Color Compensation of Multicolor FISH Images

Hyohoon Choi*, *Member, IEEE*, Kenneth R. Castleman, *Member, IEEE*, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—Multicolor fluorescence *in situ* hybridization (M-FISH) techniques provide color karyotyping that allows simultaneous analysis of numerical and structural abnormalities of whole human chromosomes. Chromosomes are stained combinatorially in M-FISH. By analyzing the intensity combinations of each pixel, all chromosome pixels in an image are classified. Due to the overlap of excitation and emission spectra and the broad sensitivity of image sensors, the obtained images contain crosstalk between the color channels. The crosstalk complicates both visual and automatic image analysis and may eventually affect the classification accuracy in M-FISH. The removal of crosstalk is possible by finding the color compensation matrix, which quantifies the color spillover between channels. However, there exists no simple method of finding the color compensation matrix from multichannel fluorescence images whose specimens are combinatorially hybridized. In this paper, we present a method of calculating the color compensation matrix for multichannel fluorescence images whose specimens are combinatorially stained.**

***Index Terms*—Chromosomes, color compensation, crosstalk, fluorescence *in-situ* hybridization (FISH), fluorescence, multicolor fluorescence *in-situ* hybridization (M-FISH), multicolor, multispectral.**

I. Introduction

THE fluorescence *in situ* hybridization (FISH) microscopic imaging modality has been widely used for the analysis of genes and chromosomes. Multiple fluorophores are often used combinatorially to visualize multiple biological components simultaneously. Using combinatorial labeling methods, $2^N - 1$ components within the specimen can be discriminated using N fluorophores. When five fluorophores are used, 31 objects can be analyzed by the binary combinations (presence or absence) of the fluorophores. N gray scale images of specimens, stained with N fluorophores, can be obtained using a monochrome camera and a set of optical bandpass filters that are specifically designed for the excitation and emission wavelengths of the fluorophores.

In particular, multiplex FISH (M-FISH) uses five fluorophores to uniquely identify all 24 chromosome types in human genome. A sixth fluorophore, DAPI (4′-6-diamidino-2-phenylindole, a blue fluorescent dye), is used to counterstain the chromosomes. Using an epifluorescence microscope and a set of optical bandpass filters, six monochrome images are captured.

Each pixel of an M-FISH image is composed of six values that correspond to the intensities of six fluorophores. By analyzing the intensity combinations of the pixels, all of the chromosome pixels in an image are identified, and a pseudocolor is assigned based on the class the pixel belongs to [1], [2].

M-FISH is an important tool for the visualization of translocations (exchange of chromosomal material between chromosomes), which is extremely common in cancer cells. A high accuracy in pixel classification has been one of the central goals to achieve the success of the M-FISH technique. The channel crosstalk complicates processing of the images and also may adversely affect the classification results, depending on the classification methods, by making less certain of the likelihood of a pixel to the correct class.

In general, there are two types of M-FISH systems: systems that use a set of optical bandpass filters as described above [3], and systems that use an interferometer [4]. While the earlier type captures a number of monochrome images which correspond to the intensities of each fluorophore, the later type captures the complete emission spectra over a range of wavelengths at every pixel. Developed techniques in this paper are pertinent to the earlier one.

When specimens are singly stained, each color channel of the image should display only one fluorophore component. However, due to the overlap of excitation and emission spectra and the broad sensitivity of imaging sensors, the obtained images contain a certain amount of crosstalk between the color channels. This phenomenon is called color spreading [5] or color spillover [6] (Note that in the field of flow cytometry, "color spreading" is termed as "color spillover"). These color spreadings introduce uncertainties and complexities into the image analysis.

The inverse process of color spreading is called color compensation. The color compensation method for FISH images was first introduced by Castleman [5], [7], [8] by modeling the color spreading effect as a linear transformation. Color compensation can be done only if the color spread matrix is defined. The color spread matrix quantifies the amount of each color being spread to the other colors. So far, the color spread matrix has been experimentally found, usually for cases where objects within the specimen are singly labeled.

By measuring the intensities across color channels at the location of an object that is singly labeled, we can easily quantify how much the single color is spilt over other channels. However, when none of objects are singly labeled, the determination

Manuscript received April 12, 2008; revised June 03, 2008. First published July 15, 2008; current version published December 24, 2008. This work was supported by the United States National Institute of Health under Grant R44 HD-038151. *Asterisk indicates corresponding author.*

*H. Choi is with the Sealed Air Corporation, San Jose, CA 95110 USA (e-mail: hyohoon@alumni.utexas.net).

K. R. Castleman was with Advanced Digital Imaging Research, League City, TX 77573 USA (e-mail: ken@castleman.org).

A. C. Bovik is with the Laboratory of Image and Video Engineering, The University of Texas, Austin, TX 78712 USA. (e-mail: bovik@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2008.928177

of color spread matrix is not easy unless singly labeled reference samples are specifically prepared.

For interferometer based systems, there also exist color compensation methods [6], [9]. The preferred color compensation method for this kind of systems is linear unmixing, which also requires prior knowledge of the reference spectra for a given dye.

In this paper, we present a mathematical method of calculating the color spread matrix for M-FISH without the need of preparing singly labeled samples. The new technique can be easily applied to other combinatorially labeled FISH images and multichannel images that exhibit a similar signal formation pattern as M-FISH images.

In Section II, a method of computing color spread matrix from the captured images is described. An example of calculating the color spread matrix is described in III. In Section IV, we present color compensation results applied to M-FISH images.

II. Color Compensation

A. Signal Model

Castleman [7], [10] has modeled the color spreading effect as a linear transformation

$$\mathbf{y} = \mathbf{E}\mathbf{C}\mathbf{x} + \mathbf{b} \tag{1}$$

where $\mathbf{C}$ is the $N \times N$ color spread matrix that specifies how the colors are spread among the channels, $\mathbf{x}$ is the $N \times 1$ vector of true fluorophore intensities at a particular pixel, $\mathbf{b}$ is the $N \times 1$ vector of black-level offsets of the imaging sensors (e.g., three channel color CCD or monochrome CCD), $\mathbf{E}$ is the $N \times N$ diagonal matrix of exposure times of each channel, and $\mathbf{y}$ is the $N \times 1$ vector containing the measured intensity values at a particular pixel. This model assumes that the gray levels are linear with brightness of the fluorophores.

Then the true intensities $\mathbf{x}$ can be found, given $\mathbf{y}$, $\mathbf{E}$, $\mathbf{C}$, and $\mathbf{b}$ by

$$\mathbf{x} = \mathbf{C}^{-1}\mathbf{E}^{-1}\{\mathbf{y} - \mathbf{b}\}. \tag{2}$$

Normally, $\mathbf{y}$, $\mathbf{E}$, and $\mathbf{b}$ are given but the color spread matrix is not. Without the color spread matrix, the true intensities cannot be recovered. When objects are uniquely stained, estimating the color spread matrix is relatively simple. We will illustrates the procedure by an example. Suppose three biological objects are uniquely stained with three fluorophores as shown in Table I. Then (1) can be written

$$\begin{bmatrix} y_{ri} \\ y_{gi} \\ y_{bi} \end{bmatrix} = \begin{bmatrix} E_r & 0 & 0 \\ 0 & E_g & 0 \\ 0 & 0 & E_b \end{bmatrix} \begin{bmatrix} C_{rr} & C_{gr} & C_{br} \\ C_{rg} & C_{gg} & C_{bg} \\ C_{rb} & C_{gb} & C_{bb} \end{bmatrix} \begin{bmatrix} x_{ri} \\ x_{gi} \\ x_{bi} \end{bmatrix} + \begin{bmatrix} b_r \\ b_g \\ b_b \end{bmatrix} \tag{3}$$

where y_{ri} is the gray level of red channel at the ith pixel, E_r is the exposure time for red channel, C_{rg} is the proportion of red color being spread to green channel, and the rest are analogous.

The color spread matrix, which is dependent on the filter sets, fluorophores, and imaging sensors, can be found from three $\mathbf{y}$ vectors from three objects. The intensities of the observed signal $\mathbf{y}_{i\in R,G,or\ B}$ are $[80, 15, 10]^T$ for red dye, $[5, 80, 30]^T$ for green dye, and $[10, 10, 160]^T$ for blue dye respectively, and

TABLE I
Color Map: Object 1 Is Stained With Red Dye, Object 2 Is Stained With Green Dye, and Object 3 Is Stained With Blue Dye

		Fluor spectra		
		R	G	B
Objects	1	x	0	0
	2	0	x	0
	3	0	0	x

$[E_r, E_g, E_b] = [1, 1, 2]$ and $\mathbf{b} = [0, 0, 0]^T$. R, G, or B is a set of indexes of the objects stained with red, green, or blue dye, respectively. Knowing that the intensities of $\mathbf{y}_i$ are originated only from the intensity of red, green, or blue dye, the true pixel values are found by $\mathbf{x}_{i\in R} = [y_r + y_g + y_b/2, 0, 0]^T = [100, 0, 0]^T$, similarly $\mathbf{x}_{i\in G} = [0, 100, 0]^T$, and $\mathbf{x}_{i\in B} = [0, 0, 100]^T$. Here y_b is divided by 2 because of the integration time. After plugging in $\mathbf{x}_i$ into (3), the nine unknowns of $\mathbf{C}$ can be found by solving nine linear equations from $\mathbf{E}^{-1}\mathbf{y}_i = \mathbf{C}\mathbf{x}_i$. Simply, calculating the intensity ratios of $\mathbf{y}_i$ is the solution in this case. Thus, the color spread matrix in this example is

$$\mathbf{C} = \begin{bmatrix} 0.8 & 0.05 & 0.1 \\ 0.15 & 0.8 & 0.1 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}.$$

The first column of the color spread matrix tells that 15% and 5% of the red intensity is spread to the green and blue channels, respectively. The inverse matrix of the color spread matrix, called the color compensation matrix, corrects these color spreadings and recovers the true signal intensities. Once the color spread matrix is computed, it can be used for other images that are captured under the similar (ideally the same) conditions using the same optical system and fluorophores.

When the specimens are combinatorially stained, the estimate of the true intensities from the observed signal cannot be done in the same way as when the specimens are uniquely stained. In M-FISH, six fluorophores are combinatorially used to discriminate 24 chromosome types. Chromosome 1, for example, is stained only with DAPI and spectrum gold dyes, chromosome 2 with DAPI and Red, chromosome 4 with DAPI, Green and Red, and so on. In the following sections, we will explain how to compute the color spread matrix $\mathbf{C}$ from only the observed signal $\mathbf{y}$ and the exposure times, $\mathbf{E}$.

The measured signal $\mathbf{y}$ of the M-FISH images can be written the same way as Castleman's model for FISH images but with slightly more details as

$$\mathbf{y} = \mathbf{E}\{\mathbf{C}\mathbf{x} + \mathbf{b}\} + \mathbf{n} \tag{4}$$

where $\mathbf{y}$ is the 6×1 vector of the observed signal intensities at a pixel, $\mathbf{x}$ is the 6×1 vector of the true signal intensities, $\mathbf{C}$ is the 6×6 color spread matrix, $\mathbf{b}$ includes the dc-offset of the CCD and various factors that cause background intensity elevation, $\mathbf{n}$ is the noise of the imaging device such as white noise and shot noise, and $\mathbf{E}$ is the 6×6 diagonal matrix of exposure times. In this model, we explicitly write that the background intensity increases linearly as the exposure time increases. Note that pixel index i for the vectors is specified only when necessary.

Six channels of the M-FISH image are first median filtered with a 3×3 kernel in order to eliminate the shot noise from $\mathbf{n}$,

and then lowpass filtered with a 3×3 kernel to remove the high frequencies which are mostly dominated by the white noise. Fortunately, the amount of shot noise and white noise from the imaging sensor in M-FISH images are not significant. Thus, the term $\mathbf{n}$ is effectively minimized from (4).

B. Background Correction

The background intensity $\mathbf{b}$ is mostly affected by the auto-fluorescence of the slide, the dc offset of the CCD, unattached free fluorescent molecules, the intensity of the defocused objects from out of depth of field, etc. Also, regions having a high density of objects usually have an elevated background intensity relative to regions without objects because of the flair effect. All these factors contribute to the nonflat intensity elevation of the background.

A 2-D cubic surface was estimated from the background pixels in order to remove $\mathbf{b}$. The surface that has the minimum mean square error relative to the background pixels is the estimated 2-D cubic surface [10]. Other various types of background correction techniques for microscope images can be found in [11].

The background pixels for each channel are found by a K-means clustering method ($K = 2$), in which the threshold is found while iteratively regrouping grayscale values into two classes until the class means converge. Given a grayscale image I, the intensity distribution is assumed to be a mixture of two gaussians: $p(x|\omega_1) \sim N(\mu_1, \sigma_1)$, $p(x|\omega_2) \sim N(\mu_2, \sigma_2)$, where x is the grayscale value in image I, ω_1 and ω_2 are the background class and foreground class, respectively, $p(x|\omega)$ is the probability density function, μ and σ are the mean and the standard deviation respectively. We further assumed that $\sigma_1 = \sigma_2$ for simplicity. Let D denote a set $\{x|x \in I\}$ of n unlabeled samples drawn independently from the mixture density

$$p(x) = p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2).$$

The decision boundary that partitions D into two groups, D_1 and D_2, is computed by minimizing the sum-of-squared error

$$J = \sum_{i=1}^{2} \sum_{x \in D_i} \|x - \mu_i\|^2.$$

μ_1 and μ_2 that minimize J are found iteratively using

$$T = \mu_1 P(\omega_1) + \mu_2 P(\omega_2) \tag{5}$$

where T is the decision boundary, μ_1 and μ_2 are the class means, and $P(\omega_1)$ and $P(\omega_2)$ are the prior probabilities ($P(\omega_1) + P(\omega_2) = 1$). Given the initial estimates of μ_1 and μ_2, the initial T is found. Using the initial T, the new means are found. This minimization process is repeated until the class means or T converges. $\min\{x\}$ and $\max\{x\}$ are chosen as the initial values for μ_1 and μ_2, respectively. The samples in D_1, pixels below the threshold T, represent the background pixels.

Given the background pixels, the 2-D cubic surface is estimated as follows. The function for a 2-D cubic surface [10] is

$$f(j,k) = c_0 + c_1 j + c_2 k + c_3 jk + c_4 j^2 + c_5 k^2 + c_6 j^2 k + c_7 jk^2 + c_8 j^3 + c_9 k^3 \tag{6}$$

where $c_0 \sim c_9$ are the coefficients that determine the surface shape, j and k are the coordinates, and $f(j,k)$ is the intensity value at jth row and kth column. The ten coefficients are estimated from the given m background pixels by solving $\mathbf{f} = \mathbf{Bc}$, where $\mathbf{f}$ is a $m \times 1$ column vector containing intensity values, $\mathbf{B}$ is a $m \times 10$ matrix containing j and k coordinates and their products with different powers, and $\mathbf{c}$ is a 10×1 column vector containing the ten unknowns

$$\mathbf{f} = \begin{bmatrix} f(j_1,k_1) \\ \vdots \\ f(j_m,k_m) \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_9 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & j_1 & k_1 & j_1k_1 & j_1^2 & k_1^2 & j_1^2k_1 & j_1k_1^2 & j_1^3 & k_1^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & j_m & k_m & j_mk_m & j_m^2 & k_m^2 & j_m^2k_m & j_mk_m^2 & j_m^3 & k_m^3 \end{bmatrix}.$$

The least squares solution for $\mathbf{c}$ that minimizes the sum of the mean square error is determined by

$$\mathbf{c} = [\mathbf{B}^T\mathbf{B}]^{-1}[\mathbf{B}^T\mathbf{f}].$$

Using the coefficients $\mathbf{c}$ and (6), a 2-D cubic surface that best fits the given background pixels is obtained. Finally, the estimated surface for each channel is then subtracted from the corresponding channel removing the above mentioned noises in $\mathbf{b}$ from (4).

C. Color Compensation

After the background correction, the signal model becomes

$$\mathbf{y} = \mathbf{ECx}. \tag{7}$$

The formation of this signal can be viewed as in Fig. 1. Six original gray levels of $\mathbf{x}$ are linearly mixed by the color spread matrix $\mathbf{C}$. Note that the vectors $\mathbf{x}$ and $\mathbf{y}$ in the signal model are the pixels only from the chromosome area now. Let's define $\mathbf{X}$ as a $6 \times n$ matrix containing n $\mathbf{x}$ vectors, and similarly $\mathbf{Y}$ is also a $6 \times n$ matrix. The observed signal before the exposure times are applied can be written as $\mathbf{E}^{-1}\mathbf{Y}$. The goal is to solve for $\mathbf{X}$ and $\mathbf{C}$ given the observation $\mathbf{E}^{-1}\mathbf{Y}$. Finding the linear mixing matrix $\mathbf{C}$ and the original signal $\mathbf{X}$ from the observed signal is a problem similar to the cocktail-party problem, where there are N speakers and N recording devices, and the N recorded signals are weighted sum of the N true signals. The recently developed technique called independent component analysis (ICA) has been used quite successfully to estimate the mixing channel parameters, $\mathbf{C}$, from the mixed signal $\mathbf{Y}$ based on the assumption that each digital signal (represented as a function of time) is statistically independent of each other at every time index [12]. ICA can also be used to separate the M-FISH mixed signal $\mathbf{Y}$ into six different statistically independent signals. However, the $\mathbf{X}$ that ICA estimates for M-FISH is not the same as the true

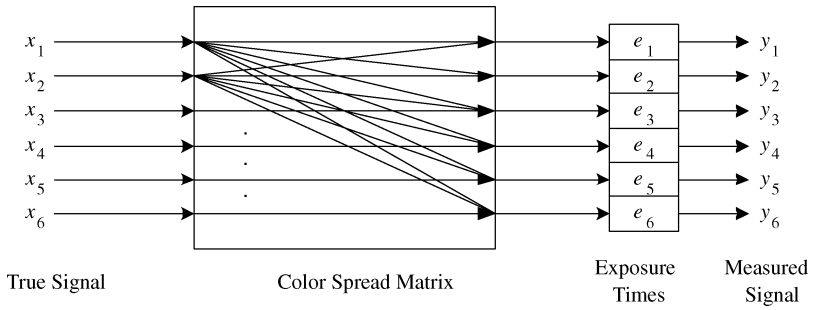


Fig. 1. Signal formation of M-FISH images. The measured intensities at a pixel is $\mathbf{y} = \mathbf{ECx}$. The original intensities $\mathbf{x} = [x_1, x_2, \ldots, x_6]$ at a pixel is linearly mixed by the color spread matrix $\mathbf{C}$, and the mixed intensities are scaled by the integration times e_1 through e_6, resulting in the captured intensities $\mathbf{y} = [y_1, y_2, \ldots, y_6]$.

signals since the combinatorial labeling causes dependencies among the true signals.

Let $\mathbf{y}_{\theta\_1,i}$ be an ith observed pixel that belongs to chromosome 1 or θ_1. A realistic set of pixel values of $\mathbf{y}_{\theta\_1,i}$ may be $[170, 65, 45, 189, 70, 76]_i^T$. From the color labeling information about chromosome 1, we know that at least four values in $\mathbf{x}_{\theta\_1,i}$ should be zero, i.e., $\mathbf{x}_{\theta\_1,i} = [x_1, 0, 0, x_4, 0, 0]_i^T$. Values in $\mathbf{x}$ should be zero where the staining is not present, and values remain unknown where fluorophores are present. This can be repeated for 24 chromosomes, creating 76 unknowns for the true signal. Also, all 36 values in the color spread matrix are unknown. Thus, the total number of unknowns are 112. Realistically, we may not find a unique solution given only the observed signal, but we can derive an optimal solution utilizing as much information as possible about the signal.

In practice, even after the careful noise removal, an observed pixel $\mathbf{y}$ will differ from $\mathbf{ECx}$, that is $\mathbf{y} = \mathbf{ECx} + \epsilon$, where ϵ may include factors not considered in our system model such as nonuniform hybridization inside of each chromosome, unremoved noise after median and lowpass filtering and background subtraction, and saturation of the pixels. Considering $\mathbf{y}$ as a multivariate random variable that belongs to a class θ_l, where $l = 1, \ldots, 24$, maximum-likelihood class parameters are estimated assuming the distribution of the random variable is normal. Therefore, we compute the mean vectors of each class from either manually or automatically classified images [2]. When using an automatic classification method, the result should be verified and misclassified pixels should be excluded. At least one image should be classified to identify pixels of each class and thus to formulate the equations. Having more pixels per class across from a number of images should reduce any possible variations between images. Once the color spread matrix is computed from one or a number of images, the color spreading of other images that are captured under the same conditions can be corrected. However, a caution should be taken when applying the color spread matrix to other images. The photo-bleaching effect, which is a well-known problem in fluorescence imaging, degrades the signal intensity as a function of time given an amount of excitation light. Thus, when capturing images, one should be attentive to applying the same total amount of excitation light at each imaging location to minimize the variation in signal intensities from image to image, and of course all other setups of the imaging system should be identical. Otherwise, a new color spread matrix has to be computed for a new set of differently captured images.

In order to minimize the effect of noise in the estimation of the color spread matrix, the 6×24 matrix $\mathbf{Y}$ is formed as

$$\mathbf{Y}^T = \begin{bmatrix} \mu_{\mathbf{y}_{\theta\_1}} = \frac{1}{P_1}\sum_{i=1}^{P_1} \mathbf{y}_{\theta\_1,i}^T \\ \mu_{\mathbf{y}_{\theta\_2}} = \frac{1}{P_2}\sum_{i=1}^{P_2} \mathbf{y}_{\theta\_2,i}^T \\ \mu_{\mathbf{y}_{\theta\_3}} = \frac{1}{P_3}\sum_{i=1}^{P_3} \mathbf{y}_{\theta\_3,i}^T \\ \vdots \\ \mu_{\mathbf{y}_{\theta\_24}} = \frac{1}{P_{24}}\sum_{i=1}^{P_{24}} \mathbf{y}_{\theta\_24,i}^T \end{bmatrix} \tag{8}$$

where $P_{l \in 1,2,3,\cdots,24}$ are the number of pixels that belong to each chromosome θ_l. If ϵ is negligible, $\mathbf{Y}$ can be expressed

$$\mathbf{Y} = \mathbf{ECX}. \tag{9}$$

The matrixes $\mathbf{C}$ and $\mathbf{X}$ contain the unknowns. In order to form a system of linear equations, (9) is written as

$$\mathbf{C}^{-1}\mathbf{E}^{-1}\mathbf{Y} - \mathbf{X} = 0. \tag{10}$$

The solution for (10) should satisfy the following constraints.

1) We assume that the intensity of all chromosomes stained with a particular dye should be the same in the original signal. For example, there are 10 chromosomes that are stained with green dye using Vysis probe, and the mean intensity of each chromosome should be the same. This assumes that all objects have the same hybridization sensitivity to the same fluorophore. However, if there are differences in the sensitivity and information is not given, then our assumption will give the best estimate. If the information is given, then the sensitivity ratios should be and can be incorporated into the equations.
2) The intensity between the input and output signals should be preserved, i.e., the sum along each columns of $\mathbf{E}^{-1}\mathbf{Y}$ should be the same as the sum along each columns of $\mathbf{X}$, $\sum \mathbf{E}^{-1}\mathbf{Y}_l = \sum \mathbf{X}_l$. To satisfy the this, each column of the color spread matrix should sum to 1.

Using these constraints, a nonhomogeneous linear system of 244 equations is formed as $\mathbf{Au} = \mathbf{h}$. The solution for 36 unknowns of $\mathbf{C}$ and 76 unknowns of $\mathbf{X}$ that optimally satisfy the equations is found. $\mathbf{A}$ is the 244×112 coefficient matrix, $\mathbf{u}$ is a column vector of the 112 unknowns, and $\mathbf{h}$ is a column vector of 214 zeros and 30 nonzero values. Among 214

TABLE II
EXAMPLE. COLOR LABELING TABLE (LEFT), COLOR SPREAD MATRIX (MIDDLE), AND EXPOSURE TIMES (RIGHT)

		Spectra				Spectra						
		1	2	3		1	2	3		1	2	3
Objects	1	x	x	0	1	0.8	0.05	0.1	1	4	0	0
	2	x	0	x	2	0.15	0.8	0.1	2	0	1	0
	3	x	x	x	3	0.05	0.15	0.8	3	0	0	2

equations, 144 equations are generated from (10), having 6×24 zeros on the right side of the equation, and 70 equations are formed by the condition that the none zero values in each row of $\mathbf{X}$ should be identical. Among 30 nonzero values in $\mathbf{h}$, 24 values are sums of intensities of each chromosome across spectra and 6 values are 1 s, representing sums of each column of the color spread matrix. The optimal solution that gives the minimum least squares error of this overdetermined problem is computed by $\mathbf{u} = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{h}$. In practice, $\mathbf{A}$ will not be error free. The amount of perturbation in $\mathbf{A}$ is directly related to the level of noise in $\mathbf{Y}$. Thus, pixels should be carefully selected, and especially pixels with saturated gray level, 255, should be avoided since the saturation is nonlinear. In actual computation, QR decomposition of $\mathbf{A}$ is used to find the solution and to avoid the calculation of $\mathbf{A}^H\mathbf{A}$, since $\mathbf{A}^H\mathbf{A}$ is strongly influenced by round off errors [13]. QR decomposition is a matrix factorization method that factorizes $\mathbf{A}$ as $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal matrix ($\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is an upper triangular matrix. The solution is then found by backsubstitution from $\mathbf{Ru} = \mathbf{Q}^T\mathbf{h}$.

The color spread matrix can only be estimated when the number of differently labeled objects is equal or larger than the number of fluorophores.

III. EXAMPLE OF COMPUTING THE COLOR SPREAD MATRIX

In this section, we show an example of how to formulate a linear system of equations from the observed signal. Suppose three objects are stained with three fluorophores combinatorially according to the color map shown in Table II, and the color spread matrix of the imaging system associated with those three fluorophores is also defined in Table II. The color map shows that object one is stained with fluorophore 1 and 2, object two is with fluorophore 1 and 3, and object three is with all three fluorophores. The color spread matrix indicates that, for each fluorophore, twenty percent of the original signal intensity is spread to the other channels. The original signal $\mathbf{X}$ is defined as

$$\mathbf{X}^T = \begin{bmatrix} 50 & 200 & 0 \\ 50 & 0 & 100 \\ 50 & 200 & 100 \end{bmatrix}.$$

Rows in $\mathbf{X}^T$ represent objects and columns represent spectra. The observed signal $\mathbf{Y}$ is defined as $\mathbf{Y} = \mathbf{ECX}$. Strictly speaking, the matrix $\mathbf{Y}$ is a set of means of each object as shown in (8). Then the matrix of the observed signal is

$$\mathbf{Y}^T = \begin{bmatrix} 200 & 167.5 & 65 \\ 200 & 17.5 & 165 \\ 240 & 177.5 & 225 \end{bmatrix}.$$

Now, given $\mathbf{Y}$, $\mathbf{E}$, and the color table, we will estimate the color spread matrix $\mathbf{C}$ and the original signal $\mathbf{X}$. We have nine unknowns for the color spread matrix and seven unknowns for the true signal. The solution for the total of sixteen unknowns can be found by solving the following equation $\mathbf{C}^{-1}\mathbf{E}^{-1}\mathbf{Y} - \mathbf{X} = 0$ with conditions defined in II-C. The equation can be written

$$\begin{bmatrix} u_1 & u_2 & u_3 \\ u_4 & u_5 & u_6 \\ u_7 & u_8 & u_9 \end{bmatrix} \begin{bmatrix} 50 & 50 & 60 \\ 167.5 & 17.5 & 177.5 \\ 32.5 & 82.5 & 112.5 \end{bmatrix} - \begin{bmatrix} u_{10} & u_{12} & u_{14} \\ u_{11} & 0 & u_{15} \\ 0 & u_{13} & u_{16} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (11)$$

Equation (11) can be written as a linear system of m equations in 16 unknowns, $\mathbf{Au} = \mathbf{h}$, where $\mathbf{A}$ is the coefficient matrix, $\mathbf{u}$ is the column vector of the unknowns, and $\mathbf{h}$ is the $m \times 1$ column vector. From (11), nine equations are formed. The sums of columns of $\mathbf{E}^{-1}\mathbf{Y}$ should be the same as the sums of columns of $\mathbf{X}$. This gives three equations. The sum of each column of $\mathbf{C}^{-1}$ should be 1. This gives three more equations. Further, nonzero values in a row of $\mathbf{X}$ should be the same, yielding four more equations. Thus, a total of 19 equations are formed. A linear system of m equations in n unknowns has a unique solution if the coefficient matrix $\mathbf{A}$ and the augmented matrix $\tilde{\mathbf{A}}$ has the same rank, and the rank equals n. In this example, $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\tilde{\mathbf{A}}) = 16$. The solution is found by the QR decomposition. $\mathbf{u}(1 \ldots 9)$ contains the solution for $\mathbf{C}^{-1}$. Then the estimated color spread matrix $\hat{\mathbf{C}}$ is

$$\hat{\mathbf{C}} = \begin{bmatrix} 0.8 & 0.05 & 0.1 \\ 0.15 & 0.8 & 0.1 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}$$

$\mathbf{u}(10 \cdots 16)$ contains the solution for the unknown $\mathbf{X}$ values. The true signal estimated is

$$\hat{\mathbf{X}}^T = \begin{bmatrix} 50 & 200 & 0 \\ 50 & 0 & 100 \\ 50 & 200 & 100 \end{bmatrix}$$

$\hat{\mathbf{C}} = \mathbf{C}$ and $\hat{\mathbf{X}} = \mathbf{X}$. Thus, the MSE between the estimation and the truth is zero. In this example, the proposed method finds the unknowns with no error.

IV. RESULTS

A. Color Compensation of M-FISH Images

In this section, we show the result of color compensating M-FISH images, and in addition we quantitatively show the improvement in image quality after the color compensation using mean squared error (MSE) and the structural similarity index (SSIM), a recently developed metric that has been shown to significantly surpass the MSE as a means for quantifying structural similarities between two images [14]. Given two images $\boldsymbol{\alpha}$ and

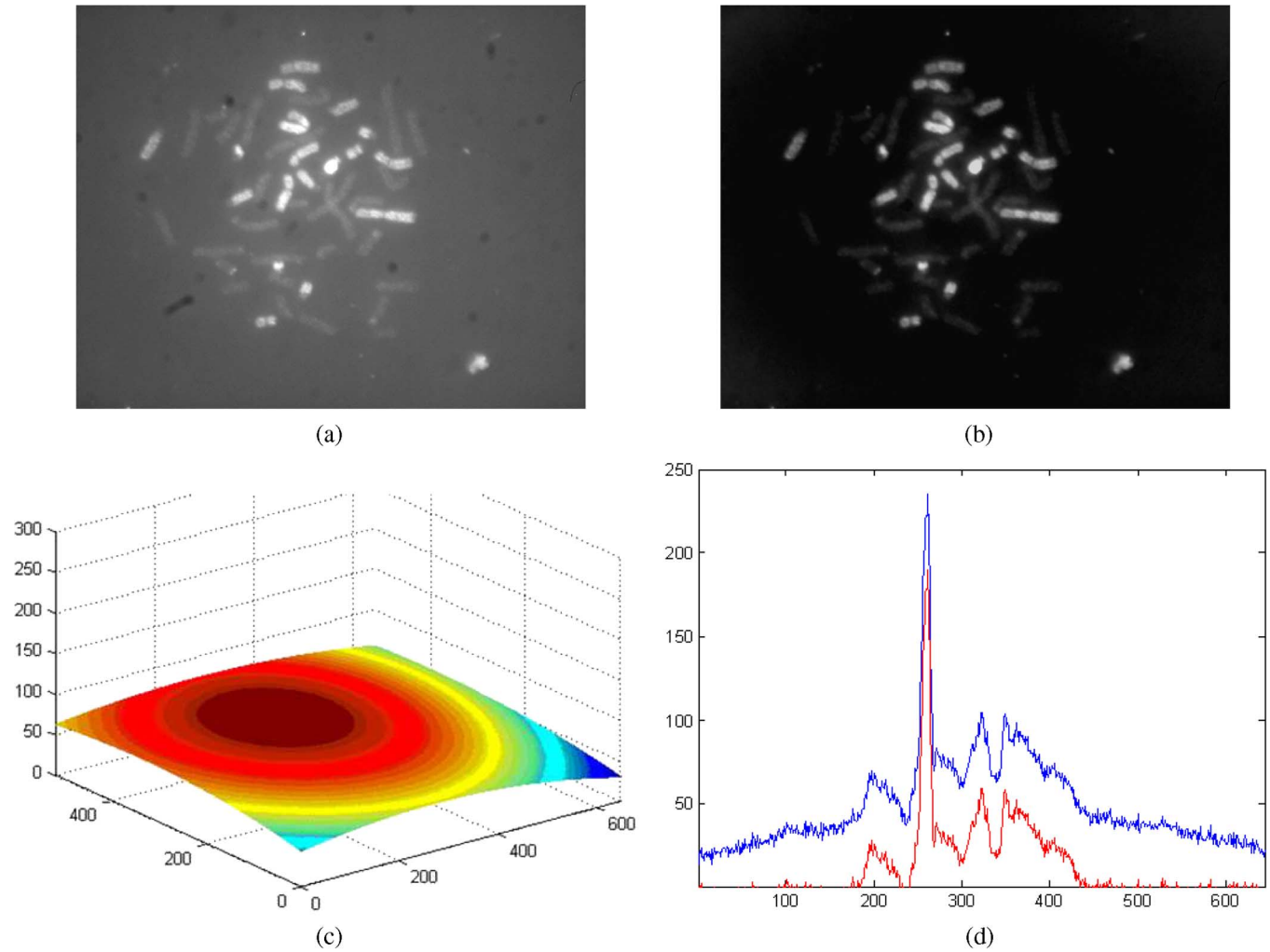


Fig. 2. Background correction. Elevated background intensity is removed after the background correction, but the channel crosstalk still remained. Profiles in (d) are drawn from middle rows of (a) and (b). (a) Original aqua channel. (b) Background corrected image. (c) Estimated cubic surface. (d) Profile: top-before, bottom-after.

$\boldsymbol{\beta}$, SSIM includes the luminance, contrast, and structure terms, and it is expressed

$$\mathrm{SSIM}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{(2\mu_{\boldsymbol{\alpha}}\mu_{\boldsymbol{\beta}})(2\sigma_{\boldsymbol{\alpha}\boldsymbol{\beta}} + C_2)}{\left(\mu_{\boldsymbol{\alpha}}^2 + \mu_{\boldsymbol{\beta}}^2 + C_1\right)\left(\sigma_{\boldsymbol{\alpha}}^2 + \sigma_{\boldsymbol{\beta}}^2 + C_2\right)} \qquad (12)$$

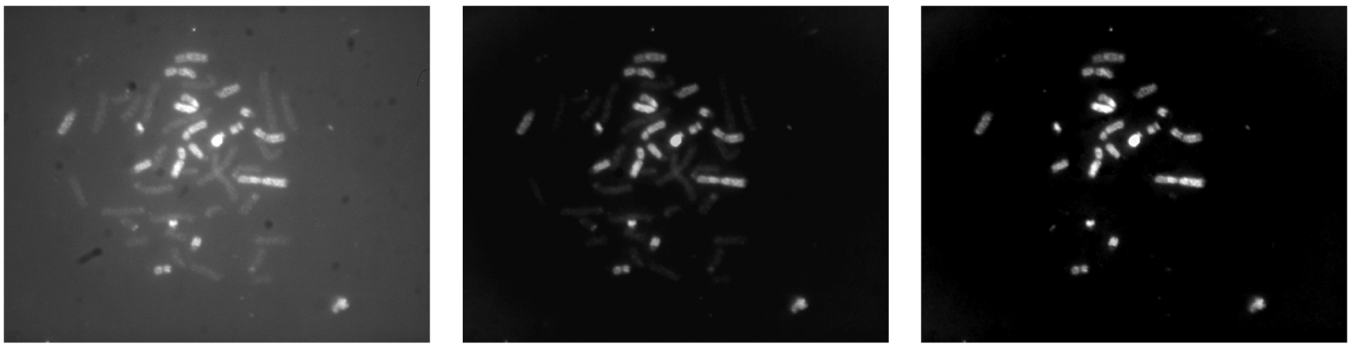
where $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, and L is the dynamic range of the pixel values (255 for 8-bit images). For the constants, we used $K_1 = 0.01$ and $K_2 = 0.03$ (refer [14] for the details). SSIM is calculated within a 11×11 circular symmetric window, which moves pixel-by-pixel over the entire image. It is easily shown [14] that $0 \leq \mathrm{SSIM}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leq 1$, where $\mathrm{SSIM}(\boldsymbol{\alpha},\boldsymbol{\beta}) = 1$ if and only if $\boldsymbol{\alpha} = \boldsymbol{\beta}$.

M-FISH images are obtained from a publicly available database at Advanced Digital Imaging Research, which contains 200 hand-segmented M-FISH images. The database is available online.[1] Fig. 2 shows the result of background correction. The original image in Fig. 2 has an elevated background and displays channel crosstalk. The estimated cubic surface of the background is shown in Fig. 2(c). The background corrected image shown in Fig. 2(b) is obtained after subtracting Fig. 2(c) from Fig. 2(a). Fig. 2(d) is a profile drawn from the middle rows of Fig. 2(a) and (b), and it clearly shows that the background elevation is effectively removed.

After correcting the background of five images that are captured from the same slide, pixels from each chromosome class are collected from those images. The means of each class are computed to form the matrix $\mathbf{Y}$. Then the color spread matrix is calculated. Fig. 3(c) shows the color compensation result on an image, which was not included in the five sample images to compute the color spread matrix, and as shown in the figure, all the crosstalk was effectively removed. A significant improvement in image quality is achieved after the color compensation.

Six-channel synthetic images, representing the ideal color compensation result, are generated using the ground truth from the database in order to quantify the image quality improvement. Ideally the images should be binary and represented by only two values, for example, 0 for nonfluorophore intensity and any reasonably large value, within the grayscale, for fluorophore

[1]http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml



(a) (b) (c)

Fig. 3. Color compensation. The color compensation removed the channel crosstalk effectively. A significant increase in image quality is achieved on image (c). (a) Original image. (b) Background correction. (c) Color compensation.

TABLE III
IMAGE QUALITY IMPROVEMENT. SUBINDEX B REPRESENTS BEFORE COLOR COMPENSATION, AND SUBINDEX A REPRESENTS AFTER COLOR COMPENSATION

Images	MSE_B	MSE_A	$MSSIM_B$	$MSSIM_A$
v1301xy	2196	534	0.074	0.693
v1303xy	922	184	0.074	0.765
v1309xy	1339	482	0.081	0.779
v1310xy	847	157	0.072	0.800
v1311xy	700	145	0.072	0.784
v1313xy	767	190	0.079	0.800
Average	1128	282	0.075	0.770

intensity. From the observation that the grayscale of chromosomes normally takes a midrange, a grayscale value of 128 is chosen for the chromosome intensity. The gray levels in areas where chromosomes overlap are usually saturated, and thus 255 is chosen for the overlapping area. The background area is set to 0. The MSE and the mean SSIM (MSSIM) are measured from before and after color compensated images against the synthetic images. An average of MSEs and MSSIMs across six channels per image is shown in Table III. MSSIM becomes one when two images are identical. As shown in Table III, the MSE reduced by a factor of 4 and the MSSIM increased by a factor of 10 after the color compensation.

V. DISCUSSION

M-FISH image formation involves much more complex processes and nonlinearities than our signal model as shown in (4). The estimation of color compensation matrix will be biased when the noise is included in the estimation. The noise terms in our signal model, such as $\mathbf{n}$ and $\mathbf{b}$, are rather effectively removed. However, other types of noise that are not included in our signal model, such as nonuniform hybridization inside chromosomes and among different chromosomes and unwashed free fluorescent molecules, and any deviations from true signal formation in our signal model affect the calculation of color compensation matrix. In order to minimize the effect of noise, images of good quality, i.e., large signal-to-noise ratio (SNR), less or no saturated pixels, and low biochemical noise, are preferred. When the level of noise is high, the solution for the color compensation matrix often becomes useless containing negative values. Noisy data should be avoided in the first place, but when inevitable a conditional optimization can be utilized to obtain a meaningful solution. Another factor that is not considered in our signal model, thus not included in the preprocessing steps, is the image misalignment. Image misalignment is a common phenomenon in multichannel images, which is mainly due to chromatic aberration of optical components. The degree of misalignment varies depending on the dataset. In ADIR M-FISH database, there are three different datasets distinguished by the labeling probes, which are Vysis, ASI, and PSI. Vysis dataset delivers high quality images—less biochemical noise, low misalignment, and large SNR—compared to the other two datasets. As the primary goal of this paper is to present the method of calculating the color compensation matrix, a subset of Vysis data is used as an example in this paper. Since the Vysis dataset has negligible amount of misalignment, the image registration was not performed in our case. However, when the amount of misalignment is significant, a proper image registration [15] must be performed as a preprocessing step before computing the mean vectors.

VI. CONCLUSION

We have shown a method of estimating the color spread matrix (or color compensation matrix) for combinatorially hybridized multicolor FISH images. The developed method does not require the preparation of singly labeled objects to obtain the reference spectra of each fluorophore. Instead, it only requires the knowledge of the mean vectors of all differently colored objects in the specimen. The mean vectors are computed, from a set of sample images, by either manually or automatically classifying objects. When automatic classification method is used, pixels below a certain posterior probability or likelihood value should be rejected. Once the color compensation matrix is computed from a set of sample images, all other images, that are taken under the same conditions as the sample images are captured, can be color compensated.

This paper also presents examples of formulating a linear system of equations in order to estimate the color spread matrix. The improvement on image quality after the color compensation is shown both qualitatively and quantitatively, and a significant improvement is observed. SSIM and MSE are used to measure the image quality improvement. MSSIM improved by a factor of 10, and MSE reduced by a factor of four on average after the color compensation.

The developed technique of computing the color spread matrix from a set of obtained images can be easily applied to other multicolor FISH images where specimens are combinatorially stained and imaged from a set of optical filter based systems, and furthermore it can be extended to other multichannel images that exhibit a similar signal formation pattern as the M-FISH images.

REFERENCES

[1] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Computer image analysis of combinatorial multi-fluor FISH," *Bioimaging*, vol. 4, pp. 52–64, 1996.

[2] W. Schwartzkopf, A. Bovik, and B. Evans, "Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images," *IEEE Trans. Med. Imag.*, vol. 24, no. 12, pp. 1593–1610, Dec. 2005.

[3] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Karyotyping human chromosomes by combinatorial multi-fluor FISH," *Nature Genetics*, vol. 12, pp. 368–375, 1996.

[4] E. Schrock, S. du Manoir, T. Veldman, B. Schoell, J. Wienberg, M. A. Ferguson-Smith, Y. Ning, D. H. Ledbetter, I. Bar-Am, D. Soenksen, Y. Garini, and T. Ried, "Multicolor spectral karyotyping of human chromosomes," *Science*, vol. 273, pp. 494–497, 1996.

[5] K. R. Castleman, "Color compensation for digitized FISH images," *Bioimaging*, vol. 1, no. 3, pp. 159–165, 1993.

[6] M. E. Dickinson, G. Bearman, S. Tille, R. Lansford, and S. E. Fraser, "Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy," *Biotechniques*, vol. 31, no. 6, p. 1272, Dec. 2001, 1274–1276, 1278.

[7] K. R. Castleman, "Color compensation with unequal integration periods," *Bioimaging*, vol. 2, no. 3, pp. 160–162, 1994.

[8] K. R. Castleman, T. P. Riopka, and Q. Wu, "FISH image analysis," *IEEE Eng. Med. Biol. Mag.*, vol. 15, no. 1, pp. 67–75, Jan.–Feb. 1996.

[9] H. Tsurui, H. Nishimura, S. Hattori, S. Hirose, K. Okumura, and T. Shirai, "Seven-color fluorescence imaging of tissue samples based on Fourier spectroscopy and singular value decomposition," *J. Histochem. Cytochem.*, vol. 48, no. 5, pp. 653–662, May 2000.

[10] K. R. Castleman, *Digital Image Processing*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[11] D. Tomaževič, B. Likar, and F. Pernuš, "Comparative evaluation of retrospective shading correction methods," *J. Microscopy*, vol. 208, no. 3, pp. 212–223, 2002.

[12] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[13] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[15] Y. Wang and K. R. Castleman, "Normalization of multicolor fluorescence in situ hybridization (M-FISH) images for improving color karyotyping," *Cytometry Part A*, vol. 64A, no. 2, pp. 101–109, 2005.