Quality Classification of Microscopic Imagery with Weakly Supervised Learning

Xinghua Lou, Luca Fiaschi, Ullrich Koethe, and Fred A. Hamprecht

HCI, IWR, University of Heidelberg, Speyererstr. 6, 69115, Germany
http://hci.iwr.uni-heidelberg.de/MIP/

Abstract. In this post-genomic era, microscopic imaging is playing a crucial role in biomedical research and important information is to be discovered by quantitatively mining the resulting massive imagery databases. To this end, an important prerequisite is robust, high quality imagery databases. This is because defect images will jeopardize downstream tasks such as feature extraction and statistical analysis, yielding misleading results or even false conclusions. This paper presents a weakly supervised learning framework to tackle this problem. Our framework resembles a cascade of classifiers with feature and similarity measure designed for both global and local defects. We evaluated the framework on a database of images and obtained a 96.9% F-score for the important normal class. Click-and-play open source software is provided.

1 Introduction

Modern biomedical research heavily relies on large scale experiments and controlling the quality of the resulting data is crucial for any meaningful analysis. Whereas this problem has been investigated thoroughly for "-omic" techniques such as microarray [13], there has not been sufficient work on controlling the quality of microscopic imagery databases [15], despite the significant role imaging techniques are playing in this post-genomic era. Existing approaches depend on manual inspection via visualization or semi-automated processing [5,1]. However, the increasing scale and resolution of biomedical experiments such as high-content screening (HCS) [4] and high-resolution 3D connectomic data [8] has raised urgent demand for scalable quality control approaches. We are seeking for automated, efficient method for detecting defect images from large scale image databases. Image defects can occur during sample preparation, such as debris contamination, and also during image acquisition, such as out-of-focus [5,1]. They will jeopardize downstream tasks including registration, segmentation, tracking as well as statistical analysis. Usually defect images are rare and exhibit large variability of appearance. For example, w.r.t. normal images (Fig. 1A and E), defects can occur at the full image scale due to out-of-focus (Fig. 1B, D and F), also at particular regions within an image due to debris comtamination (Fig. 1C, G and H).

Many challenges arise for quality control in large scale microscopic imagery databases. Firstly, supervised learning algorithms (support vector machine, random forest, etc.) [6] becomes inapplicable in practice. Because the rareness of

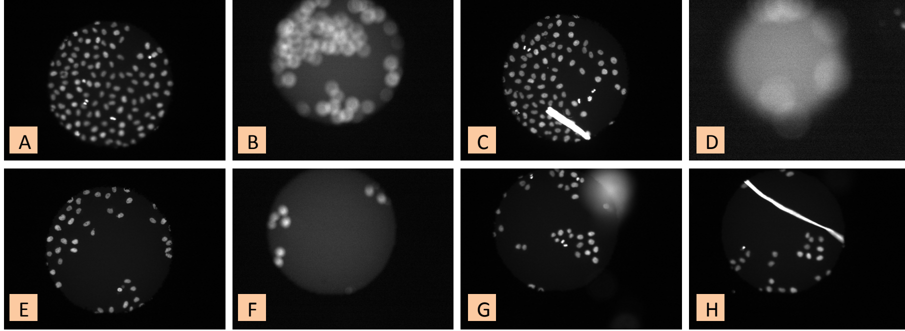

Fig. 1. Examples of normal and defect images in a high-content screening imagery database: A and E - normal images; B, D and F - out-of-focus; C, G and H - debris.

defect images makes it too time consuming to collect sufficient training samples, which may require manual screening of the entire dataset. Secondly, it is also difficult to directly model the defects because of the large variability in scale and appearance [1]. Finally, the increasing quantity and resolution of images in such databases prohibit any manual inspection and filtering, and require algorithmic scalability as well as support of parallel computing [5].

We present a framework to address this important problem and we pursue two goals: low labeling efforts and high scalability. We cast this problem as an outlier detection problem [3] (i.e. defect images as outliers) and chose to develop our framework based on the one-class SVM [18]. Briefly, one class SVM only requires training samples from the normal class and, in some projected space by kernalization, it finds the most compact "ball" to enclose those samples. Test samples outside this ball (i.e. the decision boundary) will be classified as outliers.

Several outlier detection algorithms have been proposed in the literature, such as statistical models [7], distance measure [9], density estimation [2] or space partition [11]. We opt to choose one-class SVM for its capability of implicit feature projection via the kernel trick [19], which is frequently needed when handling image data. On the contrary, for example, isolation forest [11] partitions the original feature space by decision trees and determines outliers as those samples with a short path to the root. Despite their high scalability, they are restricted to the original feature space and extension by kernalization is not obvious. This encounters problems when handling image features, which are usually histograms, and which require kernalized similarity measure (e.g. earth mover's distance based [17]).

2 Defects in Microscopic Images: Global vs. Local

We group common causes for image defects into two classes, depending on whether they affect the image globally or locally. A typical cause for global defect is out-of-focus imaging and typical examples of regional defects such as debris

contamination (e.g. hair) [1]. We handle these two types of defects differently with appropriate features and similarity measure, which allows for predicting three classes (normal, globally defect and regionally defect) even when training samples are only provided for the normal class.

For handling global defect, one important motivation is that it must be reflected in the statistics drawn from the entire image. For example, the formation of images is the convolution of the real light with the point spread function (PSF). When out-of-focus occurs, the PSF becomes wider, and this can be seen from the intensity histogram drawn from the entire image (e.g. Fig. 2A vs. Fig. 2B).

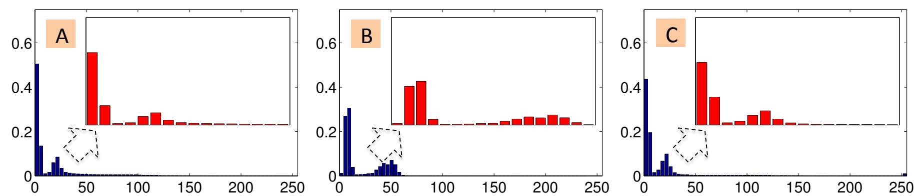

Fig. 2. Examples of the intensity histograms of normal and defect images. From left to right, the histograms corresponding to image A to C in Fig. 1. The red histogram inside is the zoomed view showing the intensity range of interest (between 0 and 64).

The task becomes more difficult when regional defects occur, because they exhibit considerable variability in scale, position and shape. A global statistic is no longer informative, e.g. Fig. 2A (normal) very similar to Fig. 2C (regional defect), and extracting information from fine regional details becomes necessary. In addition, regional defects show significant variability in appearance, implying the requirement for more features to achieve the required discriminative power.

3 Classification by One-Class SVM Cascade

To exploit the characteristics of global and regional defects and handle these two classes properly, we propose the quality classification framework shown in Fig. 3. Briefly, stage one operates on the full image level and aims at filtering out globally defect images. Stage two and three work on patch level, coupled to form a coarse-to-fine procedure for detecting regional defects. The overall framework resembles a cascade of one-class SVM classifiers.

3.1 Global Out-of-Focus Detection by Histogram Comparison

To efficient compare two images with different focus, various methods have been proposed in the computer vision community for natural image deblurring (see [12] and references therein). We follow the same intuition – out-of-focus blurring mainly affects the high frequency part (e.g. texture details) of an image. In

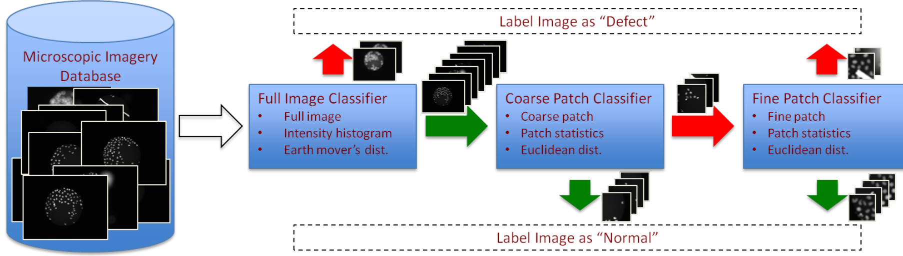

Fig. 3. Workflow of the proposed one-class SVM cascade. Red and green arrows indicate the flow of detected outlier and normal images/patches, respectively.

particular, we build a histogram of the Gaussian gradient magnitude to capture the high frequency part of an image. This histogram is used as the input feature for constructing the first one-class SVM that detects out-of-focus images. In particular, we first normalize the histogram and kernalize it using earth mover's distance (EMD) [17]. EMD describes the efforts required for transporting probability mass from one distribution (i.e. normalized histogram) to the other, and has proven superior to the Euclidean distance measure [17] (though the later is computationally much cheaper). Formally, given two normalized histograms ($\boldsymbol{h}_i$ and $\boldsymbol{h}_j$), the kernel for out-of-focus detection is

$$K^{\text{EMD}}(\boldsymbol{h}_i, \boldsymbol{h}_j) = \exp(-\lambda^{\text{EMD}} EMD(\boldsymbol{h}_i, \boldsymbol{h}_j)) \tag{1}$$

Here, λ^{EMD} adjusts the scale of the EMD response. Note that, in order to have a valid kernel for one-class SVM, the histogram must be normalized [14].

3.2 Regional Defect Detection from Patch Statistics

We have already shown the need for finer level analysis: regional defects are not possible to capture from full image statistics. Moving from image level to patch level is not direct: unlike out-of-focus, regional defects can occur at any location and scale, and exhibit arbitrary appearance. Also, one has to consider the increasing complexity: hundreds of patches may need be extracted per image from a database of thousands of images yielding a new problem of size million.

We employs two techniques for regional defect detection. Firstly, we draw basic statistics from low level features and use RBF kernel for patch similarity meaure. Secondly, we construct a coarse-to-fine procedure for speedup.

Low Level Features and Patch Statistics We use low level features from to characterize the images from different aspects including texture (structure tensor), edge (gradient magnitude), and local extreme (eigenvalues of Hessian). For each feature, the following statistics are drawn from the patch: mean, standard deviation and quantiles (10%, 50% and 90%). For patch classification, we move away from histogram and EMD kernel because of its high computational cost:

EMD is more expensive than Euclidean distance (for RBF kernel) by several orders of magnitude. .

On Feature Bagging and Classifier Ensembles The high dimensional patch statistics is used as input features to the one-class SVM. Inevitably, some have no positive contribution to the patch similarity measure. Unfortunately, we cannot perform feature selection as in supervised learning. This problem is solved using feature bagging and classifier ensemble [10]. Briefly, we sample subsets of features (viz. bagging) and train a one-class SVM on each subset individually. The intuition is: important features become more influential in a lower dimensional feature subset, and accumulating votes from the ensembles brings more robustness than a single one-class SVM trained on all features.

To illustrate the improvement in discriminative power from feature bagging, we randomly sample 500 patches for each normal and defect class and plot their RBF kernel in Fig. 4. We hope to have an ideal kernel (Fig. 4A) that makes all normal samples (first 500 rows/columns) completely similar to each other and fully distinct to defect samples (the remaining rows/columns). The kernel computed using all features (Fig. 4B) does not exhibit the desired property because the important features are averaged out by the dimension of the input. This is improved when using feature bagging. We can see the improvement from the average kernel computed with bagged features (Fig. 4C): the contrast between the normal and defect samples is apparently enhanced, which implies improved discriminative power. Note that, in the context of one-class SVM, we do not have to make the defect samples all similar to each other because they can be distributed arbitrary outside the decision boundary (ball).

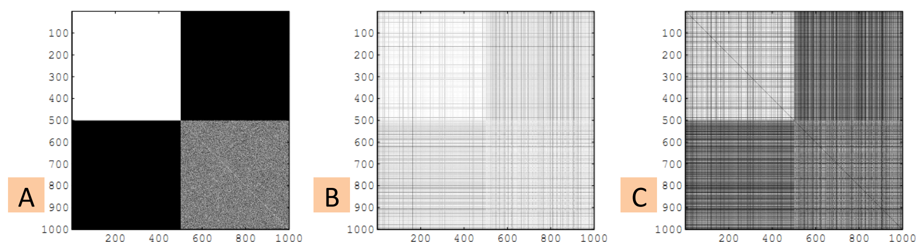

Fig. 4. Kernel matrices for 500 normal and 500 defect samples: A – ideal kernel; B – kernel using all features; C – average kernel from feature bagging.

Coarse-to-Fine Filtering Procedure Observing that a significant amount of image regions are “obviously” normal ones (such as background, regions with sparse objects), we incorporate a coarse-to-fine filtering procedure for speedup. The “fine” step (stage three in Fig. 3) operates on small patches (thus expensive). The speedup is obtained at the “coarse” step, i.e. stage two in Fig. 3, which

operates on larger patches. In particular, stage two "filters out" easy normal image regions so that they can be skipped in the expensive stage three. Large patch size may average out small defect regions and produce a false normal patch. To prevent this, we made stage two more selective on determining normal images by setting a high ν value to the one-class SVM[18].

4 Experimental Results

We evaluated our framework on an image database for mammalian cell culture study. The new 9216-microwell cell array (in a 96×96 layout) [16] was used, yielding one image per well (Fig. 1). An automated scanning microscope was used with an overall imaging time of around 10 hours (4 seconds/image). Our approach is wrapped into a click-and-play software implementation that is available to the public[1].

It is important to select normal, training images with different characteristics (e.g. cell density, illumination, etc.) such that the training features (histograms or patch statistics) are not biased towards any particular type of normal images. Also, it is helpful to train the system incrementally. That is, starting with some training images, train and predict on a small subset of images; select representative samples from the wrongly predicted ones, add them to the training set and retrain the system. We made two rounds of incremental learning and eventually found 140 (out of 9216) training images. Overall, the framework took roughly 2.5 hours to complete the prediction on the entire dataset on a 4 core (2.8G-Hz) machine. Training time is roughly 5 minutes per stage.

We define normal images as those that are useful for our cell segmentation and counting task and generate manual ground truth accordingly. Table. 1 shows the overall detection accuracy by our framework, depicted as a confusion matrix (rows being the ground truth), and the per class precision/recall is given in Table 2. The parameter settings that yields this result are biased to more tolerance of false positive rate, because it is more costly to mistake defect images for normal images. Note that the definition of "normal" may change with the task of the analysis. For example, slightly out-of-focus images are useful for cell counting but useless for phenotype classification.

	Normal	Out-of-Focus	Regional
Normal	7854	146	338
Out-of-Focus	1	426	28
Regional	19	47	357

Table 1. Classification confusion matrix.

	Precision	Recall	F-score
Normal	0.997	0.942	0.969
Out-of-Focus	0.688	0.936	0.793
Regional	0.494	0.844	0.623

Table 2. Per class precision and recall.

Some examples of detected regional defects are shown in Fig. 5. Our framework shows high accuracy on detecting regional defects, even though they exhibit

[1] `http://ilastik.org/`

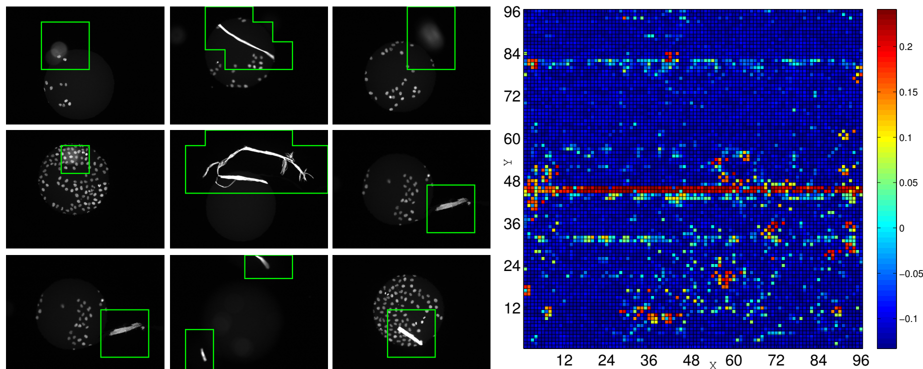

Fig. 5. Examples of regional defects found by our framework.

Fig. 6. Location of out-of-focus images on 96x96 well plate.

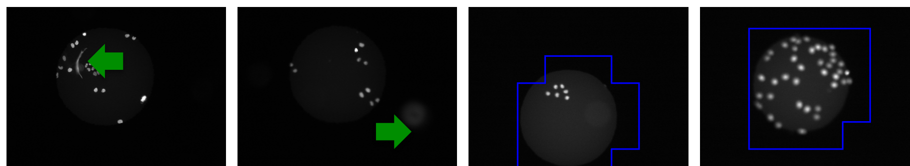

Fig. 7. Examples of errors in regional defect detection: misdetected (left two) and false positive (right two).

strong variability in size, shape, texture and other characteristics. Fig. 7 shows some errors by our framework. We notice that misdetection occurs when the regional defect is not sufficiently strong (left two images).

Fig. 6 shows the detected out-of-focus images represented by their signed distance to the classifier's decision boundary in a 96×96 cell array layout. Higher value indicates more severe out-of-focus error. As we can see from the prominent strip in the center, some systematic error caused the microscopy to malfunction during the entire acquisition time for row 45 and 46. This suggests investigation and helps to avoid such systematic errors in future experiments.

Discussion: It is worth pointing out that training data preparation would be too expensive for two-class learning. Firstly, defects exhibit huge variability in appearance, which forces users to collect "defect" images by browsing through huge databases. This task is tedious and becomes more so if there are few positive (defect) images, as is desirable from the experimental point of view. Secondly, we showed the necessity of patch level classification for detecting regional defects. Two-class learning will require users to explicitly mark each defect region/patch, thus even more expensive. We avoid this excessive labeling efforts in our approach.

5 Conclusions and Outlook

This paper presents a framework for microscopic image quality control based on one class learning. We studied the distinct properties of global and local defects in microscopic images and proposed appropriate features and similarity measures for them. At the same time we show that its possible to distinguish globally and regionally defect images with a scalable cascade of one class classifiers using only training images from the normal class.

In the future, we plan to integrate our method with the automated microscopy control. This offers the advantage that the detection results can give feedback to correct image acquisition in place and in time. Also, given our generic framework, we plan to extend our method to other biomedical imaging scenarios.

References

1. M.A. Bray, A.N. Fraser, T.P. Hasaka, et al. Workflow and Metrics for Image Quality Control in Large-Scale High-Content Screens. *J Biomol Screening*, 2011.
2. M. M. Breunig, H. P. Kriegel, R. T. Ng, et al. LOF: identifying density-based local outliers. *ACM Sigmod Record*, 29(2):93–104, 2000.
3. V. Chandola, A. Banerjee, and V. Kumar. Outlier detection: A survey. *ACM Comput Surv*, 2007.
4. C. J. Echeverri and N. Perrimon. High-throughput RNAi screening in cultured cells: a user's guide. *Nat Rev Genet*, 7(5):373–384, 2006.
5. A. Goode, R. Sukthankar, L. Mummert, et al. Distributed online anomaly detection in high-content screening. In *ISBI*, 2008.
6. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
7. A. O. Hero. Geometric entropy minimization (GEM) for anomaly detection and localization. In *NIPS*, 2006.
8. V. Kaynig, B. Fischer, and J. M. Buhmann. Probabilistic image registration and anomaly detection by nonlinear warping. In *CVPR*, 2008.
9. E. M. Knox and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *VLDB*, 1998.
10. A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *KDD*, 2005.
11. F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation Forest. In *ICDM*, 2008.
12. R. Liu, Z. Li, and J. Jia. Image partial blur detection and classification. In *CVPR*, 2008.
13. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–1161, 2006.
14. O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009.
15. R. Pepperkok and J. Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nat Rev Mol Cell Bio*, 7(9):690–696, 2006.
16. J. Reymann, N. Beil, J. Beneke, et al. Next-generation 9216-microwell cell arrays for high-content screening microscopy. *BioTechniques*, 47(4):877, 2009.
17. Y. Rubner, C. Tomasi, and L. J. Guibas. A Metric for Distributions with Applications to Image Databases. In *ICCV*, 1998.
18. B. Schoelkopf, J. C. Platt, J. Shawe-Taylor, et al. Estimating the support of a high-dimensional distribution. *Neural Comput*, 13(7):1443–1471, 2001.
19. B. Schoelkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, MA, 2002.