very nature, online anomaly detection is more susceptible to errors since early reports of anomalies are necessarily based on smaller samples of observed data. In spite of this intrinsic limitation, online detection seems to be an unavoidable choice when dealing with very large datasets and/or when the datasets are rapidly changing. There are a number of simple but robust online learning algorithms [14, 17, 15] that work well even when no statistical assumptions are made about the process producing the observed data. Many of these algorithms benefit from the early work of Littlestone [11, 12], and Vovk [16].

3 A Framework For Anomaly Detection

To enable anomaly screening at interactive speed, the adaptive learning algorithm has to be embedded in an efficient distributed infrastructure for compute-intensive tasks. Our anomaly detection framework is based on the OpenDiamond platform [9]. This is an open-source platform for distributed search of complex data that we have extended to support online anomaly detection. The platform is domain-independent: assay-specific aspects of user interaction and image processing are isolated within an OpenDiamond application. Each 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, the OpenDiamond platform 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.

As mentioned earlier, the OpenDiamond platform supports image processing on data 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 extent of the anomaly. The OpenDiamond platform conveys anomalous images along with their attributes to the client.

Our framework is well suited for parallel computation in that the OpenDiamond platform supports distribution of data and processing over a set of servers. Each server processes a subset of the data independently of other servers. A risk of this approach is that the descriptor statistics may differ substantially across servers, leading to non-uniform detection results compared to a single server. In addition, more images are needed to prime the descriptor statistics. We address these issues by sharing the priming count and descriptor statistics across servers. The client coordinates this sharing by periodically collecting, aggregating, and distributing the data across servers. The time to perform sharing is typically less than the time to process a single image; therefore, the time lag for shared data is small. The sharing period is configurable; the work in this paper used a sharing period of 5 seconds. In this way, anomaly detection applications realize the benefits of parallel processing with no loss of statistical accuracy.

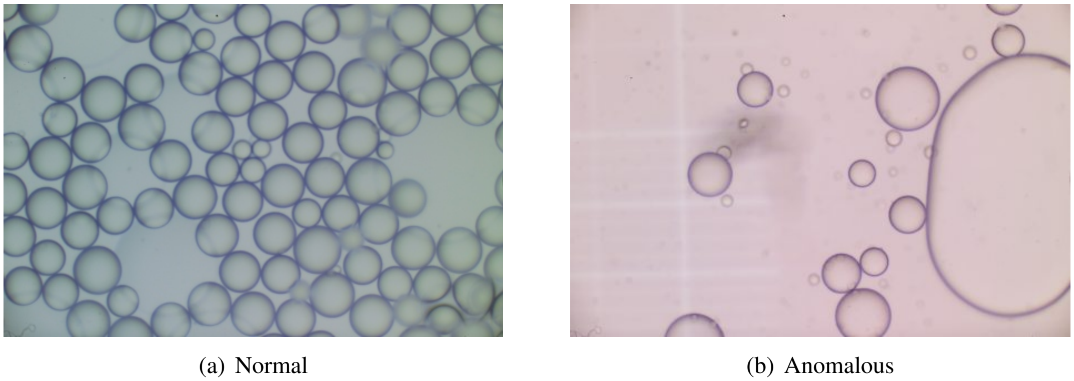
An important aspect of the framework is its ability to accomodate a wide variety of implementation methods for image processing. For example, the searchlet code for adipocyte images, described in section 4.1, is implemented in C++. In contrast, the searchlet code for neurite images, described in section 4.2, is implemented as a collection of ImageJ macros [2]. This diversity demonstrates the flexibility and extensibility of the framework.

4 Applications

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

4.1 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.



(a) Normal

(b) Anomalous

Figure 1: Example Adipocyte Images

4.1.1 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.

4.1.2 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. First, an image pyramid is built by scaling down the high-resolution images to enable efficient detection of large features. Then, a Canny-style edge detector is applied that uses color contrast gradients rather than grayscale contrast. The resulting binary edge images are used as input to an ellipse extraction algorithm that can locate overlapping and partially occluded cells [10]. The results from all pyramid levels are merged and duplicate detections are eliminated. Finally, statistics such as the cell count and cell size distribution are generated. Further details on locating adipocytes can be found in Goode *et al.* [7].

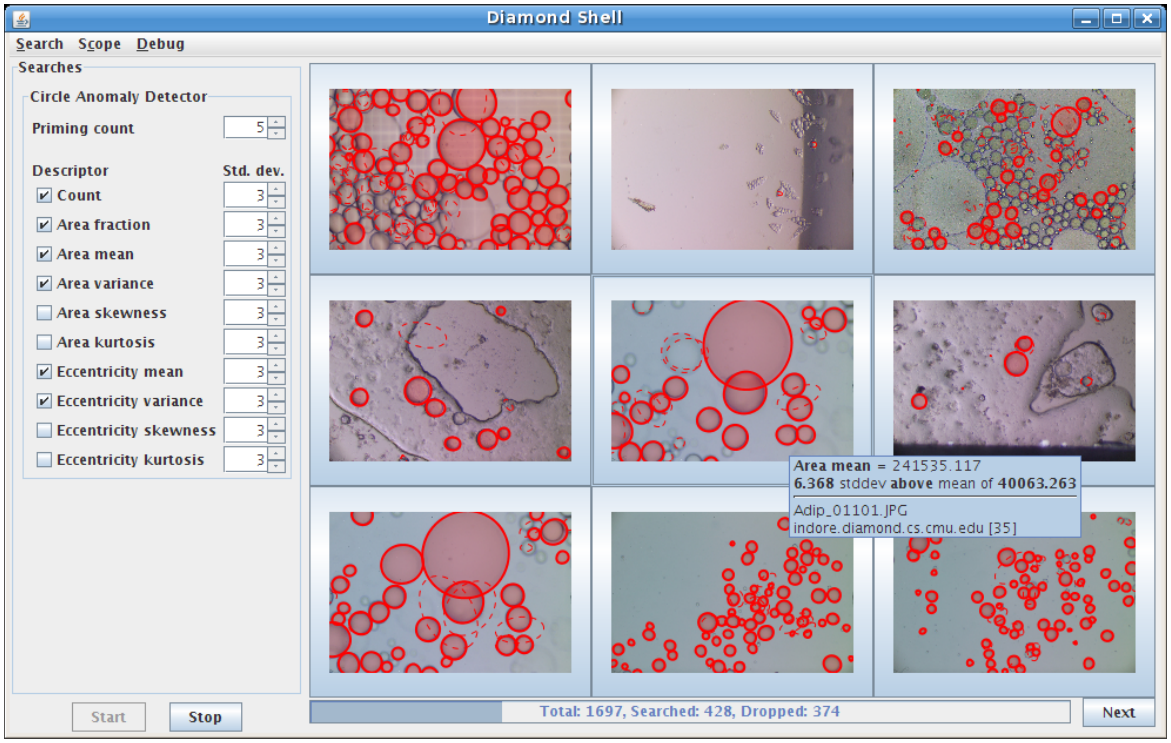
Anomalies in the adipocyte images are detected based on the cell count, the fraction of the image covered by cells, and the first four statistical moments of cell size and shape (eccentricity). Figure 2 shows an application for detecting anomalies in adipocyte images based on the framework described in Section 3. 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.

4.2 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.

4.2.1 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.



Searches

Circle Anomaly Detector

Priming count 5

Descriptor Std. dev.

☑ Count 3

☑ Area fraction 3

☑ Area mean 3

☑ Area variance 3

□ Area skewness 3

□ Area kurtosis 3

☑ Eccentricity mean 3

☑ Eccentricity variance 3

□ Eccentricity skewness 3

□ Eccentricity kurtosis 3

Area mean = 241535.117
6.368 stddev above mean of 40063.263
Adip_01101.JPG
indore.diamond.cs.cmu.edu [35]

Total: 1697, Searched: 428, Dropped: 374

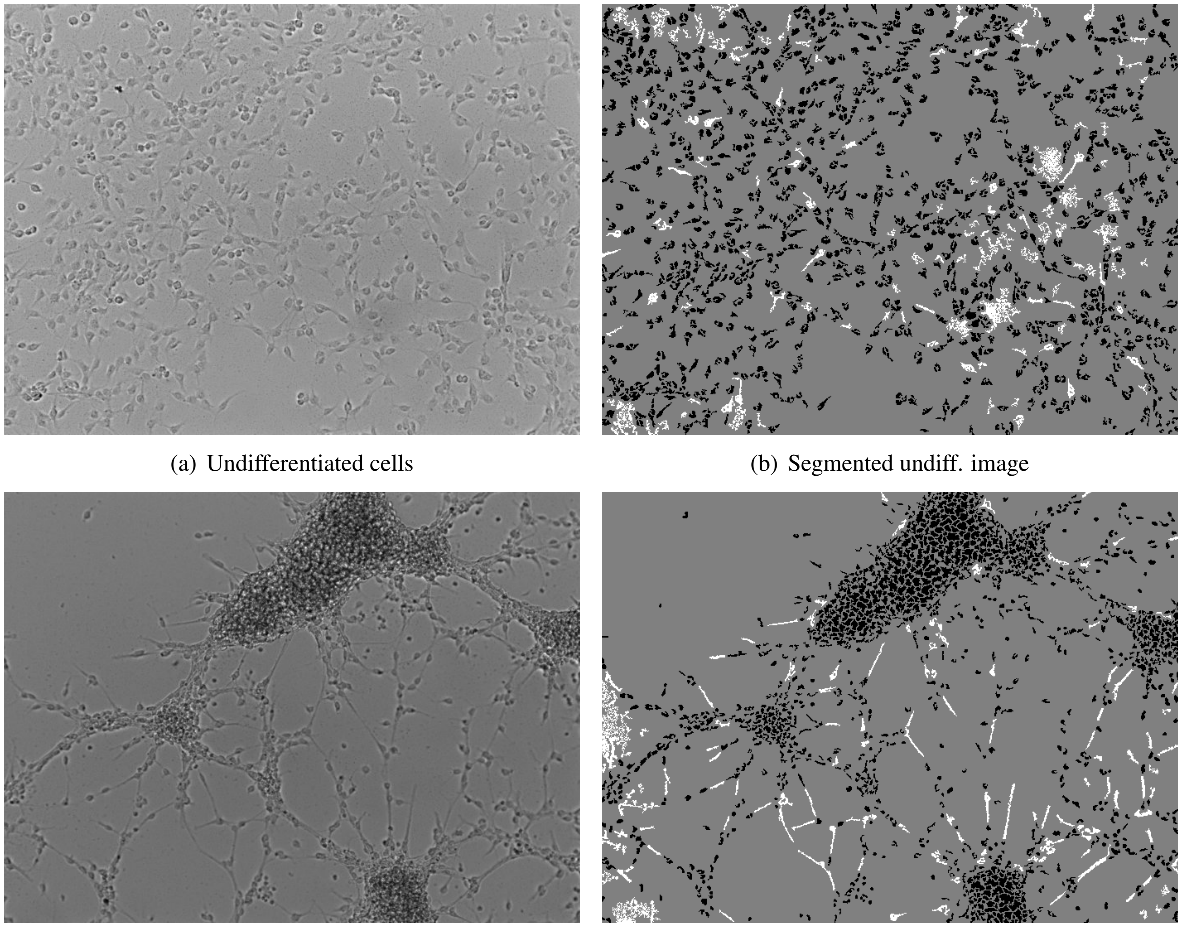
Figure 2: Screenshot of Anomaly Detection Application

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 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.

4.2.2 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 identify and mask out pixels corresponding to cell bodies. This is accomplished by a straightforward adaptive thresholding procedure that exploits the fact that cell bodies correspond to high-intensity regions in the image. Once the cell bodies have been masked out, the relatively low contrast between neurite pixels and the background becomes can be enhanced so as to segment them. 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. Finally, we compute statistics from the extracted neurites and neural cell bodies.



(a) Undifferentiated cells

(b) Segmented undiff. image

(c) Differentiated cells

(d) Segmented diff. image

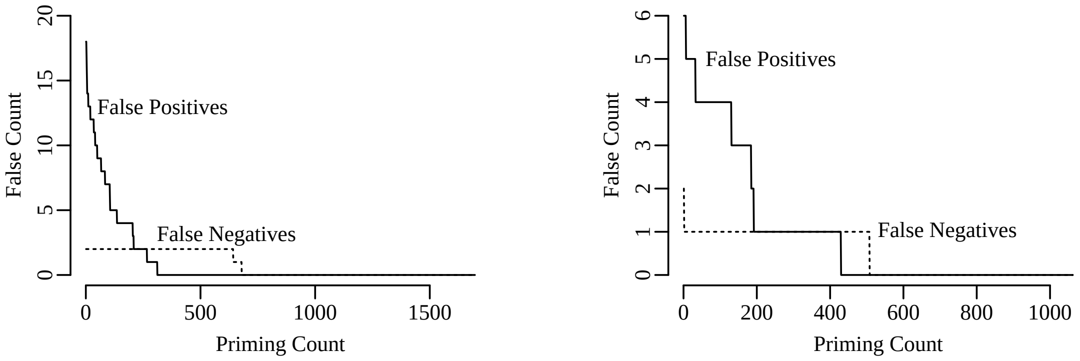
Figure 3: Examples from the neurite growth domain. Two conditions from the same well are shown (temporally spaced): undifferentiated cells (a) and differentiated cells (c). The segmented versions of those images are shown in (b) and (d), respectively; detected neurites are evident as bright linear features,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

5 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 4) 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. For neurites, where the data is stratified into several different "normal" distributions, we perform anomaly detection independently for each case (e.g., undifferentiated cells, differentiated cells). 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



(a) Adipocytes

(b) Neurites

Figure 4: Accuracy of online anomaly detection improves quickly with the size of the priming set

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).

6 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 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.

Acknowledgments

We acknowledge Dr. Laszlo Kiss (Merck Research Labs) for his encouragement with the neurite outgrowth analysis. OpenDiamond is a registered trademark of Carnegie Mellon University.

References

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

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

[3] Bace, R., and Mell, P. NIST Special Publication on Intrusion Detection Systems, August 2001.

[4] Chen, M., Kanade, T., Rowley, H., and Pomerleau, D. Anomaly Detection through Registration. In *Proceedings of Computer Vision and Pattern Recognition* (1998).

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

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

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

[8] GREGORI, M., LOMBARDI, L., SAVINI, M., AND SCIANNA, A. Autonomous Plant Inspection and Anomaly Detection. *Lecture Notes in Computer Science 1311* (1997).

[9] HUSTON, L., SUKTHANKAR, R., WICKREMESINGHE, R., SATYANARAYANAN, M., GANGER, G., RIEDEL, E., AND AILAMAKI, A. Diamond: A Storage Architecture for Early Discard in Interactive Search. In *Proceedings of File and Storage Technologies* (2004).

[10] KIM, E., HASEYAMA, M., AND KITAJIMA, H. Fast and Robust Ellipse Extraction from Complicated Images. In *Proceedings of IEEE Information Technology and Applications* (2002).

[11] LITTLESTONE, N. Learning when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning 2* (1988).

[12] LITTLESTONE, N., AND WARMUTH, M. The weighted majority algorithm. *Information and Computation 108* (1994).

[13] MINHAS, A., AND REDDY, M. Neural network based approach for anomaly detection in the lungs region by electrical impedance tomography. *Physiological Measurement 26* (2005).

[14] PARRA, L., AND SPENCE, C. On-line convolutive source separation of non-stationary signals. *Journal of VLSI Signal Processing 26*, 1/2 (2000).

[15] PETROVIC, N., JOJIC, N., FREY, B., AND HUANG, T. Real-time on-line learning of transformed hidden Markov models from video. In *Proceedings of International Workshop on Artificial Intelligence and Statistics* (2003).

[16] VOVK, V. Aggregating strategies. In *In Proceedings of the Third Annual Workshop on Computational Learning Theory* (1990).

[17] YU, K., AND LAM, W. A new on-line learning algorithm for adaptive text filtering. In *Proceedings of Information and Knowledge Management* (1998).