Sparse Non-negative Matrix Factorization (SNMF) based Color Unmixing for Breast Histopathological Image Analysis

Jun Xu^{a,b,*}, Lei Xiang^{a,b}, Guanhao Wang^{a,b}, Shridar Ganesan^c, Michael Feldman^d, Natalie Shih^d, Hannah Gilmore^e, Anant Madabhushi^f

^a*Jiangsu Key Laboratory of Big Data Analysis Technique, Nanjing University of Information Science and Technology, Nanjing 210044, China.*

^b*CICAET, Nanjing University of Information Science and Technology, Nanjing 210044, China.*

^c*Rutgers Cancer Institute of New Jersey, NJ 08901, USA.*

^d*Dept. of Pathology, Hospital of the University of Pennsylvania, PA19104, USA.*

^e*Institute for Pathology, University Hospitals Case Medical Center, Case Western Reserve University, OH 44106-7207, USA*

^f*Dept. of Biomedical Engineering, Case Western Reserve University, OH44106, USA.*

Abstract

Color deconvolution has emerged as a popular method for color unmixing as a pre-processing step for image analysis of digital pathology images. One deficiency of this approach is that the stain matrix is pre-defined which requires specific knowledge of the data. This paper presents an unsupervised Sparse Non-negative Matrix Factorization (SNMF) based approach for color unmixing. We evaluate this approach for color unmixing of breast pathology images. Compared to Non-negative Matrix Factorization (NMF), the sparseness constraint imposed on coefficient matrix aims to use more meaningful representation of color components for separating stained colors. In this work SNMF is leveraged for decomposing pure stained color in both Immunohistochemistry (IHC) and Hematoxylin and Eosin (H&E) images. SNMF is compared with Principle Component Analysis (PCA), independent Component Analysis (ICA), Color Deconvolution (CD), and Non-negative matrix Factorization (NMF) based approaches. SNMF demonstrated improved per-

*Corresponding author

Email address: xujung@gmail.com (Jun Xu)

formance in decomposing brown diaminobenzidine (DAB) component from 36 IHC images as well as accurately segmenting about 1400 nuclei and 500 lymphocytes from H & E images.

Keywords: Color unmixing; breast histology; Immunohistochemistry image; Hematoxylin and Eosin image

1. Introduction

In clinical practice, pathologists visually examine tissue samples under a microscope for the purpose of diagnosing the presence of disease. Different colour stains are usually used to indicate the presence of different proteins. Hematoxylin and Eosin (H&E) are standard stains that highlight nuclei in blue/purple and cytoplasm in pink in order to visualize the structure of interest in the tissue [28]. Hematoxylin (H) dye can stain all nuclei in the H&E histopathological images. Therefore, in order to detect or segment all the nuclei, one popular idea (such as blue ratio algorithm [7]) is to detect H stains or blue color in the image. Immunohistochemistry (IHC) is yet another type of tissues staining that attempts to capture the expression levels associated with specific proteins. For instance, CD31, CD34, and CD105 proteins are found on some specific types of cancerous cells in human breast histological tissue, respectively. As they are indicators of angiogenesis, the immunostaining of these proteins can be used to evaluate the degree of tumor angiogenesis in histological tissue sections [3, 5, 8, 22]. In clinical practice, the malignancy grading of the tumor and angiogenic activity are largely based on the intensity, architectural, and vascular density features of brown diaminobenzidine (DAB) in IHC stained breast tissue [2]. Therefore, quantitative analysis on DAB compound is the pre-requisite to evaluate the tumor growth and metastasis in breast tissue [17].

However, the IHC, H, E stains unfortunately overlap considerably which makes the problem of stain quantification within specific histologic primitives (e.g. nuclei, gland) challenging. Figure 1(a) is an example of a breast histopathological image that stained with CD31 IHC and H and E proteins. Stain or color separation in histology images could aid in automated detection of specific histologic structures such as nuclei and lymphocytes. A few approaches have been recently proposed to address this problem. Color Deconvolution (CD) is one such effective approach presented in [25] for unmixing pure stain colors. In [25], CD algorithm was employed for color

unmixing of IHC stained histological images. Similarly, CD was employed for separating DAB and H stains from IHC stained histological images in [2]. Then H and DAB components are further used for constructing vascular density and graph features, respectively. In [4, 21, 27], CD was used for H & E stained image analysis. The unmixed H stain component with CD was further used for nuclei detection and segmentation. Nuclear morphology is an important phenotypic feature for grading of disease aggressiveness. For instance, the modified Bloom Richardson (mBR) grading system for breast cancers is highly correlated with nuclei morphology. The alterations in nuclei size, shape, and topological information are important signatures for disease diagnosis and prognosis [35, 36, 37]. Recently a number of nuclear segmentation approaches have been proposed [1, 4, 7, 12, 21, 23, 26, 27]. CD was employed in [4, 21, 26, 27] as a pre-processing step for nuclei detection/segmentation. The unmixed H stain component was leveraged to detect the nuclei boundaries in [27]. Then nuclear morphological features are extracted to predict prognosis of male breast cancers. Similarly, in [4] cancer nuclei was detected based on the unmixed H stain component. Then graph features and texture features were subsequently extracted for quantifying the mBR grade in ER+ breast cancers. In [21], the unmixed H stain component was leveraged to identify different types of nuclei. In [26, 27], color unmixing and mathematical morphology operators were integrated with radial symmetry-based image transformation for nuclei detection. The watershed method was subsequently employed for nuclei segmentation. In [7], the blue ratio and multi-reference graph framework were employed for nuclei segmentation from H & E stained Glioblastoma Multiforme images. In [1], a hybrid active contour model developed on color intensities of RGB space was studied for nuclei segmentation in H & E stained breast histopathological images.

Before color unmixing, the CD approach usually transforms the H & E stained images into Optical Density (OD) images based on the Beer-Lambert Law [16]. The CD operation is subsequently applied to separate different color components by multiplication of OD image with the inverse of the color matrix. This stain matrix was defined by the measurement of relative absorption for red, green and blue components associated with each stain used in the tissue. One disadvantage of this approach is that the color matrix is pre-defined and therefore limits its applicability to other image modalities. Moreover, the CD method requires accurate estimation of the stain matrix [19].

Principle Component Analysis (PCA) [18] and independent Component Analysis (ICA) [15] are two popular unsupervised approaches that have been broadly used for representation learning [6]. Non-negative matrix Factorization (NMF) is the other important approach which decomposes a non-negative matrix into two low-rank non-negative matrices for finding a linear representation of non-negative data. In [24], an unsupervised spectral decomposition with non-negative matrix factorization (NMF) algorithm was presented for color unmixing on IHC images of human biopsy tissue. In [13], a blind color decomposition method was presented for color unmixing of histopathological images. Sparse representation is usually incorporated for improving the efficiency of decomposition. Sparsity methods have been widely investigated and applied for medical image analysis, such as in [9, 10, 11, 14, 29, 30, 31, 32, 33, 34, 38].

In this paper, we present a sparse NMF (SNMF) approach for color unmixing on histological images of the breast. The sparse representation of histological images aims to represent the image using fewer “active” components for better interpreting the staining of different components. The SNMF algorithm begins by transforming an intensity image in RGB space into the Optical Density (OD) image via the Beer-Lambert Law. Then the OD image is decomposed into a basis and a coefficient matrices, respectively. Multiplication of individual components of the basis matrix and coefficient matrix produces the OD image for each corresponding component. Finally, the intensity images of each component are represented in RGB space from the corresponding OD image. The post-processing operation is conducted on the intensity image of the resulting component. For instance, for H & E images, the unmixed H components can be leveraged for further segmenting nuclei/lymphocytes. To show the effectiveness of proposed approach in color unmixing, we qualitatively and quantitatively evaluate the SNMF against PCA, ICA, CD, and NMF on three data sets which includes 36 CD 31, CD34, and CD105 IHC and H & E stained histopathological images, 200 H & E lymph node-negative and estrogen receptor-positive (LN-, ER+) breast cancer images, and 100 human epidermal growth factor receptor-2 (HER2+) breast cancer images, respectively. SNMF is compared with other approaches in terms of decomposing brown diaminobenzidine (DAB) component from IHC images as well as accurately segmenting nuclei/lymphocytes from H & E images.

The rest of paper is organized as follows: A detailed description of Sparse Non-negative Matrix Factorization (SNMF) for color unmixing is presented in

Section 2. The experimental setup and comparative strategies are discussed in Section 3. The experiment results and discussions are reported in Section 4. Concluding remarks are presented in Section 5.

2. Methods

2.1. Image representation by the Beer-Lambert Law

It is assumed that the stained images are light absorbing and the relationship between the stain and its absorption satisfy the Beer-Lambert Law [16]. Let ζ be different color channels in RGB space, i.e. $\zeta \in \{R, G, B\}$. Each D_ζ is the corresponding OD vector of different color channels. According to Beer-Lambert Law, D_ζ of each color channel ζ for pixel p is expressed in terms of an absorbance which is defined as

$$D_\zeta(p) = -\ln \left(\frac{I_\zeta(p)}{I_0(p)} \right) = \varepsilon \times \ell \times \eta. \quad (1)$$

where ε , ℓ , and η represent respectively an absorptivity of the attenuator of the specimen, the distance the light travels through the specimen, and the concentration of staining in the specimen. I_0 is the intensity of light entering the specimen and I_ζ is the intensity of light detected in channel ζ after absorption as I_0 passes the specimen.

From Equation (1), it is easy to see that the intensities $I_\zeta \in \{I_R, I_G, I_B\}$ nonlinearly depend on the staining concentration η . Thus, these intensity values of RGB image cannot directly be used for decomposition of each of stains. However, by applying Beer-Lambert Law in (1), the corresponding optical density D_ζ for each channel is linear with η . Therefore, before employing SNMF for decomposition, the original RGB intensity image is transformed to corresponding vectors D_ζ .

2.2. Color deconvolution for color unmixing and its associated limitations

Based on color deconvolution method [25], the relationship between the amount of each stain

$$\varphi(p) = [\varphi_R(p), \varphi_G(p), \varphi_B(p)]^T$$

and OD

$$D(p) = [D_R(p), D_G(p), D_B(p)]^T$$

to pixel p can be expressed as

$$D(p) = S\varphi(p). \quad (2)$$

where

$$S = \begin{pmatrix} s_{R,1} & s_{G,1} & s_{B,1} \\ s_{R,2} & s_{G,2} & s_{B,2} \\ s_{R,3} & s_{G,3} & s_{B,3} \end{pmatrix} \quad (3)$$

is the staining matrix that defines the association between OD and the amount of each stain on pixel p . Each row in S corresponds to each staining channel.

For all pixels in the image, we can rewrite (2) as a data matrix

$$D = S\phi, \quad (4)$$

where

$$D = \begin{pmatrix} D_R(1) & D_R(2) & \dots & D_R(n) \\ D_G(1) & D_G(2) & \dots & D_G(n) \\ D_B(1) & D_B(2) & \dots & D_B(n) \end{pmatrix}, \quad (5)$$

and

$$\phi = \begin{pmatrix} \varphi_R(1) & \varphi_R(2) & \dots & \varphi_R(n) \\ \varphi_G(1) & \varphi_G(2) & \dots & \varphi_G(n) \\ \varphi_B(1) & \varphi_B(2) & \dots & \varphi_B(n) \end{pmatrix}. \quad (6)$$

The amount of stain ϕ can be obtained by solving the closed-form solution of (4).

However, although the amount of stain can be obtained with the method presented in [25], the CD framework requires that the stain matrix S in (4) is estimated in a way that is specific to the particular dataset. Once this matrix is estimated, it will be fixed and leveraged for all subsequent images. It is obvious that this approach has limitations. Aside from the fact that the estimated stain matrix S will be particular to the dataset under consideration, S might not be optimal for all images in the dataset.

2.3. Sparse Non-negative Matrix Factorization

It is easy to prove that the data matrix D in (5) is a non-negative matrix. SNMF aims to decompose a non-negative matrix into a basis matrix and a coefficient matrix. In general, PCA, ICA, and NMF/SNMF can all be seen as matrix factorization approaches, with different choices of objective function and/or constraints. Given a non-negative matrix $D \in R^{d \times n}$, where n is the number of data points (or pixels in the image) and d is dimension of each data point, it is easy to see that the desired rank of matrix D is $k \leq \min\{d, n\}$. The SNMF algorithm decomposes D into two non-negative matrix $B \in R^{d \times k}$ and $X \in R^{k \times n}$ so that

$$D \approx B \times X. \quad (7)$$

The solution of (7) can be obtained by solving the following optimization problem:

$$\begin{aligned} \min_{B, X} & \frac{1}{2} \|D - BX\|_F^2, \\ \text{s.t.} & \quad B, X \geq 0, \end{aligned} \quad (8)$$

where B and X represent basis matrix and coefficient matrix, respectively. In order to avoid the issue of scale-variance and non-unique solutions, the standard NMF algorithm (8) can be extended to Sparse NMF (SNMF) algorithm [20] by adding a regularization term and a sparsity term as follows:

$$\begin{aligned} \min_{B, X} & \frac{1}{2} \|D - BX\|_F^2 + \frac{\alpha_1}{2} \|B\|_F^2 + \frac{\alpha_2}{2} \|X\|_F^2 + \frac{\beta}{2} \sum_{i=1}^n \|X(:, i)\|_1^2 \\ \text{s.t.} & \quad B, X \geq 0 \end{aligned} \quad (9)$$

where α_1 , α_2 , and β are positive constants. The second and third terms represent the “smoothness” constraint which aims to penalize large values of the elements in basis and coefficient matrices. The fourth term of (9), the sparseness constraint is imposed on coefficient matrix by utilizing L_1 norm minimization. The sparseness constraint in Eq. (9) imposed on coefficient matrix aims to use desired dimensional of color space to better represent different components for more meaningful representation. The smoothness constraint imposed on both basis and coefficient matrices aims to smooth the image and remove white noises. The effectiveness of “smoothness” and “sparseness” constraints in (9) on color unmixing are analyzed in Section 3.

2.4. SNMF for histological image decomposition

Figure 1 shows the application of SNMF for color unmixing on an IHC image. Figure 1(a) represents the original CD31 IHC and H&E stained histological images of breast tissue. In Figure 1, the decomposition of immunostaining from original image (Figure 1(a)) is important for quantifying the brown regions, representing microvessels. By applying SNMF algorithm, the image is decomposed into antibody CD31 stained component (see Figure 1(b)) and H stained component (see Figure 1(c)).

Figure 1: SNMF based color unmixing on CD 31 IHC image. (a) Original IHC image, (b) antibody CD31 stained DAB component, and (c) Hematoxylin component.

Next, we use an H&E stained image as an example to show the principle of SNMF for color unmixing (see Figure 2). In Figure 2, each channel of original RGB image is decomposed into the constituent components and expressed as $I = [I_R^T; I_G^T; I_B^T]$. By applying Beer-Lambert Law in Eq. (1), intensity image I can be represented by OD image D in RGB space as $D = [D_R^T; D_G^T; D_B^T]$. Here we use D and D_ζ to represent OD matrix and a row vector of D , respectively. D is subsequently decomposed into a $d \times k$ basis matrix B and a $k \times n$ coefficient matrix X based on Eq. (7)–(9). Note that for two stained images (H & E), $k = 2$ and for CD31 IHC and H & E images, $k = 3$.

2.5. Image intensity reconstruction and nuclear segmentation

The OD image of H component is constructed by multiplying the first column $B(:, 1)$ of basis matrix B and the first row $X(1, :)$ of coefficient matrix X . Similarly, the E component can be reconstructed by the second column

Figure 2: The illustration of SNMF for H&E stained image decomposition. The Optical Density data D is decomposed into the constituent components: basis matrix B and coefficient matrix X .

$B(:, 2)$ of basis matrix B and the second row $X(2, :)$. Then the intensity image of the H and E components are represented in RGB space from corresponding OD images by applying Beer-Lambert Law in Eq. (1). The reconstructed intensity images (H and E components) in RGB space are shown in Fig.3.

After SNMF decomposition and reconstruction, the H and E components are expected to be discriminable and the reconstructed H component better represents the nuclei. Figure 4 shows the procedure of nuclear segmentation with the SNMF based approach. The segmentation procedure involves separating the H component from the E component via the SNMF based approach. Subsequently the intensity image of the reconstructed H component is transformed on to a binary image. Morphological operations are subsequently employed for separating connected nuclei and nuclear contour were extracted.

3. Experimental Setup and Design

To show the effectiveness of our new approach, we compare SNMF with PCA [18], ICA [15], CD[25], and NMF [24] in color unmixing of the following three data sets in Table 1:

Figure 3: The reconstruction of H and E components in RGB space by decomposed basis matrix B and coefficient matrix X . The upper row shows the multiplication of $B(:, 1)$ and $X(1, :)$ producing an OD image of the H stained component. The intensity image of the H stained component can be obtained by applying Beer-Lambert Law in Eq. (1). Similarly, the reconstruction of the E stained component is shown in the bottom row.

3.1. Experimental design and comparative strategies

3.1.1. Evaluating the effectiveness of smoothness and sparseness constraints on color unmixing with the SNMF algorithm

The aim of this experiment was to analyze the effects of smoothness and sparseness constraints in Eq. (9) on color unmixing of histopathological images with the SNMF algorithm. The experiments are conducted on the IHC images from D1 for decomposing the brown DAB component.

3.1.2. SNMF versus four other representation learning based approaches for color unmixing

For D1, since there is no ground truth for what constitutes an ideal “color unmixing” results, we only provide a qualitative comparison of SNMF and PCA, ICA, CD, and NMF in separating DAB components from original IHC and H & E stained images. For D2 and D3 we implicitly compare the “col-

Datasets	Datset 1 (D1)	Datset 2 (D2)	Datset 3 (D3)
Breast Cancer Type	invasive	lymph node-negative and estrogen receptor-positive	human epidermal growth factor receptor-2
Modality	IHC CD31, CD34, CD105 and H & E	H & E	H & E
Digital Scanner	high resolution whole slide scanner Aperio ScanScope digitizer		
Resolution	40X magnification		
N. of Patients in the studies	2	49	47
Image generation	ROIs chosen from cancerous regions in 6 WSI	ROIs chosen from cancerous regions in each WSI	ROIs chosen from lymphocytes clusters regions in WSI
N. of Images	36	200	100
Size (pixels)	800×800	800×800	100×100
Purpose	decomposing the brown DAB component	nuclei segmentation	lymphocytes segmentation
Groundtruth generated by pathologists	None	each nucleus in randomly chosen ROIs(400×400) is manually delineate	5 lymphocytes randomly chosen from each images were delineated

Table 1: The description of datasets studied in the paper. ROI and WSI represent “Region of Interest” and “Whole Slide Images”, respectively, in the table.

or unmixing” results obtained via SNMF and PCA, ICA, CD, and NMF in terms of the ability of segmentation in conjunction with the corresponding unmixed color representations to detect and segment nuclei and lymphocytes. The resultant intensity image of the H component in the RGB space is transformed into a gray scale image and then finally into a binary image. Finally, mathematical morphologic operations are employed for separating the connected nuclei and extracting the nuclear contours. Fig.4 illustrates the procedure of SNMF based nuclei/lymphocytes segmentation on H & E stained images.

3.2. Performance Evaluation

The quantitative performance of SNMF and PCA, ICA, CD, and NMF for segmentation accuracy of nuclei/lymphocyte regions are analyzed by using

Figure 4: The flowchart of SNMF based approach for nuclei segmentation. The original image (a) is decomposed into H stained component (b) with SNMF algorithm. The final segmentation results (d) is generated from binary image (c) after post-processing operation being conducted on the decomposed H stained component (b).

region-based measures:

$$\text{Dice Coefficient (DICE)} = \frac{2 \times |\mathcal{A}(S) \cap \mathcal{A}(G)|}{|\mathcal{A}(S)| + |\mathcal{A}(G)|}, \quad (10)$$

$$\text{Overlap (OL)} = \frac{|\mathcal{A}(S) \cap \mathcal{A}(G)|}{|\mathcal{A}(S) \cup \mathcal{A}(G)|}, \quad (11)$$

$$\text{Sensitivity(SN)} = \frac{|\mathcal{A}(S) \cap \mathcal{A}(G)|}{|\mathcal{A}(G)|}, \quad (12)$$

$$\text{Specificity (SP)} = \frac{|C - \mathcal{A}(S) \cap \mathcal{A}(G)|}{|C - \mathcal{A}(G)|}, \quad (13)$$

$$\text{Positive Predictive Value (PPV)} = \frac{|\mathcal{A}(S) \cap \mathcal{A}(G)|}{|\mathcal{A}(S)|}. \quad (14)$$

Here C is the total number of pixels in the image or region of interest (ROI) and $|s|$ represents the cardinality of any set s . $\mathcal{A}(S)$ and $\mathcal{A}(G)$ is the area of the segmentation results by automated and manual segmentation, respectively. The values shown in the Table 2 are average of Dice, OL, SN, SP, and PPV on nuclei/lymphocytes segmentation of all images in D2 and D3 with respect to manual segmentation.

4. Experimental results and discussion

4.1. Qualitative results

The effect of the sparseness term in (9) on color unmixing of IHC images in D1 are illustrated in Fig. 5. The first and second rows of Fig. 5 show the decomposed DAB and H components as the weight β for the sparseness term changes. When $\beta = 0$, as in Fig. 5 (a), the DAB component is not well separated within the original image (see Fig.7(a)). When the sparseness term is included, pure stains of DAB and H components are more easily separable (see Figs.5 (b-d) and (f-h)). Also, we can see that as the value of β increases, the noisy background of the DAB components are removed and the pure color of nuclei component are enhanced because of the effect of sparsity. The effects of the smoothness term in (9) on color unmixing are

Figure 5: The effect of sparseness term on color unmixing of a CD31 IHC and H&E image (Figure 7 (a)) via SNMF. The decomposed DAB (a-d) and H (e-h) components are illustrated when $\beta = 0$ (a,e), $\beta = 0.5$ (b,f), $\beta = 1.5$ (c,g), and $\beta = 2.5$ (d,h).

illustrated in Fig. 6. The images in the first and second rows of the figure show the corresponding decomposed DAB and H components as the weights

α_1 and α_2 of the smoothness term changes. When the smoothness is ignored ($\alpha_1 = \alpha_2 = 0$), and as the images and the corresponding magnified regions in Figures 6 (a) and (b) show, the decomposed components are noisy. However, when the smoothness terms are restored ($\alpha_1 = \alpha_2 = 2.5$), the resultant images (see Figs. 6 (b) and (c)) are smoothed and the most of the noise eliminated. The comparison of SNMF and PCA, ICA, CD, and NMF in

Figure 6: The effect of smoothness term on color unmixing of a CD31 IHC and H&E image (Figure 7 (a)) via SNMF. The decomposed DAB (a,c) and H (b,d) components when $\alpha_1 = \alpha_2 = 0$ (a,b) and $\alpha_1 = \alpha_2 = 2.5$ (c,d).

separating the DAB component from the original IHC images with different stains are shown in Fig.7. The first row in Fig.7 represent the original IHC CD31, CD34, and CD105 stained images. The decomposed DAB components obtained respectively via PCA, ICA, CD, NMF, and SNMF are shown in the 2nd to 6th rows in Fig.7. The results demonstrate the superiority of SNMF over PCA, ICA, CD, and NMF in separating the DAB component.

The nuclei/lymphocytes segmentation on one image from D2 (Fig.8 (a)) and D3 (Fig.8 (b)) using the different color unmixing strategies are shown in Fig.8. The 1st and 3rd rows correspond to the H component identified via the 5 different color unmixing strategies. The 2nd and 4th rows correspond

Figure 7: The DAB component of color unmixing results on a CD 31 (a), CD34 (b), and CD105 (c) IHC and H&E stained images via PCA (d,e,f), ICA (g,h,i), CD (j,k,l), NMF (m,n,o), and SNMF (p,q,r), respectively.

to the segmentation results based on the unmixed H component obtained from an H & E stained image. SNMF was able to achieve more accurate segmentation compared to PCA, ICA, CD, and NMF.

The results in Figure 8 further illustrate the importance of accurate color unmixing for nuclear and lymphocyte segmentation. As it may be evidenced by Figures 8 (c-h) and 8 (h, k, r, u) the less accurate color unmixing obtained via PCA, ICA, CD, and NMF results in correspondingly worse nuclear segmentations compared to our SNMF based approach (Figure 8 (p, q, r)).

4.2. Quantitative results

The quantitative evaluations of different approaches are conducted on manually annotated 1400 nuclei and 500 lymphocytes across all images in D2 and D3, respectively. The quantitative evaluation results on nuclei/lymphocyte segmentation of SNMF and comparison approaches are shown in Table 2. S-NMF showed better segmentation accuracy in terms of Dice and PPV on D2 as well as Dice and OL on D3.

		Dice (%)	OL (%)	SN (%)	SP (%)	PPV (%)
PCA	D2	0.70	0.55	0.65	0.99	0.78
	D3	0.73	0.59	0.63	0.99	0.90
ICA	D2	0.56	0.42	0.52	0.99	0.76
	D3	0.73	0.59	0.67	0.99	0.87
CD	D2	0.71	0.55	0.61	0.99	0.74
	D3	0.68	0.52	0.56	0.99	0.87
NMF	D2	0.75	0.60	0.76	0.98	0.71
	D3	0.79	0.65	0.72	0.99	0.87
SNMF	D2	0.76	0.60	0.74	0.99	0.78
	D3	0.80	0.67	0.73	0.99	0.87

Table 2: The average of Dice, OL, SN, SP, and PPV for the automated segmentation of about 1400 nuclei from 200 images in D2 and 500 lymphocytes from 100 images in D3 compared respect to manual segmentation.

5. Concluding Remarks

This paper presents a Sparse Non-negative Matrix Factorization (SNMF) for color unmixing on breast histology. One of major objectives for color

Figure 8: The H channel of color unmixing results (c-g) and their segmentation results (h-l) on an image (a) from D2 via PCA (c,h), ICA (d,i), CD (e,j), NMF (f,k), and SNMF (g,l) based approaches. The H channel of color unmixing (m-q) and the corresponding segmentation results (r-v) on an image (b) from D3 via PCA (m,r), ICA (n,s), CD (o,t), NMF (p,u), and SNMF (q,v).

unmixing is nuclei segmentation. As we show in Fig. 4, for quantitative evaluation on the performance of color unmixing by different approaches, the separated H component is leveraged for further segmenting the nuclei. The qualitative and quantitative evaluation on separating brown diaminobenzidine (DAB) component on immunohistochemical images and automated segmentation of nuclei and lymphocytes on H & E images suggest that SNMF can better separate individual stain components for histological images as compared with Principle Component Analysis, independent Component Analysis, Color Deconvolution, and Non-negative Matrix Factorization based approaches.

There are some limitations in this work. One limitation is that we can't quantitatively evaluate the effectiveness of color unmixing by SNMF against other approaches because of lacking ground truth. We alternatively choose to evaluate the segmentation accuracy for different approaches. In this application, color unmixing can be considered as a pre-processing step for nuclei/lymphocyte segmentation. Any advanced color-based nuclei segmentation methods can be used upon the unmixing results in post-processing step. Therefore, the accuracy of nuclei/lymphocytes segmentation by different approaches would dependent on the performance of both color unmixing for better separation of pure color of H component and subsequent segmentation module. This is the limitation of this evaluation strategy. Finally, the SNMF algorithm in the paper is only conducted on small region of interests from whole slide images. In the future work, we will apply SNMF algorithm for color unmixing problem in whole slide images.

Acknowledgement

This work is supported by National Natural Science Foundation of China (No. 61273259); Six Major Talents Summit of Jiangsu Province (No. 2013-XXRJ-019), and Natural Science Foundation of Jiangsu Province of China (No. BK20141482); the National Cancer Institute of the National Institutes of Health under award numbers (R01CA136535-01, R01CA140772-01, R21CA167811-01, R21CA179327-01; R21CA195152-01); the National Institute of Diabetes and Digestive and Kidney Diseases under award number R01DK098503-02; the DOD Prostate Cancer Synergistic Idea Development Award (PC120857); the DOD Lung Cancer Idea Development New Investigator Award (LC130463); the DOD Prostate Cancer Idea Development Award; the Ohio Third Frontier Technology development Grant; the CTSC

Coulter Annual Pilot Grant; the Case Comprehensive Cancer Center Pilot Grant VelaSano Grant from the Cleveland Clinic; the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University.

References

- [1] Ali, S., Madabhushi, A., 2012. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. *Medical Imaging, IEEE Transactions on* 31, 1448–1460. doi:10.1109/TMI.2012.2190089.
- [2] Basavanthally, A., Feldman, M.D., Shih, N., Mies, C., Tomaszewski, J.E., Ganesan, S., Madabhushi, A., 2011. Multi-field-of-view strategy for image-based outcome prediction of multi-parametric estrogen receptor-positive breast cancer histopathology: Comparison to oncotype dx. *J Pathol Inform* 2, Suppl S1:1. doi:10.4103/2153-3539.92027.
- [3] Basavanthally, A., Ganesan, S., Agner, S.C., Monaco, J.P., Feldman, M.D., Tomaszewski, J.E., Bhanot, G., Madabhushi, A., 2010. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on Biomedical Engineering* 57, 642–653.
- [4] Basavanthally, A., Ganesan, S., Feldman, M.D., Shih, N., Mies, C., Tomaszewski, J.E., Madabhushi, A., 2013. Multi-field-of-view framework for distinguishing tumor grade in er+ breast cancer from entire histopathology slides. *Biomedical Engineering, IEEE Transactions on* 60, 2089–99.
- [5] Belien, J.A., Somi, S., de Jong, J.S., van Diest, P.J., Baak, J.P., 1999. Fully automated microvessel counting and hot spot selection by image processing of whole tumour sections in invasive breast cancer. *Journal of Clinical Pathology* 52, 184–192.
- [6] Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35, 1798–1828.

- [7] Chang, H., et al., 2013. Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. *Medical Imaging, IEEE Transactions on* 32, 670–682. doi:10.1109/TMI.2012.2231420.
- [8] Doyle, S., Feldman, M.D., Tomaszewski, J.E., Madabhushi, A., 2012. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE transactions on bio-medical engineering* 59, 1205–18.
- [9] Fang, R., Chen, T., Sanelli, P.C., 2013. Towards robust deconvolution of low-dose perfusion ct: Sparse perfusion deconvolution using online dictionary learning. *Medical Image Analysis* 17, 417 – 428.
- [10] Fang, R., Karlsson, K., Chen, T., Sanelli, P.C., 2014a. Improving low-dose blood-brain barrier permeability quantification using sparse high-dose induced prior for patlak model. *Medical Image Analysis* 18, 866 – 880.
- [11] Fang, R., Sanelli, P., Zhang, S., Chen, T., 2014b. Tensor total-variation regularized deconvolution for efficient low-dose ct perfusion, in: *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2014*. Springer. volume 8673 of *Lecture Notes in Computer Science*, pp. 154–161.
- [12] Filipczuk, P., et al., 2013. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *Medical Imaging, IEEE Transactions on* 32, 2169–2178. doi:10.1109/TMI.2013.2275151.
- [13] Gavrilovic, M., JC, A., J., L., C., W., Bengtsson E, B. C, C.I., 2013. Blind color decomposition of histological images. *Medical Imaging, IEEE Transactions on* 32, 983–994.
- [14] Huang, J., Zhang, S., Li, H., Metaxas, D., 2011. Composite splitting algorithms for convex optimization. *Computer Vision and Image Understanding* 115, 1610 – 1622.
- [15] Hyvarinen, A., Karhunen, J., Oja, E., 2004. Independent component analysis. volume 46. John Wiley & Sons.

- [16] Ingle, J., Crouch, S., 1988. Spectrochemical analysis. Englewood Cliffs, NJ , 106.
- [17] Janowczyk, A., Chandran, S., Singh, R., Sasaroli, D., Coukos, G., Feldman, M., Madabhushi, A., 2012. High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts. Biomedical Engineering, IEEE Transactions on 59, 1240–1252. doi:10.1109/TBME.2011.2179546.
- [18] Jolliffe, I., 2005. Principal component analysis. Wiley Online Library.
- [19] Khan, A., Rajpoot, N., Treanor, D., Magee, D., 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. Biomedical Engineering, IEEE Transactions on 61, 1729–1738.
- [20] Kim, H., Park, H., 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23, 1495–1502.
- [21] Li, G., Sanchez, V., Patel, G., Quenby, S., Rajpoot, N., 2014. Localisation of luminal epithelium edge in digital histopathology images of ihc stained slides of endometrial biopsies. Computerized Medical Imaging and Graphics , 1729 – 1738.
- [22] Mete, M., Hennings, L., Spencer, H.J., Topaloglu, U., 2009. Automatic identification of angiogenesis in double stained images of liver tissue. BMC bioinformatics 10, S13.
- [23] Qi, X., et al., 2012. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. Biomedical Engineering, IEEE Transactions on 59, 754–765. doi:10.1109/TBME.2011.2179298.
- [24] Rabinovich, A., , Agarwal, S., Laris, C., Price, J.H., Belongie, S., 2003. Unsupervised color decomposition of histologically stained tissue samples, in: Advances in Neural Information Processing Systems.
- [25] Ruifrok, A.C., Johnston, D.A., 2001. Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23, 291–299.

- [26] Veta, M., et al., 2013. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. PLoS ONE 8, e70221.
- [27] Veta, M., Kornegoor, R., Huisman, A., Verschuur-Maes, A., Viergever, M.A., Pluim, J., van Diest, P., 2012. Prognostic value of automatically extracted nuclear morphometric features in whole slide images of male breast cancer. *Modern Pathology* 25, 1559–1565.
- [28] Veta, M., Pluim, J.P., van Diest, P.J., Viergever, M.A., 2014. Breast cancer histopathology image analysis: A review. *Biomedical Engineering, IEEE Transactions on* 61, 1400–1411.
- [29] Wang, G., Zhang, S., Xie, H., Metaxas, D.N., Gu, L., 2015. A homotopy-based sparse representation for fast and accurate shape prior modeling in liver surgical planning. *Medical Image Analysis* 19, 176 – 186.
- [30] Wu, G., Wang, Q., Zhang, D., Nie, F., Huang, H., Shen, D., 2014. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Medical Image Analysis* 18, 881 – 890.
- [31] Yu, Y., Zhang, S., Li, K., Metaxas, D., Axel, L., 2014. Deformable models with sparsity constraints for cardiac motion analysis. *Medical Image Analysis* 18, 927 – 937.
- [32] Zhang, S., Huang, J., Metaxas, D., Wang, W., Huang, X., 2010. Discriminative sparse representations for cervigram image segmentation , 133–136.
- [33] Zhang, S., Zhan, Y., Dewan, M., Huang, J., Metaxas, D.N., Zhou, X.S., 2012a. Towards robust and effective shape modeling: Sparse shape composition. *Medical Image Analysis* 16, 265 – 277.
- [34] Zhang, S., Zhan, Y., Metaxas, D.N., 2012b. Deformable segmentation via sparse representation and dictionary learning. *Medical Image Analysis* 16, 1385 – 1396.
- [35] Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S., 2015a. Towards large-scale histopathological image analysis: Hashing-based image retrieval. *IEEE Transactions on Medical Imaging* 34, 496–506.

- [36] Zhang, X., Liu, W., Zhang, S., 2014. Mining histopathological images via hashing-based scalable image retrieval, in: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, pp. 1111–1114. doi:10.1109/ISBI.2014.6868069.
- [37] Zhang, X., Su, H., Yang, L., Zhang, S., 2015b. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition.
- [38] Zheng, Y., Daniel, E., III, A.A.H., Xiao, R., Gao, J., Li, H., Maguire, M.G., Brainard, D.H., Gee, J.C., 2014. Landmark matching based retinal image alignment by enforcing sparsity in correspondence matrix. *Medical Image Analysis* 18, 903 – 913.