Published in final edited form as:

IEEE Trans Med Imaging. 2007 October ; 26(10): 1401–1411. doi:10.1109/42.650887.

Automated Segmentation and Classification of High Throughput Yeast Assay Spots

Kourosh Jafari-Khouzani,

Image Analysis Laboratory, Radiology Department, Henry Ford Health System, Detroit, MI 48202 USA and also with the Department of Computer Science, Wayne State University, Detroit, MI 48202 USA (phone: 313-874-4378; fax: 313-874-4494; e-mail: kjafari@rad.hfh.edu)

Hamid Soltanian-Zadeh[Senior Member, IEEE],

Image Analysis Laboratory, Radiology Department, Henry Ford Health System, Detroit, MI 48202 USA and also with the Control and Intelligent Processing Center of Excellence, Electrical and Computer Engineering Department, Faculty of Engineering, University of Tehran, Tehran, Iran (e-mail: hamids@rad.hfh.edu)

Farshad Fotouhi,

Department of Computer Science, Wayne State University, Detroit, MI 48202 USA (e-mail: fotouhi@wayne.edu)

Jodi R. Parrish, and

Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, MI 48201 USA (e-mail: jparrish@genetics.wayne.edu)

Russell L. Finley Jr.

Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, MI 48201 USA (e-mail: rfinley@wayne.edu)

Abstract

Several technologies for characterizing genes and proteins from humans and other organisms use *yeast growth* or *color development* as read outs. The yeast two-hybrid assay, for example, detects protein-protein interactions by measuring the growth of yeast on a specific solid medium, or the ability of the yeast to change color when grown on a medium containing a chromogenic substrate. Current systems for analyzing the results of these types of assays rely on subjective and inefficient scoring of growth or color by human experts. Here an image analysis system is described for scoring yeast growth and color development in high throughput biological assays. The goal is to locate the spots and score them in color images of two types of plates named “X-Gal” and “growth assay” plates, with uniformly placed spots (cell areas) on each plate (both plates in one image). The scoring system relies on color for the X-Gal spots, and texture properties for the growth assay spots. A maximum likelihood projection-based segmentation is developed to automatically locate spots of yeast on each plate. Then color histogram and wavelet texture features are extracted for scoring using an optimal linear transformation. Finally an artificial neural network is used to score the X-Gal and growth assay spots using the extracted features. The performance of the system is evaluated using spots of 60 images. After training the networks using training and validation sets, the system was assessed on the test set. The overall accuracies of 95.4% and 88.2% are achieved respectively for scoring the X-Gal and growth assay spots.

Index Terms

Color image analysis; high throughput yeast assays; neural networks; texture classification; wavelet transforms

I. Introduction

High throughput assays have become an important part of the discovery process in biological and biomedical research. A variety of high throughput technologies have been developed, for example, to discover how genes and their encoded proteins function together in normal and disease states [1], [2]. Many of these technologies use the baker's yeast, *Saccharomyces cerevisiae*, either as a model organism to study biological processes that yeast have in common with humans, or as an assay system to test the functions of genes and proteins from humans or other model organisms [3]. For example, a technology known as the "yeast two-hybrid system" uses yeast cells essentially as test tubes to assay interactions between specific proteins from other organisms [4], [5]. Other assays have been developed that use yeast cells to screen for drugs or new drug targets [6]–[8]. A common feature of these and other yeast-based assays is that they measure the ability of the yeast to grow or to change color on solid media. In the yeast two-hybrid system, a protein-protein interaction results in activation of two different reporter genes. One reporter gene is required for growth and the other encodes an enzyme that converts a colorless compound to a color in the visible spectrum. Therefore, growth and color in the yeast cells are the characteristics analyzed in the yeast two-hybrid system. Generally, digital photographs are taken of the plates containing the solid media and yeast, and the level of yeast growth and the color of the yeast colony for each assay are scored manually. This is a tedious, time-consuming process that is particularly rate-limiting for high throughput screens that conduct hundreds to thousands of these assays in each experiment; see e.g., [9], [10].

We have developed an image analysis system for scoring yeast growth and color development in images of 96-well plates, a common format for high throughput assays. We employed the system to analyze the results of high throughput yeast two-hybrid screens that use a modified yeast two-hybrid system [12]. In these screens, yeast cells are spotted on two flat solid media plates in the common 96-well format, with twelve columns and eight rows of spots. Fig. 1 shows examples of these plates. Each spot represents a separate assay for an interaction between two specific proteins. One of the plates (the "growth assay" plate shown on the right side of images in Fig. 1) lacks a specific nutrient so that the yeast will not grow unless the interacting proteins activate a reporter gene required for synthesis of the nutrient. The other plate (left on images in Fig. 1) contains the chromogenic substrate "X-Gal", which will be converted to a blue compound in yeast that express the other reporter gene. This configuration provides two independent assays for protein-protein interactions; the same 96 spots are placed on both the X-Gal and growth assay plates. After the yeast are given a chance to grow, the plates are photographed. Note that the two plates are generated at the same time and must be photographed together since they represent the same 96 protein-protein interactions assayed in two different ways, and their scores must be reported together. The spots tend to have a circular shape as shown in Fig. 1 (a), though some may have irregular shapes as shown in Fig. 1 (b).

On the X-Gal plate, the level of reporter activity is scored based on the amount of blue color of the yeast on a scale of integers $\{0, 1, \dots, 5\}$, where 0 is white and 5 is dark blue. Fig. 2(a)

¹The score of an X-Gal plate could be considered as a real value in the interval $[0, 5]$, since in reality the score continuously changes. However, according to the convention and for simplicity the score is reported as an integer value.

shows examples of spots on X-Gal plate with their scores. In some circumstances, the yeast grow poorly on the X-Gal plate. In this case, if the spot has score 0 it will be darker than usual and looks like the background (see the spot in upper left corner of Fig. 2 (a)). Examples of background spots in X-Gal plates are the spots in location (1, 5), i.e., row 1 and column 5, in Fig. 1 (a), and locations (2, 6), (6, 9), and (8, 10) in Fig. 1 (b). In contrast, other spots on the X-Gal plates in which the growth is normal but the color score is still 0 we refer to as *non-background* spots. The reason that we distinguish these two types of spots (background and non-background) is that despite having the same score, they have different colors. This distinction will result in a more accurate classification.

Several studies have shown that the amount of growth and the amount of blue color correlate with the level of activation of the reporter genes, and in turn with the general affinities of the protein-protein interactions being assayed [11]. This quantitative information promises to be particularly useful for developing computational approaches to interpret protein interaction data.

To classify the X-Gal spots, color features are desirable. However, as shown in Fig. 2(a), the area of the color-changed region varies within each score. This causes the mean values of the color components (e.g., RGB or HSV) of the spots with the same scores to vary, making them unreliable features for classification. This suggests that using color histograms instead of color mean values will be more effective. Although the color histogram may vary within each score, it contains information that the mean values of the color components may not have. For example, two spots with different scores may have similar mean values but different color histograms. In this paper, we propose a new histogram technique to extract color features.

According to convention [12], the amount of yeast on the growth assay plate is scored on a scale of integers $\{0,1,2,3\}$ ¹, where 0 indicates no growth and 3 indicates maximum growth. Fig. 2(b) shows examples of the spots on the growth assay plates with their scores. Similar to the X-Gal plates, the intensity and area of the growth region on the growth assay plate may not determine the score. Due to inevitable color variations in the media and consequently growth assay spots, the mean intensity may vary significantly within the spots of the same score, even after adjusting the illumination of the plate image. This also causes the manual scores to be highly subjective. Fig. 3 shows the range of average intensities of spots with the scores of 0, 1, 2, and 3 based on measurements on 40 spots (10 spots from each score). As shown, there exist overlaps between the intensity ranges (especially between scores 0 and 1). Therefore, the mean intensity alone cannot determine the score. Indeed, the density of growth determines the score regardless of the total growth area. For example, some spots of score 3 in Fig. 2(b) have smaller but more dense growth areas compared with some spots of score 2 in the same figure. However, their *texture properties* change as the score increases, which suggests that texture features should be used to classify the spots. We use the wavelet transform for texture feature extraction.

In summary, the problem can be formulated as follows. We have color images of two types of plates named “X-Gal” and “growth assay” plates, with uniformly placed spots (cell areas) on each plate. The goal is to locate the spots and score them even when there are variations in the media color or illumination. The scoring system relies on color for the X-Gal spots, and texture properties for the growth assay spots. The block diagram of the proposed system to segment and score the spots is depicted in Fig. 4. As shown, the image is first segmented into two plates. The spots on each plate are segmented and features are extracted separately

¹Similar to X-Gal scores, the growth assay scores also could be considered as real values, but conventionally the integer values are reported.

from spots of each plate. The spots are finally scored using an artificial neural network (ANN) as classifier.

Related image analysis and quantification problems have been addressed for other bioinformatics applications in [13]–[19]. A substantial amount of work has been done, for example, to automate analysis of DNA and protein microarrays. Such analysis involves segmentation of the microarray images to identify the spots of DNA or protein, and processing of each spot based on intensity information. In [13] and [14], an initial grid is manually overlaid on the array of spots. A clustering technique is then employed to extract spot intensities. Katzer *et al.* [15] model the gridding problem as a Markov random field labeling problem. Liew *et al.* [16] use adaptive thresholding and statistical intensity modeling to locate the good-quality spots as guide spots and generate the grid structure based on the geometry of the guide spots. Hirata *et al.* [17] use a projection-based technique and mathematical morphology to create the grid structure. The orientation of the grid is manually adjusted before applying the technique. In [18] the need for a grid is eliminated using a dynamic model based on pixel clustering.

More related to analysis of microbial growth, an automated method has been presented for identification of bacterial types [19]. In this work, the spots of growing bacteria are found and analyzed based on features of “normalized area” and “shape index”. However, to the best of our knowledge, no automated methods to extract quantitative information directly from images of high throughput yeast assays have been reported.

The problem of analyzing yeast plates, addressed in this paper, shares some of the same issues (i.e., segmentation of a set of uniformly placed spots) with these other applications but also has several unique features. First, each image in the yeast application contains two different plates each with unique spot features. We use projection-based segmentation techniques to segment the plates and then locate the spots in each plate. Second, although all images are taken in similar environments, the lighting conditions, spot shapes and background may vary from plate to plate, unlike the much more uniform backgrounds that can be achieved with microarrays. For example, the color of the media on which the yeast is growing, which forms the background of the image, inevitably varies with each new preparation of media. Therefore, illumination must be adjusted before feature extraction. Furthermore, the size and positions of the spots cannot be made more exact due to the nature of the experimental system. Although the suspension of yeast cells is spotted robotically, the fluid dynamics cannot be controlled more precisely to ensure that the yeast cells are uniformly distributed in every spot. The thickness of the media can also vary, leading to small variations in spot shapes and positions relative to each other and to the plate edges. In addition, the plates may be rotated slightly due to imperfect photography settings, which may deteriorate the segmentation process. We use a Radon transform orientation estimation technique to correct the orientation. This makes the segmentation result more accurate. Finally, the color and texture information determine the score on the plates. Our system must extract color information and texture features from the spots on the X-Gal and growth assay plates, respectively. The sets of extracted features are then fed into an ANN for classification. The proposed techniques for segmentation and scoring are generic in nature and can be extended to different types of plate images.

The outline of this paper is as follows. In Section II, we describe the proposed technique to segment the plates and locate the spots in each plate. In Section III, we present the proposed features to be extracted from each spot. The classification method is presented in Section IV. Experimental results are explained in Section V. Conclusions are given in Section VI.

II. Segmentation

As shown in Fig. 1, the image contains two plates with 8×12 spots located in a dark background (a black screen underneath the plates). The plates are placed in parallel. The images are captured using a Kodak DC290 (2.1 Mega Pixel resolution), with the output images of 1200×1792 pixels in JPEG format. The resolution and quality of the images are adequate for visual analysis. We accomplish segmentation and localization of the spots of each plate in two steps. In the first step, the plates are separated from the background. In the second step, the spots are located in each plate. These steps are explained in the following subsections, and depicted in Fig. 4.

A. Plate Segmentation

We use a projection-based technique to separate the plates. However, if the plates are rotated, this segmentation will not be accurate. Therefore, the image must be rotated and corrected before segmentation. To this end, we pre-segment the image to find the rotation angle. The image is then rotated and corrected, and the segmentation is repeated.

To separate the plates from background, the projection of the gray-scaled image along the horizontal lines is employed to find the location of the upper edges of the plates. As shown in Fig. 5, the projection resembles a step function at the vicinity of the plate's upper edge. The mean value of the projection in this vicinity is used as a threshold to locate the edge. Since all images are taken at the same conditions, the plates have a constant size. Therefore, after finding the upper edge, we locate the lower edge and remove the upper and lower background. Similarly, using the vertical projection, we remove the left and right background of the resulting image (see Fig. 6).

A vertical middle line of the resulting image divides the image into two plates. At this point, we can check if the image is rotated (as shown in Fig. 6 the plates are slightly rotated). In general, the correct orientation of an object in the image can be estimated from some *directional* regions (like edges) of the object that always reflect the correct orientation. For example, in the current application, edges of the plate frames always indicate the orientation of the plates. Using a subimage of this directional region and the Radon transform orientation estimation technique proposed in [20], the orientation can be estimated. In this technique, the Radon transform of the subimage is first calculated, which is in fact the projections of the images in different directions. The variance of each projection is calculated to create a 1-D function of the orientation. The projection of the image along the principal direction has local maximum variations. Therefore, locating the sharpest change in the generated 1-D signal results in finding the principal direction of the image.

In the current application, this subimage is taken from where the upper right corner of the left plate meets the upper left corner of the right plate (see Fig. 7). This subimage contains edges of the two plates. Following orientation estimation, if the image is rotated, we correct the orientation of the original image and then repeat the previous steps to segment the plates. Fig. 8(a) shows the plate segmentation result for the X-Gal plate shown in Fig. 1 (a).

B. Spot Segmentation

Following segmentation of the plates, we segment the spots. First, the edges of each plate's frame are removed by knowledge of their size. Fig. 8 (b) shows the result. Next, a grid is placed on the image to separate the spots by using the horizontal and vertical projections of the image. As shown in Fig. 9, we calculate the horizontal and vertical projections of the segmented plates. After smoothing the projections, we can locate the separating grid lines by calculating the local minima using the derivative of the signal and placing grid lines at the location of local minima. However, since the spots may not have uniform intensities, we

may miss or have extra local minima of the projection after smoothing. This happens when spots in one column or row have large differences in intensity compared with the neighboring spots or when the spots have irregular shapes. Furthermore, due to the same problem, we may not have uniform spacing between the local minima. This problem has been stated in [17] and a rule-based algorithm has been proposed to locate the correct grid lines. The algorithm searches and tests all the local minima and distances between them and eliminates the mis-positioned lines and adds new lines wherever a line is missed. The proposed algorithm is complex to implement and also does not correct the grid line locations (the lines are either added or removed). We propose the following method which is simpler and more accurate to locate the grid lines.

Suppose the correct locations of grid lines are $X_k = \mu + kd$, $k = 1, 2, \dots, N-1$ where d is the distance between centers of the two adjacent spots and N is the number of spots in the horizontal or vertical direction. The local minima of the projection are estimations of X_k 's. Suppose we denote them by random variables with a normal distribution, i.e., $\hat{X}_k = \mu + kd + \sigma_x \eta_k$ where $\eta_k \sim N(0,1)$ is a random variable with the standard normal distribution. The goal is to estimate μ , since by knowing this parameter we are able to find the correct locations of the grid lines using $X_k = \mu + kd$. We use maximum likelihood estimation (MLE) to estimate μ , i.e.,

$$\hat{\mu} = \arg \text{Max} \left\{ p(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_{N-1} | \mu) \right\} \quad (1)$$

Knowing that

$$p(\hat{X}_k | \mu) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left(-\frac{1}{2} \left(\frac{\hat{X}_k - kd - \mu}{\sigma_x} \right)^2 \right) \quad (2)$$

and assuming X_k 's are independent, we have:

$$\begin{aligned} p(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_{N-1} | \mu) &= \prod_{k=1}^{N-1} p(\hat{X}_k | \mu) \\ &= \prod_{k=1}^{N-1} \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left(-\frac{1}{2} \left(\frac{\hat{X}_k - kd - \mu}{\sigma_x} \right)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma_x} \right)^{N-1} \exp \left(-\frac{1}{2} \sum_{k=1}^{N-1} \left(\frac{\hat{X}_k - kd - \mu}{\sigma_x} \right)^2 \right) \end{aligned} \quad (3)$$

Therefore,

$$\hat{\mu} = \arg \text{Min} \left\{ \sum_{k=1}^{N-1} \left(\frac{\hat{X}_k - kd - \mu}{\sigma_x} \right)^2 \right\} = \arg \text{Min} \left\{ \sum_{k=1}^{N-1} (\hat{X}_k - kd - \mu)^2 \right\} \quad (4)$$

By setting the first derivative with respect to μ equal to zero we get:

$$\hat{\mu} = \frac{1}{N-1} \sum_{k=1}^{N-1} \hat{X}_k - \frac{N}{2}d \quad (5)$$

The above provides an MLE estimate for μ if all values of $\hat{X}_k$ are available (i.e., for $k = 1, 2, \dots, N-1$). However, as explained before, we may miss or have extra local minima due to the differences in spots brightness. To overcome this problem, we note that we still can use (4) for the available $\hat{X}_k$ values. However, since in practice, we have a number of $\hat{X}_k$ values without knowing their indices (i.e., k values), the value in (4) cannot be directly calculated. To resolve this problem, we assume $\hat{X}_k \in (kd + \mu - d/2, kd + \mu + d/2)$, which is practically accurate. Then (4) can be rewritten as follows for the available $\hat{X}_k$'s:

$$\begin{aligned} \hat{\mu} &= \arg \text{Min} \left\{ \sum_k (\hat{X}_k - kd - \mu)^2 S(\hat{X}_k - kd - \mu) \right\} \\ &= \arg \text{Min} \left\{ \int_{-\infty}^{\infty} \sum_{k=1}^{N-1} (x - kd - \mu)^2 S(x - kd - \mu) \sum_m \delta(x - \hat{X}_m) dx \right\} \end{aligned} \quad (6)$$

where

$$S(x) = \begin{cases} 1 & -d/2 < x < d/2 \\ 0 & \text{o.w.} \end{cases} \quad (7)$$

If the correlation of two functions is denoted by $\text{corr}(\cdot, \cdot)$, then (6) can be written as:

$$\hat{\mu} = \arg \text{Min} \left\{ \text{corr} \left(\sum_{k=1}^{N-1} (x - kd - \mu)^2 S(x - kd - \mu), \sum_m \delta(x - \hat{X}_m) \right) \right\} \quad (8)$$

Notice that in (8) we do not need to know which $\hat{X}_k$'s are available. Therefore, μ can be estimated by numerically solving (8), which can also be interpreted as a template matching. Define the template signal

$$f_T(x) = - \sum_{k=1}^{N-1} (x - kd)^2 S(x - kd) \quad (9)$$

and

$$l(x) = \sum_k \delta(x - \hat{X}_k) \quad (10)$$

Then (8) implies $\hat{\mu} = \arg \text{Max} \{ \text{corr}(f_T(x - \mu), l(x)) \}$. The values of correlation are calculated for different values of μ , then the maximum of correlation is calculated to estimate μ . In practice, the correlation function can be approximated by $\sum_x f_T(x - \mu) l(x)$, where $l(x)$ is a train of discrete delta functions, i.e. $l(x) = \sum_k \delta(x - \hat{X}_k)$, where

$$\delta[n] = \begin{cases} 1, & n=0 \\ 0, & \text{o.w.} \end{cases} \quad (11)$$

Fig. 10 (a) shows an example of $l(x)$ and $f_T(x)$. The parameter μ is estimated by finding where these two signals have maximum match. Application of this technique to the image in Fig. 9 generated the grid lines shown in Fig. 11.

After locating the grid lines, we locate the spots. For each block in Fig. 11, we calculate the center of mass and consider a circle with a constant radius. We use the intensity information inside the circle for feature extraction. Although the actual spot shape usually is not strictly circular, we constrained the spot shape to be circular to ensure that the spot extraction procedure is robust (see Fig. 1 (b) for some examples of non-circular spots). All the information for scoring the spot can be found in this circular region with a constant radius. The final segmentation of the spots on the X-Gal and growth assay plates is presented in Fig. 12. The estimated location of each spot has been shown by a circle around it.

C. Illumination Adjustment

Ideally, to achieve a robust classification, the images should be taken under exactly the same conditions. However, in practice, the images have different illuminations. Furthermore, the illumination of the spots inside each plate may vary based on the lighting conditions. In addition, the apparent illumination can differ from plate to plate due to unavoidable differences in the media darkness. Therefore, before calculating the features, we need to adjust the illumination.

To locally adjust the illumination of the spots, we use the intensity information of the background in which the spots are located. This is because the background has a uniform color within each plate and therefore can provide us with the illumination pattern. Since the location of each spot is known, by excluding the spots (circular regions) from each plate, we get the background pixels. However, since the spots have different sizes, we may still have pixels of spots in the background. To remove these pixels, we map all pixels into a fixed color map, whose colors are already classified into two groups of background and non-background (yeast growth areas). If the color into which a pixel is mapped belongs to background, then the pixel is classified as background, otherwise, it is classified as non-background. In this way non-background pixels are found and removed. Samples of background and yeast growth areas from a number of plate image samples (10 in our experiments) are used as training sets to classify the colors in the color map into the background and non-background classes. A color is classified as background if it has higher occurrence in the background areas. Otherwise, it is a non-background color. Finding the background pixels, we fit a polynomial surface of order 3 to their intensities. The model and its order are selected empirically. Using the produced surface, the intensities of pixels are adjusted by dividing the intensity of each pixel by the value of the surface at that point and multiplying by a desired fixed value. For visualization purposes, we chose the value of 110 in our experiments, which is the intensity mean value of the background pixels from a number of plate image samples. The surface is smooth enough to avoid following the background artifacts and noise, and flexible enough to follow the illumination variations. Note that in growth assay plates, the intensity of the background pixels around the high score spots may be higher, which causes sudden changes in the background intensity. However, the intensity changes due to the illumination problems are smoother, and that is why we use a smooth surface to correct the illumination.

III. Feature Extraction

As mentioned earlier, on the X-Gal plates the color determines the score, whereas on the growth assay plates, intensity variations mainly specify the score. Therefore, we employ different techniques to extract features from the spots on each plate.

A. X-Gal Plates

Since in X-Gal plates, the scores depend on the color variations, we may use the color histogram of the image as a feature vector. To calculate the histogram, the color image is

mapped onto a predefined and fixed color map. This map has been optimized and fixed using training samples of plate images. However, since the color map is usually large (689 colors in our experiments), it is not appropriate to directly use it as a feature vector. To reduce the feature space dimension, we employ an optimal linear transformation. Linear transformation is desirable since it preserves the geometric properties between clusters. This problem can be formulated as follows.

Assume there is a collection of data points within the subspace of $V \subset \mathbb{R}^n$ that can be classified into M groups. A linear transformation $T: V \rightarrow W$ with $W \subset \mathbb{R}^m$ and $m < n$ is desirable such that for the mapped points in the new subspace W , the ratio of inter class distance (D_{inter}) to the intra class distance (D_{intra}) is maximized. Assuming we would like to map the class centers to predefined locations, this problem can be considered as the following constrained optimization problem:

Maximize: $\frac{D_{\text{inter}}}{D_{\text{intra}}}$, subject to $\mathbf{c}_j = T \bar{\mathbf{v}}_j$, $1 \leq j \leq M$, where T is the transformation matrix and $\mathbf{c}_j$ is the target position for the average vector of the j th group $\bar{\mathbf{v}}_j$, defined by

$$\bar{\mathbf{v}}_j = \frac{1}{N_c(j)} \sum_{l=1}^{N_c(j)} \mathbf{v}_j^l \quad (12)$$

with $\mathbf{v}_j^l$ being the l th data point in the j th group, and $N_c(j)$ the number of data points in the j th group. The D_{inter} reflects the average distance between different data groups and can be defined by:

$$D_{\text{inter}}^2 = \frac{2}{M(M-1)} \sum_{j=1}^M \sum_{i=j+1}^M \|T \bar{\mathbf{v}}_j - T \bar{\mathbf{v}}_i\|^2 \quad (13)$$

The D_{intra} reflects the average variance of the data point in each group and can be defined by:

$$D_{\text{intra}}^2 = \frac{1}{M \sum_{j=1}^M N_c(j)} \sum_{j=1}^M \sum_{i=1}^{N_c(j)} \|T \bar{\mathbf{v}}_j^i - T \bar{\mathbf{v}}_j\|^2 \quad (14)$$

A special case can be defined for $M = m$, i.e., when the dimension of the target feature space is equal to the number of groups. If we further decide to assign the target position of $\mathbf{c}_j$ to be on the axes of the new subspace, i.e., $\mathbf{c}_j = [0, 0, \dots, c_j, \dots, 0]^T$, then

$$D_{\text{inter}}^2 = \frac{2}{M} \sum_{j=1}^M c_j^2 \quad (15)$$

which is a constant. The problem then reduces to the minimization of D_{intra} . This problem has been solved in [21] assuming a Gaussian model for the data points of each group. The optimal solution is given by

$$T_{m \times n} = \begin{bmatrix} \bar{v}_1 - \bar{v}_1^p \\ \bar{v}_2 - \bar{v}_2^p \\ \dots \\ \bar{v}_m - \bar{v}_m^p \end{bmatrix} \quad (16)$$

where $\bar{v}_i^p$ is the projection of $\bar{v}_i$ onto the subspace spanned by $\{\bar{v}_k, k = 1, 2, \dots, m, k \neq i\}$. This means to calculate the rows of the matrix T , we need to calculate the average of data points in each group and make them orthogonal using the Gram-Schmidt orthogonalization procedure. The features in the new subspace are the inner products of these vectors with the original feature vector, i.e.,

$$w_{m \times 1} = T_{m \times n} v_{n \times 1} = \begin{bmatrix} (\bar{v}_1 - \bar{v}_1^p) \cdot v_{n \times 1} \\ (\bar{v}_2 - \bar{v}_2^p) \cdot v_{n \times 1} \\ \dots \\ (\bar{v}_m - \bar{v}_m^p) \cdot v_{n \times 1} \end{bmatrix} \quad (17)$$

where $v_{n \times 1}$ and $w_{m \times 1}$ are the feature vectors in the original and new feature spaces, respectively.

In the problem of scoring X-Gal plates, we have seven groups of $S_k, k \in \{0, 1, 2, 3, 4, 5, b\}$ where k indicates either a spot having scores $\{0, 1, 2, 3, 4, 5\}$ or background (indicated by b). Thus $m = 7$ and $n = 689$, which shows a significant reduction in the feature space dimension.

B. Growth Assay Plates

For the growth assay plates, the score is specified based on the pattern of intensity variations. In practice, spots with the same score may have slightly different intensity values even after illumination adjustment, which is due to unavoidable variations in the media. Therefore, in addition to illumination adjustment explained in II-C, we need to individually normalize the intensities of each spot. The mean value of the spot is not a suitable choice for normalization, since the growth area (the bright areas of the spot) may vary from spot to spot and thus change the mean value. To find a proper value, we first remove the background pixels from the spot and then find the mean intensity of the remaining pixels. The resulting value can be used to normalize all pixel intensities before feature extraction. To find the background pixels, we use the method of color map explained in II-C. If all the pixels in a spot belong to the background, then no normalization is performed.

After normalization, texture features are extracted from a circular region around the spot. We calculate the wavelet transform of each spot up to three levels of decomposition using Daubechies wavelet basis with length 5 (db5) [22]. The choice of db5 is made after applying wavelet bases db3, db4, db5, and db6, and choosing the one that produces highest correct classification rate. Note that other wavelet bases (not studied in this work) may lead to higher correct classification rates. However, wavelet bases with larger lengths are not suitable, since the spots images are small (around 50×50). Wavelet bases with larger lengths have longer and smoother $h(n)$ sequences and therefore are not suitable for small images (they may not capture sharp intensity variations).

¹We have shown the non-background and background spots of score zero with “0” and “b” respectively. This separation is because these spots look different.

The resulting wavelet transform has 10 subbands. We ignore the three highest resolution subbands, since they are dominated by noise. For the remaining 7 subbands, from each subband I , we calculate the following feature:

$$e^{(l,s)} = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N |I_{i,j}^{(l,s)}|^{\frac{1}{2}} \quad (18)$$

where M and N are the dimensions of the subband, l is the decomposition level, and s represents the subband number within the decomposition level l . This feature shows the amount of signal energy in a specific resolution and is different from the traditional energy feature in the sense that instead of using the squared values we use the square root of the

absolute values of $I_{i,j}^{(l,s)}$. The energy feature has been shown to be useful for texture analysis [23],[24]. In addition to the above features, we calculate color features using (17) assuming classes of scores $\{0,1,2,3\}$. Although color is not the main source of information in scoring growth assay spots, due to slight variations of color in different plates, the above features help to obtain more accurate scores. The total number of features is $7+4=11$ for the growth spots.

IV. Scoring

The simplest method for scoring using the extracted features is k -nearest neighbors [25]. However, due to the diversity of the spots' appearance and high number of classes in our application, the number of training samples must be high, which causes the classifier to be rather slow. To avoid this problem, we propose using ANN to classify the spots. Although this may not be the optimal choice for this application, it provides acceptable performance with reasonable generalization. For X-Gal spots, we use a multilayer perceptron with 7 inputs, 1 output, 3 layers and 5, 3, and 1 neurons in the first, second, and third layers, respectively. The first two layers use a tan-sigmoid transfer function whereas the last layer uses a saturated linear transfer function. Since the output of the network is in the interval $[0,1]$, we multiplied it by 5 to get values in the range $[0,5]$. To score a spot, the features are calculated and fed to the neural network, and the output of the network is multiplied by 5, then rounded to the nearest integer. Therefore, the final output values are $\{0,1,2,3,4,5\}$. The network size is selected by trying different sizes and choosing the smallest in which the test error starts increasing after reaching its minimum, with increasing the number of epochs.

For the growth assay spots, we use a similar ANN with the differences that it has 11 inputs and the output is multiplied by 3 to generate values of $\{0,1,2,3\}$. The network size is selected in the same way that is done for X-Gal spots.

V. Experimental Results

In our experiments, we evaluated the segmentation and scoring units separately. To evaluate the performance of the proposed segmentation technique, we used a data set of 3,000 plate images (see some samples in Fig. 1). In only one case, the plate segmentation failed. In only three cases, the spots segmentation failed. Failures in these cases were due to either anomalous imaging or plates. The remaining 2,996 images were segmented correctly (verified through inspection)¹.

To evaluate the performance of the proposed scoring techniques, we selected a subset of 60 images from the above data set. The images were selected as follows. Since in practice the

¹Note that manual segmentation of these images in order to compare with automatic segmentation was impracticable.

majority of the X-Gal spots have the score of 0, we selected the images with the highest number of blue spots (high scores) in their X-Gal plates. Thus, we have $60 \times 96 = 5760$ spots from X-Gal plates, and $60 \times 96 = 5760$ spots from growth assay plates. All spots were then scored manually by an expert. The numbers of spots within each group of scores for X-Gal and growth assay plates are presented in Tables I and II. Spots of score 0 in X-Gal plates include two groups of background and non-background spots as explained in Section I. The training, validation, and test sets for X-Gal plates were prepared in the following way.

For the growth assay plates the data sets were prepared in the similar way. For the X-Gal plates, the data set was randomly divided into two halves, one for training and validation, and the other for testing. Thus, we used 2484 (2364+120), 221, 103, 51, 12, and 8 spots, respectively with scores 0 (non-background+background), 1, 2, 3, 4, and 5 for training and validation, and the rest for testing. Here, we had the problem of unbalanced data. Repeating the data points could be an option but it might not provide a good generalization for the ANN classifier. To resolve this problem, based on the assumption that the data sets of each group were convex, we added new samples to each class by interpolation. This assumption will be assessed later by testing the generalization of the ANN classifier.

New samples were created as follows. Two samples were randomly selected from one class, and a point was randomly picked between them using linear interpolation. Then a Gaussian noise was added to the sample with an appropriate standard deviation in order to make the samples more realistic and uniformly distributed. The new added sample participated in the next sample generations. This procedure produces more samples in the denser areas, which reduces the effect of outliers. New samples were added to the groups of scores 0 (background), 1, 2, 3, 4, and 5, until the number of data points in each group was equal to the number of data points in the group of score 0 (non-background) which had the maximum number of data points (i.e., 2364). Note that we considered background and non-background spots of score zero as two separate sets for sample generation, since they had different feature properties and could not be considered in one convex set. Therefore, for training and validation, we had the total number of 2364×7 data points from the X-Gal plates.

Fig. 13 shows the scatter plots of the extracted features from spots of scores 0, 1 and 2. This shows that these data points have clusters with regular shapes. To evaluate the performance of the classification, we randomly split the above data set (which included generated and non-generated data points) into two sets of training and validation (50% each). The neural network was trained on the training set and the performance was monitored on the validation set. Once the network started overfitting, learning was stopped and the final classification accuracy was evaluated on the test set [26]. We used a back-propagation technique to train the network. Table III shows the correct classification results for training, validation, and test (which includes original samples not used in training and validation) sets for the X-Gal plates. As shown, the overall accuracies for the training and validation sets are 98.6% and 98.3% respectively. Accuracies for the test set are between 66.7% and 97.1%. The overall accuracy for the test set is 95.4%. As shown, the accuracy is decreased for higher scores. This can be attributed to the low number of original samples for these scores. Although the interpolated samples are added to make the data sets balanced, with the low number of samples we may not get an accurate distribution of the samples in the feature space. Table IV shows the confusion map of the classification result for the test set of X-Gal plates. As shown, the majority of the misclassifications are due to an error of one score unit.

For the growth assay plates, a similar procedure was employed to create training, validation, and test sets. We used 1808, 386, 282, and 403 spots respectively with scores 0, 1, 2, and 3 for training and validation, and the rest for testing. Sample generation for scores 1, 2, and 3 of the growth assay plates continued until the number of samples in each class was equal to

1808, which was the number of zero score spots. Therefore, a number of 1808×4 spots were created for training and validation. Table V shows the correct classification results for training, validation, and test sets for the growth assay plates. The overall accuracies for the training and validation sets are 93.0% and 91.0% respectively. As shown, the accuracies for the test set are between 70.0% and 92.9%. As shown, the accuracies for scores 1 and 2 are low. The main reason for the low accuracies for growth assay plates is the imperfect data preparation and photography conditions, which result in relatively large variations in the color and intensities of the spots. Another factor can be imperfect manual scoring. Note that this analysis compares the automatic scores with scores obtained by manual inspection, which may include subjective bias and inconsistencies.

We observe that the error is higher for the growth assay plates compared with the X-Gal plates. This is due to the large variability of the spot intensities and appearances, which causes uncertainty in both manual and automatic scoring. Using higher quality images may increase the accuracy.

The proposed method has limitations for scoring uncommon spots. Fig. 14(a) shows a color-changed area between two adjacent spots in an X-Gal plate, which creates ambiguity in scoring both spots. The system assigns some part of that area to each spot and scores the spots based on the extracted features from those areas. In general, color changes in one spot may affect the adjacent spots. However, this rarely happens in practice. Fig. 14(b) shows a spot on the growth assay plate with the growth area on the boundary (out of a normal spot area). This spot has score 2. However, the system removes the growth area outside the boundary and assigns a lower score to that spot. Fig. 14(c) shows a fungus-contaminated area, which covers some spots. No scoring should be done in these cases. However, the segmentation does not fail due to the robustness of the techniques described in Section II-B to the anomalies, and the spots are scored. Fortunately, the majority of these plates are detected when photographing them and are not used for scoring. Fig. 14(d) shows a spot with a small area of color change. Since most of the pixels in this spot belong to the background, the estimated score may be lower than the real score. In Fig. 14(e) the growth area of one spot has leaked to the neighboring spot regions, which affects the features extracted from the adjacent spots and consequently their scores. In general, since the image acquisition is done manually, most of these anomalies (especially the fungus-contaminated plates) can be detected by a quick visual inspection of the plates before image acquisition.

VI. Conclusion

An automated method was presented to segment and score spots of yeast growing on the X-Gal and growth assay plates. Color features using a color histogram technique and wavelet texture features were extracted respectively from the spots on each plate. Artificial neural networks were used for scoring the spots using the extracted features. Our experimental results showed the reliability of the system for segmenting and scoring the spots.

To have more accurate scores for X-Gal plates, more features may be extracted from each spot. The easiest way to do this is to map the introduced color histogram to a larger feature vector. For the growth assay plates, images with higher quality and more accurate manual scores might improve the results.

Our system should profoundly enhance the efficiency of scoring yeast two-hybrid assays. In its first practical test, we used it to score 3,000 images from a high throughput two-hybrid screen. The scoring took 20 person hours, including the time to upload images, process, and download results. By comparison, the same 3,000 images were scored manually, which took about 1,200 person hours. In practice the plates must be scored twice when scoring

manually. Assuming we score only once, the program results in approximately 30-fold increase in efficiency. Moreover, by proper modifications and refinements, the system could be adapted to a variety of other yeast-based assays for characterizing genes and biological systems. Although the 96-well plate is now a standard for high throughput assays, other research groups may use plates with different sizes (e.g. 10×10 instead of 8×12 arrays). Our system could be easily modified to accommodate different plates. Assuming all images are taken in similar lighting and magnification conditions, with the knowledge of the sizes of the plate, array, and spots, and the distances between centers of spots, the program could be adjusted for the new plates and used to quantify growth and color in a variety of settings.

Acknowledgments

This work was supported in part by the National Institute of Health under Grant R01-EB002450, Grant HG001536, and Grant RR18327 from the National Center for Research Resources, a component of the National Institutes of Health.

We would like to thank members of the Finley laboratory for helping with data acquisition and system evaluation, and Mostafa Ghannad-Rezaie for the comments about data classification.

References

1. Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S. Protein analysis on a proteomic scale. *Nature*. 422(6928):208–215. 2003. [PubMed: 12634794]
2. Fields S, Kohara Y, Lockhart DJ. Functional genomics. *Proc Natl Acad Sci U S A*. 1999; 96(16): 8825–8826. [PubMed: 10430853]
3. Hughes TR, Robinson MD, Mitsakakis N, Johnston M. The promise of functional genomics: completing the encyclopedia of a cell. *Curr Opin Microbiol*. 7(5):546–554. 2004. [PubMed: 15451511]
4. Brent R, Finley RL Jr. Understanding gene and allele function with two-hybrid methods. *Annu Rev Genet*. 31:663–704. 1997. [PubMed: 9442911]
5. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature*. 340(6230): 245–246. 1989. [PubMed: 2547163]
6. Weiss A, Delproposto J, Giroux CN. High-throughput phenotypic profiling of gene-environment interactions by quantitative growth curve analysis in *Saccharomyces cerevisiae*. *Anal Biochem*. 327(1):23–34. 2004. [PubMed: 15033507]
7. Hughes TR. Yeast and drug discovery. *Funct Integr Genomics*. 2(4–5):199–211. 2002. [PubMed: 12192593]
8. Butcher RA, Schreiber SL. Identification of Ald6p as the target of a class of small-molecule suppressors of FK506 and their use in network dissection. *Proc Natl Acad Sci U S A*. 2004; 101(21):7868–7873. [PubMed: 15146068]
9. Stanyon CA, Liu G, Mangiola BA, Patel N, Giot L, Kuang B, Zhang H, Zhong J, Finley RL Jr. A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol*. 5(12):R96, 2004. [PubMed: 15575970]
10. Uetz P, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*. 403(6770):623–627. 2000. [PubMed: 10688190]
11. Estojak J, Brent R, Golemis EA. Correlation of two-hybrid affinity data with in vitro measurements. *Mol Cell Biol*. 15(10):5820–5829. 1995. [PubMed: 7565735]
12. Zhong J, Zhang H, Stanyon CA, Tromp G, Finley RL Jr. A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating. *Genome Res*. 13(12):2691–2699. 2003. [PubMed: 14613974]
13. Nagarajan R, Peterson CA. Identifying spots in microarray images. *IEEE Trans Nanobiosci*. June; 2002 1(2):78–84.
14. Nagarajan R. Intensity-based segmentation of microarray images. *IEEE Trans Med Imag*. July; 2003 22(7):882–889.

15. Katzer M, Kummert F, Sagerer G. Methods for automatic microarray image segmentation. *IEEE Trans Nanobiosci.* Dec; 2003 2(4):202–214.
16. Liew AWC, Yan H, Yang M. Robust adaptive spot segmentation of DNA microarray images. *Pattern Recognition.* May; 2003 36(5):1251–1254.
17. Hirata R Jr, Barrera J, Hashimoto RF, Dantas DO, Esteves GH. Segmentation of microarray images by mathematical morphology. *Real-Time Imaging.* Dec; 2002 8(6):491–505.
18. Damiance APG Jr, Zhao L, Carvalho ACPLF. A dynamical model with adaptive pixel moving for microarray images segmentation. *Real-Time Imaging.* August; 2004 10(4):189–195.
19. Trattner S, Greenspan H, Tepper G, Abboud S. Automatic identification of bacterial types using statistical imaging methods. *IEEE Trans Med Imag.* July; 2004 23(7):807–820.
20. Jafari-Khouzani K, Soltanian-Zadeh H. Radon Transform Orientation Estimation for Rotation Invariant Texture Analysis. *IEEE Trans Pattern Anal Mach Intell.* June; 2005 27(6):1004–1008. [PubMed: 15945146]
21. Soltanian-Zadeh H, Windham JP, Peck DJ. Optimal linear transformation for MRI feature extraction. *IEEE Trans Med Imag.* Dec; 1996 15(6):749–767.
22. Daubechies, I. *Ten Lectures on Wavelets.* Philadelphia, PA: Soc. Ind. Appl. Math.; 1992.
23. Laine A, Fan J. Texture classification by wavelet packet signatures. *IEEE Trans Pattern Anal Mach Intell.* 15(11):1186–1191. 1993.
24. Sam-Deuk K, Udpa S. Texture classification using rotated wavelet filters. *IEEE Trans Syst, Man, Cybern, A, Syst, Humans.* Nov; 2000 30(6):847–852.
25. Gose, E.; Johnsonbaugh, R.; Jost, S. *Pattern Recognition and Image Analysis.* Englewood Cliffs, NJ: Prentice-Hall; 1996.
26. Bishop, CM. *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press; 1995.

Fig. 1. Sample images containing X-Gal (left) and growth assay (right) plates. The spots in (a) have circular shapes, whereas some spots in (b) have irregular shapes.

Fig. 2. Spot samples of (a) X-Gal plates, and (b) growth assay plates with their scores. As shown, in X-Gal plates the scores depend on color information, and in growth assay plates the scores depend on intensity and texture information of the spot.

Fig. 3.
Range of intensities for growth assay spots having scores 0–3.

Fig. 4. Block diagram of the proposed system. As shown, the image is first segmented to separate X-Gal and growth assay plates. Each plate is then segmented separately using a projection-based technique. Color features are extracted from the X-Gal spots, and texture and color features are extracted from the growth assay spots. The spots are scored using artificial neural networks based on the extracted features.

Fig. 5.
Horizontal projection of the image to find the upper edge of the plates.

Fig. 6. Vertical projection of the image to find the left and right edges of the plates. As shown, the plates are slightly rotated.

Fig. 7. Orientation adjustment is done using the specified location of the image. The Radon transform of this area can be used to estimate the orientation of the plates as described in [20].

Fig. 8.

(a) The segmentation output for X-Gal plate before removing the frame edges, (b) after removing the frame edges.

Fig. 9.
The horizontal and vertical projections of the plate image to locate the spots.

Fig. 10.

Locating the grid lines. (a) Local minima of the projection, (b) a template signal to locate accurately the grid lines. After calculating the correlation of (a) with shifts of (b), the local maximum of the result is found in a suitable interval, to locate the grid lines.

Fig. 11.
Overlaying the separating grid lines on the X-Gal plate shown in Fig. 8 (b).

Fig. 12.
The final spot segmentation for (a) X-Gal plate, and (b) growth assay plate.

Fig. 13.

Scatter plots of some selected features for X-Gal spots, where $[f_1, \dots, f_7]$ is the feature vector. (a) and (b) data points of score 0, (c) and (d) data points of score 1, and (e) and (f) data points of score 2. The scatter plots in one column correspond to the same features. The scatter plots of the other scores were not shown since they had few samples, which may not visualize distribution of the samples clearly.

Fig. 14. Limitations of the proposed scoring system. (a) The color-changed area is between two spots, (b) the growth area is on the spot boundary, (c) fungus contamination, (d) a spot with a small color-changed area, (e) the growth area of one spot has leaked to the neighboring spots.

TABLE I

The number of data points within each group of scores in X-Gal plates.

0						
Score	Non-background	Background	1	2	3	4 5
Number of spots	4727	240	443	206	103	24 17

TABLE II

The number of data points within each group of scores in growth assay plates.

Score	0	1	2	3
Number of spots	3616	772	565	807

TABLE III
Correct classification percentages for training, validation, and test sets for X-Gal plates.

Data set	Score					Overall	
	0	1	2	3	4		5
Training	98.1	97.2	97.2	99.4	99.9	100	98.6
Validation	98.2	96.5	96.5	98.8	99.9	99.8	98.3
Test	97.1	86.9	85.4	78.8	83.3	66.7	95.4

Confusion map of the classification result for the test set of X-Gal plates. The numbers on each row correspond to spots of one class and show how many times they have been classified to different classes.

TABLE IV

Score	0	1	2	3	4	5
0	2411	71	1	0	0	0
1	21	193	8	0	0	0
2	0	5	88	10	0	0
3	0	1	6	41	4	0
4	0	0	0	2	10	0
5	0	0	0	0	3	6

TABLE V
Correct classification percentages for training, validation, and test sets for growth assay plates.

Data set	Score			
	0	1	2	3
Overall				
Training	95.9	94.6	88.7	92.7
Validation	93.5	92.1	86.3	92.0
Test	92.9	78.8	70.0	88.9