Automatic Analysis Method of Protein Expression images based on Generalized Data Field

Shuliang Wang
International School of Software
Wuhan University
Wuhan, China
slwang2005@gmail.com

Ying Li
State Key Laboratory Engineering in Surveying, Mapping and Remote Sensing
Wuhan University
Wuhan, China
lyljhappy@163.com

Wenchen Tu
International School of Software
Wuhan University
Wuhan, China
tuwc369@hotmail.com

Peng Wang
International School of Software
Wuhan University
Wuhan, China
wangpengtoo@gmail.com

***Abstract*—For detection of protein expression in biomedical image, shape measurement of protein expression mostly depends on semi-automatic analysis of image analysis software which makes the results vulnerable to subjective factors, since the automatic analysis is too complicated to operate. Therefore, a novel algorithm based on generalized data field (GDF) is proposed to determine the region of protein expression. Instead of being directly divided into the measured object and background, all the data objects, namely pixels of an image, are naturally clustered into multiple classes based on potential distribution in generalized data field. Each class represents protein expression in different degree, which precisely describes the details of protein expression. Compared with image-pro plus software analysis, KM and EM, experiment results demonstrate that the protein expression can be extracted easily and objectively from an image by GDF. Furthermore, noises of background are eliminated by the smoothing procedure of GDF.**

Keywords-detection of protein expression; clustering; generalized data field

I. INTRODUCTION

Protein Expression is a molecular biology technique applying model organism such as bacteria, yeast, animal cell or plant cell to express foreign genes proteins. And image processing software, such as quantity one, image J, image-pro plus, is further applied to carry on quantitative analysis of protein expression. No matter which software is chosen, the first step is to analyze pixels and detect the right parts of the image reflecting measured object, namely finding the areas of protein expression. Because the automatic analyzing procedure of image processing software is quite complex, shape measurement of protein expression are usually based on semi-automatic analysis which makes the results suffer from subjective factors and simultaneously lower the quality of quantitative analysis. To tackle these problems, we propose a novel clustering algorithm GDF. The proposed algorithm can straight forwardly and automatically recognize the marked areas of protein expression image meanwhile reducing artificial error and improving the quality of analysis.

In the following sections, section II substantially introduces the background of generalize data field and section III specifically describes the basic process of the proposed algorithm. In section IV, several experiments are conducted to reveal that GDF algorithm is highly useful to determine area of protein expression.

II. BACKGROUND

Inspired by the idea of physical fields, mutual interaction between the particles of matters and its description method are introduced to the abstract number field space. In space $\Omega \subseteq R^P$, let data set $D=\{X_1, X_2, \ldots, X_n\}$ denote a P-dimensional independent random sample, where data object $X_i = (X_{i1}, X_{i2}, \ldots, X_{iP})^T$ with $i = 1, 2, \ldots, n$. Each data object X_i is viewed as a particle or a nucleon with certain mass m_i .Thus, a virtual field exists around these objects and any data object is affected by all the other objects in the field. The field existing in the data space is called data field[1][2]. For any given point x in the data field, the potential value is defined as

$$\hat{\varphi}(x) = \frac{1}{n\sigma}\sum_{i=1}^{n} m_i \times K\left(\frac{x-x_i}{\sigma}\right). \quad (1)$$

Where $K(x)$ is the unit potential function, σ is an impact factor and m_i is the mass of data object x_i.

From (1), it can be known that the potential value equation is similar to the equation of kernel density estimation. In order to overcome some difficulties from proving the feasibility of the data field theory, by making a reference to the conditions of kernel density estimation, the unit potential function $K(x)$ must satisfy the following conditions: $\int K(x)\mathrm{d}x = 1$, $\int xK(x)\mathrm{d}x = 0$ and $0 < R(K) = \int K(x)^2\, dx < \infty$. In addition, the mass m_i with $i = 1,2,\cdots,n$ satisfies

$$\sum_{i=1}^{n} m_i = 1, m_i \geq 0 \; and \; \lim_{n\to\infty} n \sup_{1\leq i\leq n}\{m_i\} = 1.$$

For the existent multi-dimensional data, considering that the impact factor σ should be anisotropic, it is replaced by a matrix H. In this case, for any given point x in space, the potential value estimation is defined as:

$$\hat{\varphi}(x) = \frac{1}{|H|}\sum_{i=1}^{n} m_i \times K\left(H^{-1}(x - X_i)\right). \quad (2)$$

Where $K(x)$ is a multivariate unit potential function and H is a positive definite $p\times p$ matrix that is a non-singular constant matrix.

III. ALGORITHM DESCRIPTION

Basically, the outline of the proposed algorithm can be developed as the following. Firstly, we divide hierarchical grid in feature space for the succeeding procedures. Then by clustering partitioned cells with GDF, data objects are divided into different clusters, specifically descripted in the parts B and C of this section. Finally, clusters are merged into two classes: measured object and background.

A. Potentioal value estimation

Using hierarchical grid structure in feature space, we propose a new technique for the potential value estimation. In feature space Ω, divide $2N \times 2N \times 2N$ and calculate the mean of data points located in each cell as the feature value of the cell and the mean of the corresponding spatial coordinates as spatial coordinate of the cell. Thus a new feature space Ω_s is formed based on feature space Ω. Then merge every eight adjacent cells into a cell which forms the second grid structure and, similarly, obtain another new feature space Ω_b. For feature space Ω_b, let $x_{i_u j_u k_u}$ denote the feature of the (i_{th}, j_{th}, k_{th}) cell with $x_{i_u j_u k_u} = \left(x^{(1)}_{i_u j_u k_u}, x^{(2)}_{i_u j_u k_u}, x^{(3)}_{i_u j_u k_u}\right)$ and $w_{i_u j_u k_u}$ is the quality of (i_{th}, j_{th}, k_{th}) cell. Thus for $\forall\, f = \left(x^{(1)}, x^{(2)}, x^{(3)}\right)^T \in \Omega_s$, the potential value estimation (1) can be rewritten as

$$\hat{\varphi}(f)_G = \sum_{\Omega_b} w_{i_u j_u k_u} \times K\left(\frac{x - x_{i_u j_u k_u}}{\sigma}\right). \quad (3)$$

Where $w_{i_u j_u k_u} = Q_{i_u j_u k_u}$, $Q_{i_u j_u k_u}$ is the quantity of points located in (i_{th}, j_{th}, k_{th}) cell. To improve accuracy, the impact factor σ should be different in each dimension and correspondingly, multivariate unit potential function K is defined as the product of three one-dimensional unit potential functions. Thus the equation (11) becomes

$$\hat{\varphi}(f)_G = \sum_{\Omega_b} \prod_{j=1}^{3} w_{i_u j_u k_u} K\left(\frac{x^{(j)} - x^{(j)}_{i_u j_u k_u}}{\sigma_j}\right). \quad (4)$$

Where $\sigma = \left(\sigma_1, \sigma_2, \sigma_3\right)^T$, $\sigma_i, i = 1,2,3$ denotes the i^{th} dimensional impact factor.

As the potential value estimation is similar to kernel density estimation, improvements can be obtained by setting the impact factorσ as a multiple of the window width h, that is $\sigma = ch = c\left(h_1, h_2, h_3\right)^T$ where c is the proportionality coefficient and $h = \left(h_1, h_2, h_3\right)^T$ is the window width of the kernel estimation. The window width h is calculated with Sheather-Jones' plug-in method in feature space Ω_s [3][4]. The proportionality coefficient c can be self-tuned to obtain different levels of image segmentation.

B. Clustering mehod

We also propose a new clustering algorithm based on hierarchical grid-based potential estimation to improve the running speed, which is called downhill method. Contrary to the hill-climbing clustering, the first step is to detect the modes of potential estimation (maximal point) which are located among the zeros of the gradient, that is, $\nabla\hat{\varphi}(x) = 0$. The gradient of potential value estimation can be obtained by calculating partial derivative of (4),

$$\nabla\hat{\varphi}(f) = \left(\frac{\partial\hat{\varphi}(x)}{\partial x^{(1)}}, \frac{\partial\hat{\varphi}(x)}{\partial x^{(2)}}, \frac{\partial\hat{\varphi}(x)}{\partial x^{(3)}}\right)^T. \quad (5)$$

Then searching along each mode to find the clusters until the gradient is no longer rising, and finally, all cells are clustered to one of the modes. This procedure is an elegant way to locate these zeros without estimating the potential value and simplify the searching process without repeated work compared with the hill-climbing.

To acquire area of protein expression, we need to merge clusters into two classes, that is, measured object and background. Consequently, we can determine the distance between each cluster and black, and meanwhile, adaptively set distance threshold T to merge clusters.

C. Description of the proposed algorithm procedure

In this paper, we selected the space L*u*v* as the feature space motivated by a linear mapping property to pixels classification [5][6][7].

The algorithm can be described in following steps:

1) Divide hierarchical grids and then calculate the impact factor $\sigma = ch = c\left(h_1, h_2, h_3\right)^T$ where the proportionality coefficient c is the algorithm parameter and the window width h is obtained with Sheather-Jones' plug-in method. The initial values of N and c are respectively set to 10 and 2.0. For color image, RGB color space should be transformed into L*u*v* color space in the first instance.
2) Calculate the derivative of each cell according to (5) in space Ω_s to find all the modes. Delineate the initial clusters $\{C_k\}_{k=1,\cdots,v}$ by grouping together all modes which are adjacent within six-neighborhood where C_k can contain more than a mode.
3) For each $k = 1, \cdots, v$, starting from modes in C_k as centers, search cells toward the direction of the gradient ascent until the gradient is not rising. The searched cells are added to C_k. As many cells are empty, the complexity of this step is relatively lower.
4) Merge clusters $\{C_k\}_{k=1,\cdots,v}$ into two classes by adjusting threshold T and Map classified points to spatial region of an image and merge those spatial regions which consist of less than M points.

IV. CASE STUDY

As for images of protein expression, the key point is to analyze pixels in the image and determine the part reflecting measured object (protein expression). Then by calculating the measurement parameters of the recognized object, the corresponding data is obtained to carry on statistical analysis on measured objects.

A. Analysis of the gray-scale images of protein expression detected by Western blots

The method of Western blots is applied in protein analyzing to study protein molecules in vitro, examining

whether the target protein is in the sample or the status of protein expression. Fig. 1(a) is a picture showing protein expression. The proposed algorithm can be used to easily and automatically recognize each lane. Besides, the interference of background noise can be eliminated to make sure its correction during the identifying process, displayed in Fig. 1. From Fig. 1(b), it can be known that each lane is partitioned into dark gray part and light gray part because of differences of gray value, which results in two different equipotential lines. Thus we can choose an appropriate equipotential line to recognize lanes based on actual requirements.

B. Analysis of the color images of protein expression

In Fig.2-1, we show two groups of VASP proteins (red) expression and f-action proteins (green) expression in patients with cancer. Fig. 2-1(a) and Fig. 2-1(b) are one group while the other group is Fig. 2-1(c) and Fig. 2-1(d). The proposed algorithm clusters pixels in original images of protein expression and the clustering results are shown in Fig. 2-2, in which the numbers of clusters in Fig. 2-2(a), Fig. 2-2(b), Fig. 2-2(c), Fig. 2-2(d) are 17, 21, 18, 20, respectively. It can be seen that clustering results accurately reflect the real distribution of protein expression since we select the larger parameter N of grid division, that is, N=20. And then, the clusters are merged to obtain the area of protein distribution. To get satisfied results shown in Fig. 2-3(a), Fig. 2-3(b), Fig. 2-3(b), Fig. 2-3(d), the user can adaptively adjust the threshold T, a significant parameter in the procedure of merging clusters, according to their own requirements. Results in Fig. 2-3 basically accord with those in original images.

KM[8][9] and EM[10] have been successfully applied to medical images processing. In Fig. 2-4 and Fig. 2-5, we use KM method and EM method respectively to acquire the area of protein expression. Compared with the original image, it can be found that the area obtained by EM has a tendency to be excessively larger. For example, as a matter of fact, f-action protein does not exist in the nucleus region in Fig. 2-1(b) and Fig. 2-1(d), nevertheless, the results with EM contains the region of the nucleus, as shown in Fig. 2-5(b) and Fig. 2-5(d). On the contrary, areas extracted by EM are too small to represent the protein expression in some cases, though some experimental results are similar to the real protein expression. For example, the protein expression area abstracted by KM in Fig. 2-4(a) is apparently not in accord with that in Fig. 2-1(a).

Their accuracy of the clustering results obtained were evaluated by measuring the ratio between number of correctly classified pixels and total number of pixels compared with the ground truth performed by an expert. That is, their performance of clustering is measured as

$$Accuracy=\frac{number\ of\ correctly\ classified\ pixels}{total\ number\ of\ pixels}. \quad (6)$$

In order to check sensitivity and accuracy to parameter choice of the proposed algorithm, the parameter c is set to 1, 1.5, 2 and 2.5 to obtain clustering results for calculating accuracy according to (6). Comparison with the ground truth in Table I and Fig. 3 suggests the proposed algorithm can provide high accuracy for an ample parameter range. And to assess performance of different algorithms objectively, highest accuracy is calculated for different algorithms and the results are listed in Table II, which show that GDF is better than KM and EM.

From the analysis above, it can be concluded that GDF not only has simpler operations but also can accurately examine the shape of measuring objects, providing accurate measuring data for subsequent statistical analysis in accordance with users' own requests.

V. CONCLUSION

In this paper, we propose a clustering algorithm based on generalized data field (GDF) to obtain the shape of various measuring objects, for example, protein expression or nucleus. The data objects are naturally clustered according to the potential distribution in field. According to their own requests, the users merge the clusters into two groups, one is measuring objects and the other is background. Compared with Image-pro plus software, KM method and EM method, the algorithm based on GDF operate easily and automatically recognize the measured object. Moreover, it can eliminate background noise's interference and improve the accuracy of determining the measured object.

ACKNOWLEDGMENT

We would like to thank Ke Su for use of her images and discussion about experiment results. We would also like to thank JinFei Yin and QiLiang Chen for revising the code of the proposed algorithm.

REFERENCES

[1] D. Li, S. L. Wang, D. Y. Li, Spatial data mining theories and applications, Science Press, 2006.

[2] S. L. Wang, W. Y. Gan, D. Y. Li, D. R. Li, "Datafield for hierarchical clustering," International Journal of Data Ware housing and Mining, vol. 7, pp. 43-63,2011.

[3] M. C. Jones, J. S. Marron, S. J. Sheather, "A brief survey of bandwidth selection for density estimation," Journal of the American Statistical Association, vol. 91, pp. 401-407, 1996.

[4] S. J. Sheather, M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," Journal of the Royal Statistical Society, vol. 53, pp. 683-690, 1991.

[5] Gaurav Sharma, H. Joel TRussel, "Digital color imaging," IEEE transactions on image processing, vol. 6, no. 7, pp. 901-932, 1997.

[6] CIE Colorimetry, CIE Pub. no. 15. 2, Center. Bureau CIE, Vienna, Austria, 1986.

[7] M. Mahy, L. Van Eyckden, A. Oosterlinck, Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV, Color Res. Appl. , vol. 19, no. 2, pp. 105-121, Apr. 1994.

[8] C. W. Chen,J. B. Luo, K. J. Parker, "Image segmentation via adaptive mean clustering and knowledge-based morphological operations with biomedical applications," IEEE Transactions on Image Processing, vol. 7, no. 12, pp. 1673-1683, 1998.

[9] H. M. Moftah, A. E. Hassanien, M. Shoman, "3D brain tumor segmentatin scheme using K-mean clustering and connected component labeling algorithms," Proc. Intelligent Systems Design and Applications(ISDA 10), IEEE Press, Dec. 2010, pp. 320-324, doi: 10.1109/ISDN.2010.5687244.

[10] N. M. Phuong, N. X. Vinh, "Normalized EM algorithm for tumor clustering using gene expression data," Proc. IEEE BioInformatics and BioEngineering (BIBE 08), IEEE Press, Dec, 2008, pp. 1-7, doi:10.1109/BIBE. 2008. 4696683.

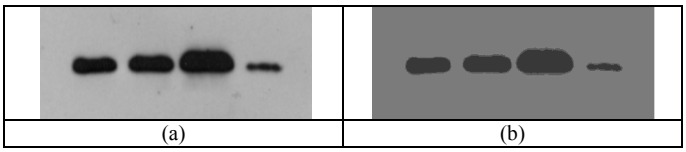

Figure 1. Gray image analysis of protein expression. (a) is an original image and (b) is its corresponding result.

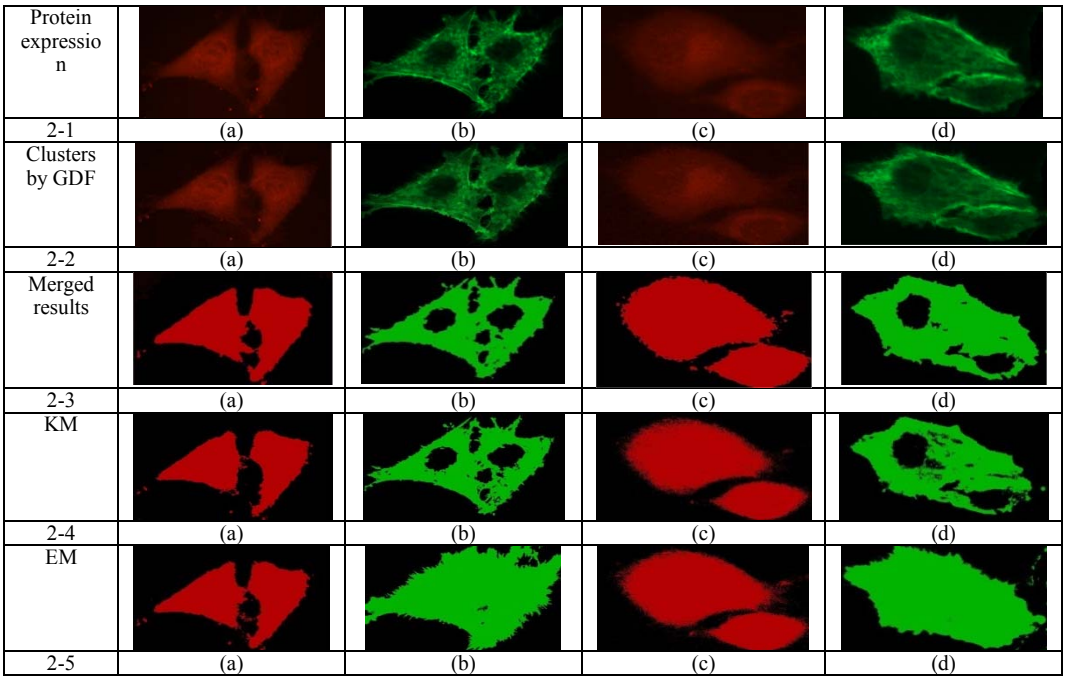

Figure 2. Measured protein expression with different methods. The original four images of protein expression are shown in 2-1. Natural clusters and merged results of GDF are respectively displayed in 2-2 and 2-3. 2-4 and 2-5 are correspondingly measured results by using KM and EM, where c=2.0.

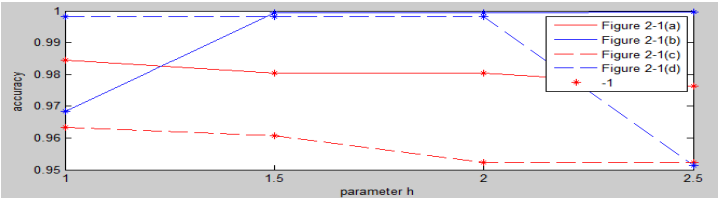

Figure 3. Clustering accuracy with different parameter h.

TABLE I. CLUSTERING ACCURACY (%)

GDF Parameters	***Figure 2-1 (a)***	***Figure2-1(b)***	***Figure 2-1(c)***	***Figure 2-1(d)***
GDF(c=1)	98.44	96.85	96.34	99.83
GDF(c=1.5)	98.04	99.94	96.07	99.83
GDF(c=2.0)	98.04	99.94	95.23	99.83
GDF(c=2.5)	97.64	99.99	95.23	95.13

TABLE II. ACCURACY COMPARISION (%)

Methods	***Figure 2-1 (a)***	***Figure2-1(b)***	***Figure 2-1(c)***	***Figure 2-1(d)***
GDF	98.44	99.99	96.34	99.83
KM	91.36	95.68	95.94	94.92
EM	94.46	81.05	95.10	84.66