Spot Counting on Fluorescence In Situ Hybridization in Suspension Images using Gaussian Mixture Model

Sijia Liu[1], Ruhan Sa[1], Orla Maguire[2], Hans Minderman[2], and Vipin Chaudhary[1]

[1]Department of Computer Science and Engineering, University at Buffalo
[2]Flow and Image Cytometry Facility, Roswell Park Cancer Institute

ABSTRACT

Cytogenetic abnormalities are important diagnostic and prognostic criteria for acute myeloid leukemia (AML). A flow cytometry-based imaging approach for FISH in suspension (FISH-IS) was established that enables the automated analysis of several log-magnitude higher number of cells compared to the microscopy-based approaches. The rotational positioning can occur leading to discordance between spot count. As a solution of counting error from overlapping spots, in this study, a Gaussian Mixture Model based classification method is proposed. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) of GMM are used as global image features of this classification method. Via Random Forest classifier, the result shows that the proposed method is able to detect closely overlapping spots which cannot be separated by existing image segmentation based spot detection methods. The experiment results show that by the proposed method we can obtain a significant improvement in spot counting accuracy.

Keywords: Fluorescent In Situ Hybridization Image, Gaussian Mixture Model, Spot Counting, Pattern Recognition

1. INTRODUCTION

Cytogenetic abnormalities are important diagnostic and prognostic criteria for acute myeloid leukemia (AML). Karyotyping and Fluorescent In Situ Hybridization, (FISH), are the conventional methods by which these abnormalities are detected. A flow cytometry-based imaging approach for FISH in suspension (FISH-IS) was established that enables the automated analysis of several log-magnitude higher number of cells compared to the microscopy-based approaches[1].

The ImageStream platform captures up to 12 simultaneous, spectrally separated images from each cell at rates up to 1,000 cells/sec. With the ImageStream approach, the collected images are 2-D projections of a 3-D cell in suspension. Thus, the rotational positioning of a cell relative to the camera can lead to an inadvertent overlap of spot-like hybridization sites. In addition, segregation of a single hybridization site into two or three derivate smaller and dimmer spots can occur; all leading to a discordance between spot count and ploidy. The correlation of a spot count in a cell with the simultaneous measurement of fluorescence intensity in the same cell allows the necessary correction for overlapping spots.

One of the challenges in accurately detecting the spots is that in a large fraction of FISH-IS image dataset, not all the spots are visible due to the issue discribed above, especially in disomy (two-spot) and trisomy (three-spot) images. However, based on human viewing of the images, there are many instances where one or two visible spots are truly trisomy and one spot is a disomy. Thus, whether a practical classification system based on the FISH-IS images can be found is critical to the effectiveness of proposed FISH-IS method.

The aim of this study is to design an image classification approach that automatically, which means it involves no human labelling on the spot position, and accurately obtained the expected spot counts without prior knowledge of the spot quantification.

In this paper, we proposed an image classification method within FISH-IS dataset to classify monosomy, disomy and trisomy images according to proposed features extracted from Gaussian Mixture Model density estimation.

Medical Imaging 2015: Image Processing, edited by Sébastien Ourselin, Martin A. Styner, Proc. of SPIE Vol. 9413, 94132H · © 2015 SPIE · CCC code: 1605-7422/15/$18 · doi: 10.1117/12.2081026

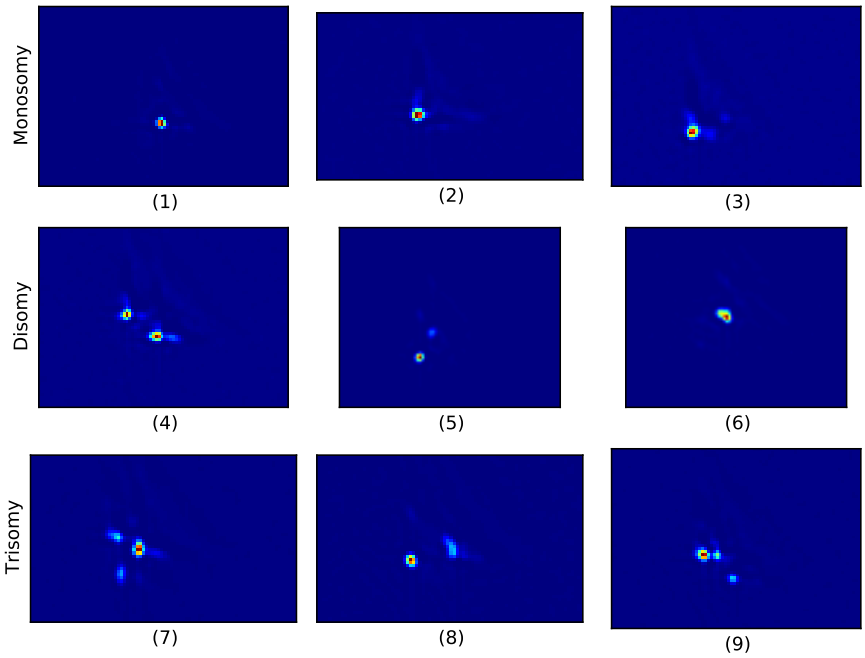

Figure 1. Samples of FISH-IS images. Figure 1(1)-(3) are monosomy (one-spot), (4)-(6) disomy (two-spot) and (7)-(9) trisomy (three-spot) images. Figure (5) is disomy with one clear spot and one dimmed spot that results into two spots but will likely be marked as one spot. A similar pattern appears in (7). Figure (6) is an example of two close spots overlapping as one. Although (8) is provided as a trisomy image which should contains three spots, only one strong spot and another dimmed spot can be found. This figure is challenging for human experts to diagnose. In (9), two close spots have different range of maximum intensity values. Except for monosomy images, most of the existing methods can not achieve high classification accuracy in the disomy and trisomy images.

2. RELATED WORKS

Although there are some existing studies in this field, all existing algorithms have either drawbacks or do not applicable to our problem. A typical FISH image analysis system has two separated steps, nuclei segmentation and spot detection. Solutions are normally closely related to specific dataset, which may have differences in the size of image, pixel intensity ranges, techniques on image formation and numbers of cell nuclei in a single image[2].

Netten et al. [3] proposed an automatic fluorescent dot counting system in interphase cell nuclei. Tophat threshold is used to remove the noise on the background. To exclude the impact of merged spots, a nonlinear Laplacian threshold is performed following the tophat thresholding, which is used to generate the mask image. After the spots detected by thresholding methods, several features based on spot intensity statistics such as maximum intensity, total intensity and average intensity are extracted. The shape of spots is also considered by calculating the eccentricity of each spot, which will be one for circularly symmetric shapes. To detect the overlapping spots, a nearest neighbor classifier is used to classify if the spots are 'overlapping'. The study shown a good result on single spots but only 54% of the overlapping spots can be classified correctly by 6-NN classifier.

Lerner et al.[4] developed a clustering-based method on spot detection. Under the assumption that subsignals of a signal are closer to each other than to subsignals of another signal, the detected subsignals from preprocessing procedures are combined into non-dot-like signals. The global K-means algorithm is adopted, which starts with the mixture mean as the first cluster center and adds incrementally the subsignal that as the next cluster center, together with previous cluster centers. In the feature selection step, a set of 21 features is measured for signals. Besides using Neural Network, naive Bayesian Network is also used to estimate class-conditional densities.

For specific thresholding method, Raimondo et al.[5] applied a modification thresholding selection method of [6] to estimate tophat thresholds for both the red and green channel. The algorithm assumes that there is one

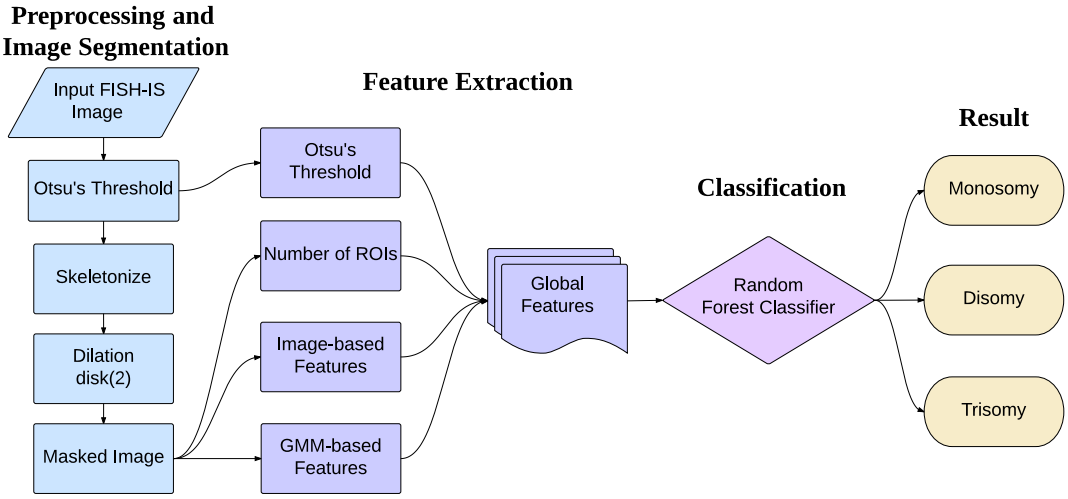

Figure 2. FISH-IS classification procedure workflow

dominant mode in the image histogram. A straight line is drawn from the peak to the high intensity end of the histogram. More precisely, the line starts at the largest frequency bin A and finishes at the first empty bin B of the histogram following the last filled bin. The threshold is selected as the histogram index that maximizes the perpendicular distance between the line and histogram curve. The authors also used a 7×7 window estimate the center position of every spot. This template will eliminate the spot with size not similar with it.

Sagonas et al.[7] described that both red and green spots can be modeled by radial basis function (RBF) with small variance because spots have circular shape. Thus, the spots are represents by a pixel cluster obtained from RBF. Although the method presents a significant improvement compared with [5], the requirement of the expert manually labeling restricts it from further practical applications. Two-dimensional Gaussian function curve fitting methods have also been discussed in [8]. The component fitting of two dimensional Gaussian functions are validated in [8] as a spot quantification method. In addition to these spot detection and classification methods, spot segmentation accuracy are also regarded as a factor by [9]. Besides, the dataset in this study is 3D images, thus all features extracted after the first image segmentation stage are 3D features. Their result has a better ROC curve of spot detection true positive rate than several existing algorithms. Segmentation accuracy evaluated by Tanimoto coefficient and its standard deviation among datasets are calculated and compared to show the method promising.

3. PROPOSED METHOD

In this paper, we discussed an automatic workflow of FISH-IS image classification system. The system comprises three parts: image preprocessing, feature extraction and feature classification. All these components are discussed in detail in this section.

3.1 Preprocessing Methods

Since the FISH-IS images are with low signal-to-noise ratio and have low resolution, several preprocessing methods have to be applied before the feature extraction step. Generally, each image has approximately 80×80 pixels. For better discrimination of further global feature extraction, the input images are neither normalized nor resized.

An arbitrary noise threshold ratio $T_N \in [0, 1)$ is set and the maximum intensity in each image is I_{max}. After taking the maximum value of each image, each pixel with intensity value $I(x, y)$ where $I(x, y) < T_N I_{max}$ is set as zero. Afterwards, the image is skeletonized[10], and each skeleton is regarded as a seed. A morphological dilation operation with a radius 2 disk-shape template is applied. The binary image after these procedures is set as the mask image. The images in the following discussions are all masked images.

3.2 Feature Extraction

Although after preprocessing procedures, most of the single spots can be separated and counted by labeling the remaining Regions of Interest(ROIs), the counting error from ROIs are very large. In our FISH-IS dataset, only 57.2% of the disomy images contain exactly two ROIs. For trisomy images, the situation is even worse: there are only 19.2% trisomy images have three separated spots. The counting error as large as what is described above can not be accepted in a diagnosis method. However, according to our observations into the misclassified images by the preprocessing procedure, it is found that the overlapping is the main cause of the error. As the consequence of this observation, it is necessary to develop an algorithm to eliminate the merge of ROIs due to overlapping spots. In this study, the solution is using classification method to classify each image in the whole dataset instead of directly using the counting of ROIs.

Before performing classification methods within three categories of monosomy, disomy and trisomy from FISH-IS dataset, effective numeric features have to be extracted from two dimensional intensity images. The preprocessing methods will remove noisy pixels and areas before the feature extraction procedure.

3.2.1 Global Features

The global image features are measured from each image and used as the input to the classifier in the classification step: total intensity, average intensity, maximum intensity, second maximum intensity, ratio between maximum intensity and second maximum intensity, Otsu's threshold, number of ROIs, minimum AIC and BIC from GMM.

1. *Total intensity I_T:* The sum of intensity of all pixels in each image.
2. *Average intensity I_{avg}:* The total intensity divided by the total number of pixels.
3. *Maximum intensity I_{max}:* The maximum intensity value of the whole image. That pixel usually locates within the most obvious spot in each image.
4. *Second maximum intensity $I_{max,2}$:* Once the spot containing the maximum intensity value is removed, it is the maximum intesity value of the remaining pixels.
5. *Ratio of maximum vs. second maximum intensityI_r:* This is the ratio of the largest peak divided by the second largest peak, defined as
$$I_r = \frac{I_{max,2}}{I_{max}}$$
.
6. *Otsu's Threshold T_O:* Otsu's threshold[11] is a commonly used threshold method to automatically calculate a threshold to transform a grayscale image to a corresponding binary image. It is used as a measure of image intensity scale in our method.
7. *Total energy:* The sum of square of all pixel intensities. This is regarded as the optical energy.
8. *Number of ROIs* After preprocessed image segmentation step, the number of ROIs is used as a global image feature. The other of the features can be regarded as the correction of the errors from number of ROIs.
9. *Minimum AIC $k_{min,AIC}$ and BIC $k_{min,BIC}$ for GMM:* The number of components in GMM which minimizes Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) and its corresponding values from GMM. The details will be discussed in the following section.

The features described above can fall into several categories. Feature 1-7 are features from statistical information on image intensity. Feature 8 is the result from image segmentation, which is treated as the original results from preliminary studies. Feature 9 is the proposed new feature based on GMM.

All the first eight features can be represented by scalar values, while the eighth feature contains four scalar values itself. Then they are padded into a 12×1 vector as the image descriptor.

3.2.2 Serialization of Two-dimensional Images in Gaussian Mixture Model

Gaussian Mixture Model (GMM)[12] is widely applied in areas of pattern classification, data mining and natural language processing. A Gaussian Mixture distribution has the form of

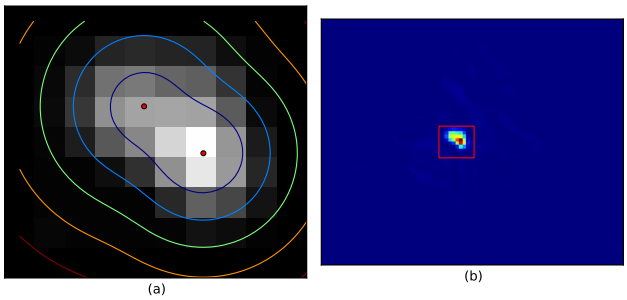

Figure 3. Gaussian Mixture Model in normal disomy image. The original image is (b) and the red rectangle is the bounding box of the image segmentation results. The two spots are closely overlapped. The red dots in (a) are the mean of the Gaussian kernels, and the estimation of kernel intensities are illustrated by the colored contours. Using Gaussian Mixture Model and choosing the minimum BIC helps us decide the number of kernels to be 2.

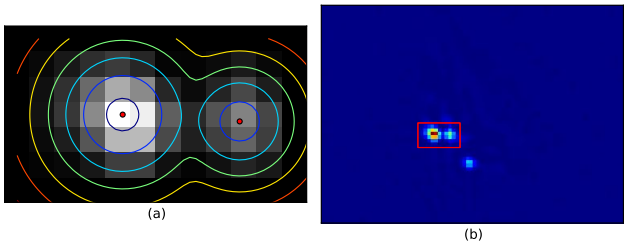

Figure 4. Gaussian Mixture Model in trisomy image with two close spots. Two Gaussian components are detected in the red rectangle area in (b) and illustrated in (a). The spot on the right has less than half the maximum intensity of the left one. The result shows that GMM is robust in the detection of dynamic range of intensities.

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \phi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}), \tag{1}$$

where the parameters $\{\phi_k\}, k = 1, \ldots, K$ must satisfy $\phi_k \in [0, 1]$ with $\sum_{k=1}^{K} \phi_k = 1$. $K = 1, \ldots, K_M$ is the number of Gaussian components and the maximum possible number of components K_M should be set as 3 in our case to sufficiently fit the image with at most three real spots. For each component, the parameters of Gaussian kernel are the mean $\boldsymbol{\mu_k}$ and the covariance $\boldsymbol{\Sigma_k}$, corresponding to the location and the shape, respectively. These parameters will be obtained by Expectation Maximization algorithm (EM)[12].

To fit GMM into our two-dimensional intensity images, each intensity unit in pixels is treated as a two-dimensional data point $\boldsymbol{X_i} = (x_{i,1}, x_{i,2})$, where $x_{i,1}$ and $x_{i,2}$ are the coordinate of the pixel location, $\boldsymbol{X}$ is the training data of GMM and $\boldsymbol{X}_i$ is the i-th data point in $\boldsymbol{X}$. Since the training data of GMM should be of the dimension $N \times 2$ while the images are two-dimensional array with the value as the image intensity, a serialization operation must be performed for each image. The data point $\boldsymbol{X}_i$ are inserted into $\boldsymbol{X}$ by the times the same as the number of the intensity value at $\boldsymbol{X}_i$.

The complexity of EM algorithm can be linearly reduced before performing the transformation described above if the intensity value of all the images will be divided by an arbitrary value M. The pseudo-value of the pixel located at $(x_{i,1}, x_{i,2})$ $\widetilde{\boldsymbol{X_i}}$ is calculated by the following equation:

$$\widetilde{\boldsymbol{X_i}} = \left\lfloor \frac{I(x_{i,1}, x_{i,2})}{M} \right\rfloor, \tag{2}$$

where $I(x_{i,1}, x_{i,2})$ is the intensity value at $(x_{i,1}, x_{i,2})$. For example, if the intensity of pixel at (15,28) is 240 and M is set as 30, there will be $240/30 = 8$ data points with the value of (15, 28) inserted into the training data $\boldsymbol{X}$.

After the serialization transformation of the two-dimensional image, we should get the training data with the size of $\widetilde{N} \times 2$. Here $\widetilde{N} = \lfloor I_T/M \rfloor$, where I_T is the total intensity defined in Section 3.2.1.

3.2.3 Minimum AIC and BIC of Gaussian Mixture Model

Given a GMM with the parameter $\theta = \{K, \phi_1, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}, \ldots, \phi_K, \boldsymbol{\mu_K}, \boldsymbol{\Sigma_K}\}$, the log-likelihood function $\ln \mathcal{L}$ can be obtained by

$$\ln \mathcal{L}(\boldsymbol{X}|\theta) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \phi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \right\}. \tag{3}$$

Then AIC[13] and BIC[14] are defined as

$$\mathrm{AIC}(K) = 2K - 2 \ln \mathcal{L}$$

and

$$\mathrm{BIC}(K) = -2 \cdot \ln \mathcal{L} + K \cdot (\ln(n) - \ln(2\pi))$$

respectively, where K is the number of Gaussian components in GMM and n is the number of data points in X. Both AIC and BIC are calculated with respect to K from 1 to K_M and K with the minimum value of AIC and BIC are chosen as the two image features denoted as

$$K_{min,AIC} = \underset{K \in 1,\ldots,K_M}{\arg\min}\ \mathrm{AIC}(K)$$

and

$$K_{min,BIC} = \underset{K \in 1,\ldots,K_M}{\arg\min}\ \mathrm{BIC}(K)$$

.

One of the major drawbacks of other spot detection methods is the uncertainty of the thresholds, especially when overlapping spots have different intensity range. Figure 4 shows the larger spot with an intensity of approximately 1000 while the smaller spot has its maximum intensity value of approximately 300. Both the static or dynamic thresholds are likely to reduce the area or even filter the spot out. Using the inference of Gaussian Mixture Model, we can get the parameters of each spot and then apply spot detection rules while getting rid of the risk of deleting positive spots during the early preprocessing procedures. With model selection methods like AIC and BIC, the minimum number of components with acceptable log-likelihood can be chosen by selecting the component number with the minimum value.

GMM is extremely effective in the detection of overlapping spots. Some spots may be very close to each other. By the component location, the overlapping of part of spots will not impact the component number determination.

3.3 Classification

This step aims at classifying an FISH-IS image into one of the three categories: monosomy, disomy and trisomy. In this study, the 10-fold cross validation is used while testing the classifiers. The measure of the performance for the classifier is the average accuracy among three categories. The accuracy is defined as the proportion of correctly classified samples to the total number of samples.

Ground Truth/ Detected as	Monosomy	Disomy	Trisomy
Monosomy	99.3%	0.7%	0.0%
Disomy	42.8%	57.2%	0.0%
Trisomy	29.7%	51.0%	19.2%

Table 1. The counting results from preprocessed ROIs. The result demonstrates that only 57.2% of disomy and 19.2% of trisomy images contain exactly two or three clear spots.

4. EXPERIMENTAL RESULTS

Once a FISH-IS image is retrieved from FISH-IS dataset, the global features are extracted right after the preprocessing procedures. The image processing methods are implemented in Python based on the open source image processing library scikit-image[15]. All global features are flattened as a vector to the classifier. In our FISH-IS image dataset, we have 6247 monosomy images, 5371 disomy images, and 7089 trisomy images. To evaluate the performance of Random Forest classifier, a 10-fold cross validation is used to obtain average accuracy of trained models.

The result from preprocessing and ROI segmentation methods in Table 1 indicates that the detection rate of single spot image is high but only 57.2% of disomy and 19.2% of trisomy images contain exactly two or three clear spots. This is caused by overlapping of the spots. While choosing a higher noise ratio T_N can eliminate some of the overlapping cases, the thresholding method will remove some of the valid spots as well. As a result, it will not be a right solution to our problem.

Ground Truth/ Detected as	Monosomy	Disomy	Trisomy
Monosomy	99.1%	0.6%	0.3%
Disomy	0.6%	79.5%	19.9%
Trisomy	0.9%	18.9%	80.2%

Table 2. Detection accuracy of classification via Random Forest classifier

Table 2 shows the final classification accuracy of the Random Forest classifier. The number of trees is set as 10, which is the default value from the Python package Scikit-learn[16].

From Table 2 it is demonstrated that by the classification system we propose, the detection rate of disomy and trisomy are significantly improved compared with only image processing methods or classifiers based on image features while the high accuracy of the monosomy classification is still retained. Besides, from the classifier we trained during our experiments, the monosomy images can be well classified from disomy and trisomy images.

5. CONCLUSION

In this paper, we proposed a GMM-based feature for spot counting in FISH-IS images, the pattern of the spots in which are not specified prior to the study. This method does not require any manual labeling of spots by experts. The features generated by the GMM kernel density estimation method can accurately locate the spot even for largely overlapping cases. The result of our method can be regarded as a important reference for the diagnosis by clinical experts. Practically, considering the poor image quality and the statistics from the image segmentation results which indicates a large proportion of disomy and trisomy images do not manifest as distinct spots, the accuracy is satisfactory.

References

[1] Minderman, H., Humphrey, K., Arcadi, J. K., Wierzbicki, A., Maguire, O., Wang, E. S., Block, A. W., Sait, S. N. J., George, T. C., and Wallace, P. K., "Image cytometry-based detection of aneuploidy by fluorescence in situ hybridization in suspension," *Cytometry A* **81A**, 776784 (Sep 2012).

[2] Theodosiou, Z., Kasampalidis, I. N., Livanos, G., Zervakis, M., Pitas, I., and Lyroudia, K., "Automated analysis of fish and immunohistochemistry images: a review," *Cytometry Part A* **71**(7), 439–450 (2007).

[3] Netten, H., Van Vliet, L. J., Vrolijk, H., Sloos, W. C., Tanke, H. J., and Young, I. T., "Fluorescent dot counting in interphase cell nuclei," *Bioimaging* **4**(2), 93–106 (1996).

[4] Lerner, B., Lev, K., and Josepha, Y., "Segmentation and classification of dot and non-dot-like fluorescence in situ hybridization signals for automated detection of cytogenetic abnormalities," *Information Technology in Biomedicine, IEEE Transactions on* **11**(4), 443–449 (2007).

[5] Raimondo, F., Gavrielides, M. A., Karayannopoulou, G., Lyroudia, K., Pitas, I., and Kostopoulos, I., "Automated evaluation of her-2/neu status in breast tissue from fluorescent in situ hybridization images," *Image Processing, IEEE Transactions on* **14**(9), 1288–1299 (2005).

[6] Rosin, P. L., "Unimodal thresholding," *Pattern Recognition* **34**(11), 2083 – 2096 (2001).

[7] Sagonas, C., Marras, I., Kasampalidis, I., Pitas, I., Lyroudia, K., and Karayannopoulou, G., "Fish image analysis using a modified radial basis function network," *Biomedical Signal Processing and Control* **8**(1), 30–40 (2013).

[8] Brauner, J. M., Groemer, T. W., Stroebel, A., Grosse-Holz, S., Oberstein, T., Wiltfang, J., Kornhuber, J., and Maler, J. M., "Spot quantification in two dimensional gel electrophoresis image analysis: comparison of different approaches and presentation of a novel compound fitting algorithm.," *BMC Bioinformatics* **15**, 181 (2014).

[9] Ram, S., Rodrguez, J. J., and Bosco, G., "Segmentation and detection of fluorescent 3D spots," *Cytometry A* **81A**, 198212 (Mar 2012).

[10] Zhang, T. and Suen, C. Y., "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM* **27**(3), 236–239 (1984).

[11] Otsu, N., "A threshold selection method from gray-level histograms," *Automatica* **11**(285-296), 23–27 (1975).

[12] Bishop, C. M., [*Pattern recognition and machine learning*], vol. 1, Springer-Verlag New York, Inc. (2006).

[13] Akaike, H., "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on* **19**(6), 716–723 (1974).

[14] Schwarz, G., "Estimating the dimension of a model," *The annals of statistics* **6**(2), 461–464 (1978).

[15] van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ* **2**, e453 (6 2014).

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research* **12**, 2825–2830 (2011).