An Artificial Neural Network Model for Multi Dimension Reduction and Data Structure Exploration

Chee Siong Teh, Ming Leong Yii & Chwen Jen Chen
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak (UNIMAS)
Kota Samarahan, Sarawak, Malaysia
e-mail: csteh@fcs.unimas.my, yiimingleong@gmail.com, cjchen@fcs.unimas.my

***Abstract*—This paper proposes an hybrid Artificial Neural Network (ANN) with Self-Organizing Map (SOM) and modified Adaptive Coordinates (AC) for multivariate dimension reduction and data structures exploration. SOM, being a prominent unsupervised learning algorithm, is often used for multivariate data visualization. However, SOM only preserved input space inter-neurons distances and not in the output space because of SOM rigid grid. SOM grid provides little information for visual exploration of the clustering tendency of the multivariate data. Modified AC is therefore proposed to remove SOM's map rigidity and provides better data topology preserved visualization. Empirical study of the hybrid yielded promising topology preserved visualizations for synthetic and benchmarking datasets.**

Keywords- Self-Organizing Map; Adaptive Coordinates; multivariate data visualization; multi-dimension reduction

I. INTRODUCTION

Visual information is essential for human intuitive decision making. However, data with dimension higher than three is not possible to be visualized directly. Dimensionality reduction is required in order to visualize the underlying data structure. To do this, classical method such as Sammon's Non-linear Mapping (NLM) [1] and Multidimensional Scaling (MDS) [2] provide excellent ways. But due to their point-to-point mapping nature and calculation complexities, they are not practical for real life applications where databases are always expanding. Algorithms with data compression or vectors quantization and visualization abilities are more preferable in real life applications. Self-Organizing Map (SOM) proposed by Kohonen [3] met these requirements excellently. It since, became a standard analytical tool for many real world problems.

SOM can be used for dimension reduction, vector quantization and visualization. Recent application of SOM in real life problems can be found in [11-14]. Although SOM has become a very popular analytical tool, it has one major drawback due to it rigid grid used in the output space. Only the input space data topology is preserved. SOM's output space is represented by rectangular or hexagonal grid which obviously does not preserve the inter-neuron distances. The studies in [4-7] pointed out this drawback and proposed new algorithms that preserve the inter-neuron distances in the output space. Visualization induced SOM (ViSOM) [4] and Probabilistic Regularized SOM (PRSOM) [5] are two of the popular ones. Both of these algorithms introduced a regularization control parameter so that the distances between two neighborhood neurons can be controlled. By regularizing the inter-neuron distances of the input space with suitable control parameter, the output space of the projected map is able to preserve the data topology. However, large amount of neurons are required in order to produce accurate data visualization. Large number of neurons increases computation cost and the projected map becomes more vulnerable to dead neurons problem [8-9].

Adaptive Coordinates (AC) was proposed as an extension to the original SOM [6-7]. AC did not modify SOM, instead by using virtual adaptive units to mirror SOM neurons movement, AC is able to produce topology preserved output map. This is an advantage as compare to ViSOM and PRSOM in terms of number of neurons utilized. The details of AC algorithm can be found in [6-7]. Nevertheless, AC's projection ability is very much depended on a magic number or free parameter that triggers the starting to the adaptation process. During initial training, the adaptation tends to be too strong and will cause all adaptive units to move towards single point [7]. But if the adaptation starts too late, the remaining neurons weight vectors movements, before SOM converge, will be too little to produce meaningful visualization. This magic number can only be found heuristically. To lift this limitation, the original AC is modified and is hybrid with SOM to produce an algorithm for multivariate data dimension reduction and data visualization in this paper.

Section II gives an overview of SOM, and the modified AC is presented in section III. The proposed hybrid of SOM with modified AC is presented in section IV. Experiment results and discussions are presented in section V, and section VI concludes the work.

II. SELF-ORGANIZING MAP (SOM)

Kohonen's Self-Organizing Map (SOM) [3] has the desirable property of topology preservation, which captures an important aspect of the feature maps in the cortex of highly developed animas brains. It is widely used for projection of multivariate data, density approximation, and clustering. It has been successfully applied in the areas of speech recognition, image processing, robotics, telecommunication, and process control [3].

978-0-7695-3879-2/09 $26.00 © 2009 IEEE
DOI 10.1109/SoCPaR.2009.59

SOM network architecture basically consists of a two-dimensional array of units, each connected to all *n* input nodes with weights vector

$$\mathrm{w}_j = (w_{j1}, w_{j2}, ..., w_{jd})$$

where *d* is the dimension of the dataset being analysed. The learning process of SOM involved finding the Best Matching Unit (BMU), for each sample **x** drawn from dataset, using Euclidean norm

$$\mathrm{i}^* = \underset{\mathrm{i}}{\operatorname{argmin}} \left\| x\text{-}w_j \right\| \quad (1)$$

and updates the weights of all the connections according to the unsupervised "on-line" competitive learning rule

$$w_j(t+1) = w_j(t) + \eta . h_{i^* j}(t)\left[x(t) - w_j(t)\right] \quad (2)$$

Where η is the learning rate and $h_{i^* j}(t)$ is the Gaussian neighborhood function

$$h_{i^* j}(t) = \exp\left[-\frac{(r_{i^*} - r_j)^2}{2\sigma(t)^2} \right] \quad (3)$$

$(r_{i^*} - r_j)$ is the distance between BMU *i** and neuron *j* and

$$\sigma_\wedge(t) = \sigma_{\wedge 0} \exp\left(-2\sigma_{\wedge 0} \frac{t}{t_{\max}} \right) \quad (4)$$

is the neighborhood range with initial value $\sigma_{\wedge 0}$ which is initialized with the half of the SOM lattice size.

Although SOM provides good input space topology preservation, but due to the rigid predefined rectangular or hexagonal grid it uses on the output space, the SOM map gives little intuitive information about the data topology. SOM grid does not preserve the data topology. Various studies [4-7] pointed out this limitation.

III. Adaptive Coordinates (AC)

The Adaptive Coordinates (AC) [6] was proposed to removes the rigidity of SOM grid map, allowing a more intuitive recognition of the input data cluster boundaries. The basic idea of AC is to mirror the movement of neurons' weight vectors, in each of the SOM training iterations, into two dimension adaptive coordinates $\langle ax_i, ay_i \rangle$. For each iteration *i*, the distances between neurons weights before SOM's weight vectors adaptation $Dist_i(t)$ and after the adaptation $Dist_i(t+1)$ are used to compute the relative AC adaptation factor.

$$\Delta Dist_i(t+1) = \frac{Dist_i(t) - Dist_i(t+1)}{Dist_i(1)} \quad (5)$$

The adaptive coordinates, except the winner node, is then moved towards the winner neuron *c* according to the equations below:

$$\begin{aligned} ax_i(t+1) &= ax_i(t) + \Delta Dist_i(t+1).\left(ax_c(t) - ax_i(t)\right) \\ ay_i(t+1) &= ay_i(t) + \Delta Dist_i(t+1).\left(ay_c(t) - ay_i(t)\right) \end{aligned} \quad (6)$$

As highlighted in [8] and [9], AC suffered from inconsistent adaptive units movements due to the use of relative adaptation factor as shown in (5). The initial SOM training tends to be too strong and causes the movements of the adaptive units to fall into single point. Following that, after about one fifth of the predefined training epochs, when SOM is converging, there will be too little weight vectors movements to give notable mirroring movement of the adaptive units. It means that the remaining training epoch, after the SOM converged, will not improve the visual projection on the AC. It will simply waste of computational cost. This threshold value that triggers the starting of adaptation can only be found heuristically. This is not desirable because it reduced SOM robustness. Therefore to ensure a better visual projection on the AC and SOM, a few approaches to modify the original AC is proposed in this study. They are presented in section IV.

IV. Proposed Hybridization Of Som And Modified AC (SOM With Modified AC)

SOM's original algorithm is extremely robust. No parameter is required in order for it to produce good topological preserved map. But as highlighted in previous section, SOM does not preserve the inter-neuron distance in the output space due to the rigid grid. A modified AC is proposed in this paper to remove the rigidity so that a better topology preserved map can be produced.

In order to successfully hybrid modified AC into SOM while retaining SOM robustness, an extra set of coordinates $\langle ax_i, ay_i \rangle$ are used as the adaptive units. These adaptive units will be used to mirror the neurons' movement. To overcome the inconsistent movements of these adaptive units for every iteration, the input and output spaces are normalized so that all movements are within 0 to 1 scale. Equation (7) shows the modified adaptation factor.

$$\Delta Dist_i(t+1) = d_{out}(t) - d_{in}(t) \quad (7)$$

where $d_{out}(t)$ is the Euclidean's distance of adaptive coordinates in the output space and $d_{in}(t)$ is the Euclidean's distance of the respective neurons weights. Instead of mirroring directly the movements of neurons as proposed in [6] and [7], the modified adaptation factor will approximate the distances of neurons and their respective adaptive units. The polarity of the adaptation factor will determine whether the adaptive coordinates will be pull closer to the winner or push away from it through the coordinate update formula in (8).

$$\begin{aligned} ax_i(t+1) &= ax_i(t) + \Delta Dist_i(t+1).\sigma n_\wedge(t).\left(ax_c(t) - ax_i(t)\right) \\ ay_i(t+1) &= ay_i(t) + \Delta Dist_i(t+1).\sigma n_\wedge(t).\left(ay_c(t) - ay_i(t)\right) \end{aligned} \quad (8)$$

where $\sigma n_\wedge(t) = \sigma_{\wedge 0} \exp\left(-2\sigma_{\wedge 0} \frac{t}{t_{\max}} \right)$ is the adapted neighborhood range as in (4). Its value is exponentially decreasing between 1~0 but never reached zero or exceed 1. Since the proposed algorithm removed the need for threshold starting the adaptation, the adaptation process can

be started after SOM converged. This reduces overall computational cost.

The proposed algorithm can be summarized as follow:

Step 1: Find BMU for each sample selected according to (1)
Step 2: Update Codebook weights according to (2)
Step 3: Find the adaptation factor for a neuron according to (7)
Step 4: Update the adaptive units according to (8)
Step 5: Repeat Step 1 to 4 according to a pre-defined number of epochs.

V. EXPERIMENTS

The performance of the hybrid SOM with modified AC is first demonstrated using synthetic 2D and 3D Gaussians dataset. These datasets are simple and can be visualized directly by human observer. Then benchmarking Wine and Wincousin Breast Cancer (WBC) datasets [10] are used to test the dimension reduction and topology preserved projection ability for higher dimension. In all the experiments, the codebook vectors of SOM are initialized by random selection of samples from the dataset being evaluated while the adaptive coordinates $\langle ax_i, ay_i \rangle$ are initialized based on the SOM grid but with normalized values. The SOM lattice is set to 10x10 and the learning rate is set to be linearly decreasing from 0.9 to 0.01 for all experiments. The visualization results are compared with SOM, SOM with original AC [6-7], and ViSOM [4] for the benchmarking datasets. The results of ViSOM projections are adapted from [5].

A. Synthetic 2-D Gaussians

The synthetic 2-D dataset consists of three well separated Gaussians with 100 samples each. Their mean vectors are [2 3], [-4 2], and [0 -2] and covariance matrices are $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The 2D Gaussians input dataset is shown in Fig. 1 (a).

Visual inspection of the projected maps depicted that the proposed SOM with modified AC projection is much better than SOM, and SOM with original AC. SOM with modified AC removed the SOM rigid grid (Fig. 1 (c)) and produced data topology preserved output map as shown in Fig. 1 (b). Clustering tendency of the projected map in SOM with modified AC is much better as compared with the original AC as in Fig. 1 (d). Three distinct clusters that resemble very closely to the original dataset's clusters are revealed in the projection of SOM with modified AC. This is very useful information for data clustering process.

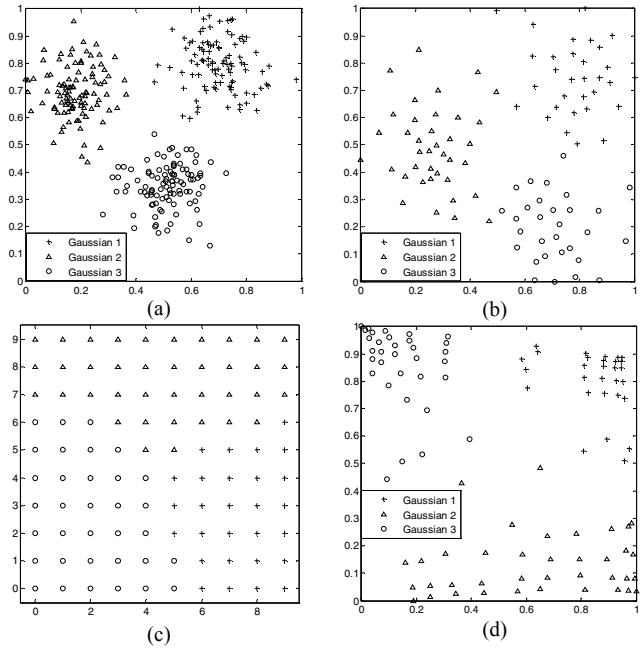

Figure 1. The visualization of three synthetic 2D Gaussians dataset. (a) 3 Synthetic 2D Gaussians, (b) SOM with modified AC, (c) SOM, and (d) SOM with original AC

B. Synthetic 3-D Gaussians

The labeled synthetic 3-D Gaussians from [5] is used. It consists of three Gaussians sources with 100 samples each. Their mean vectors are $[5.0\ 7.0\ 6.0]^T$, $[-2.0\ 5.0\ -3.0]^T$, and $[-10.0\ 6.0\ 2.0]^T$. Their respective covariance matrixes are

$$\begin{bmatrix} 5.0 & -1.0 & 0.3 \\ -1.0 & 0.3 & -1.0 \\ 0.3 & 1.5 & 4.0 \end{bmatrix}, \begin{bmatrix} 1.0 & 0.1 & 0.5 \\ -2.0 & 5.0 & 1.0 \\ 1.3 & -2.0 & 3.0 \end{bmatrix}, \text{ and } \begin{bmatrix} 0.1 & -1.0 & -0.2 \\ 1.2 & 2.3 & 1.4 \\ -0.3 & 1.2 & 4.0 \end{bmatrix}.$$

It is shown in Fig. 2 (a). It consists of two clouds of stretched longitudinally 3D Gaussians and one cloud of normal 3D Gaussian.

The projected maps (Fig. 2(b-d)) are results of dimension reduction from 3D to 2D. As in 2D synthetic dataset, SOM with modified AC removed SOM rigid grid (Fig. 2 (c)) and produced a better data topology preserved visualization (Fig. 2 (b)) as compare to the original AC (Fig. 2 (d)). Fig. 2 (b) revealed two stretched clusters and one normally distributed cluster. These 2D clusters resembled the dataset's 3D clusters. It shows the SOM with modified AC is able to reveal the clustering tendency of the original dataset even after performed the dimension reduction from 3D to 2D.

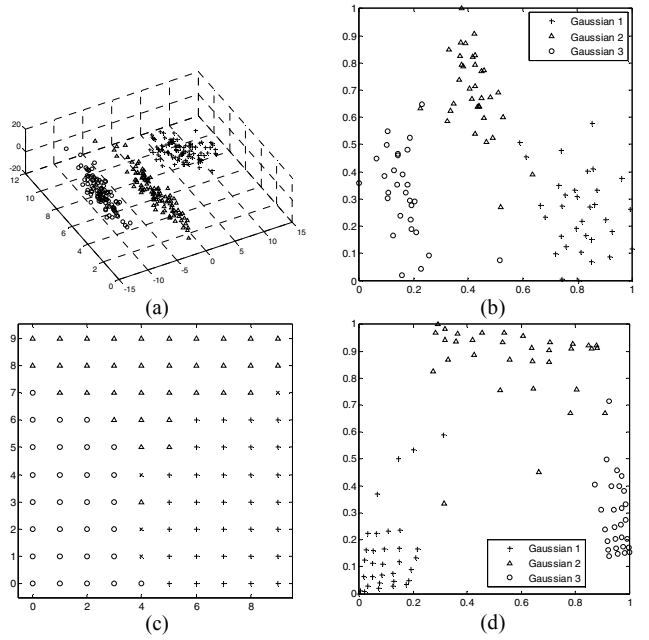

Figure. 2. The visualizations of three synthetic 3D Gaussians dataset. (a) 3 Synthetic 3D Gaussians, (b) SOM with modified AC, (c) SOM, and (d) SOM with original AC

C. Winconsin Breast Cancer Dataset

Winconsin Breast Cancer (WBC) [10] consists of 683 labeled samples with 9 dimension and 2 different classes. The samples with missing value were removed for easier processing.

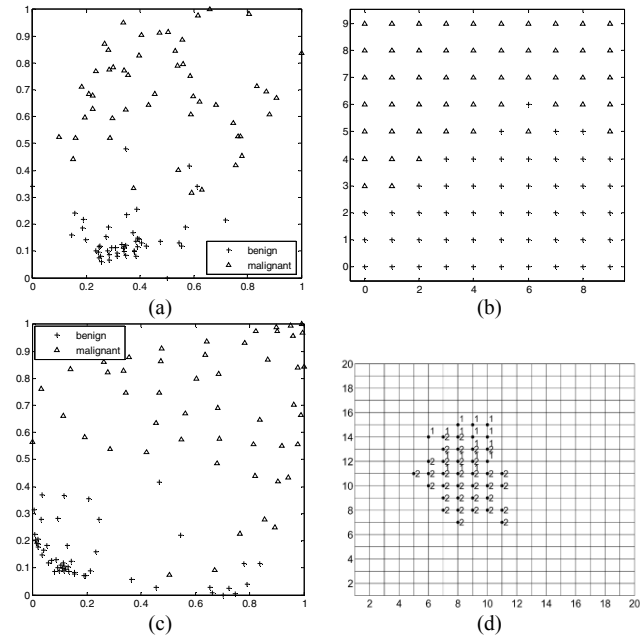

Figure 3. 2D visualizations of Winconsin Breast Cancer dataset. (a) SOM with modified AC, (b) SOM, (c) SOM with original AC, and (d) ViSOM (map size 20x20, λ=3.0) (adapted from [5])

After performing dimension reduction, SOM grid, as shown in Fig. 3 (b), is able to reveal some information about two clusters that represented the two diagnosed classes. Those are the benign and malignant classes. ViSOM regularized visualization in Fig. 3(d) is more informative than SOM. Two overlapping clusters are shown in the middle of the map. By using the proposed SOM with modified AC, even more information about the clustering tendencies is revealed as shown in Fig. 3 (a). It shows that benign class is much denser as compare to malignant class. Visual judgment shows that the SOM with modified AC (Fig. 3 (a)) is able to produce a much better topology preserved map than the SOM with original AC as shown in Fig. 3 (c), which is based on the mirroring effect.

D. Wine Data Set

Wine dataset [10] consists of 178 labeled samples with 13 dimension and 3 different classes. This dataset has higher dimensions as compare to previous datasets. SOM grid, as shown in Fig. 4 (b) is able to reveal three different clusters after performed the dimension reduction from 13D to 2D. But no information about the clustering density is available from the rigid grid.

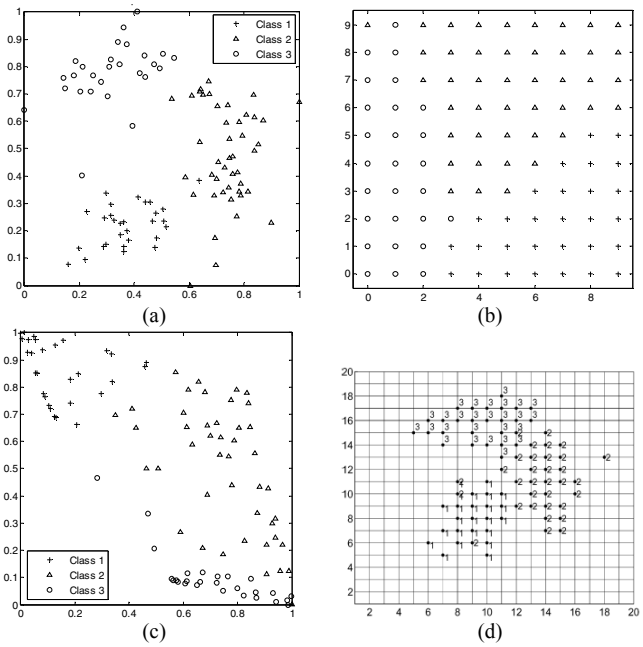

Figure. 4. 2D visualizations of Wine dataset. (a) SOM with modified AC, (b) SOM, (c) SOM with original AC, and (d) ViSOM (map size 20x20, λ=0.8) (adapted from [5])

ViSOM regularized visualization as shown in Fig. 4 (d) is able to show three clusters in the middle of the map. However, it is still confined to the rigid grid and most of the neurons were not utilized. By removing the rigid grid through SOM with modified AC, visual inspection revealed three distinct clusters as shown in Fig. 4 (a). SOM with modified AC provides better projection as compare to SOM with original AC as shown in Fig. 4 (c) where the clusters are rather indistinct.

VI. CONCLUSION

This paper proposed a new hybrid ANN by using SOM and modified AC for data dimension reduction and data visualization. It removes the rigid grid projection of SOM and thus produces topological preserved map. Empirical results demonstrated the hybrid algorithm is able to produce promising data structure and inter-neuron distances preserved visualization. No predefined parameter is required for the proposed algorithm. Therefore, the proposed hybrid technique shows potential for real life applications where topology preserved visualization and computationally efficient algorithm is required. Besides, it also shows the potential for fully automated intelligent system where little or no human intervention (nonparametric) is required. Inter-neurons distance preservation enhancement and probability density estimation for the projected map will be good candidates for future investigation.

ACKNOWLEDGMENT

This research was supported by Fundamental Research Grant Scheme through FRGS/02(06)/662/2007(27). The authors would also like to thank Universiti Malaysia Sarawak for supporting this work.

REFERENCES

[1] J. W. Sammon, "A nonlinear mapping for data structure analysis," IEEE Trans. on Computer, C-18, 401-409, 1969.

[2] R. A. Shepard and J. D. Carrol, "Parametric representation of nonlinear data structures," in Proc. of. Int'l. Symp. Multivariate Analysis (Krishnaiah, P. R. Ed.). New York: Academic, 1965, pp. 561-592.

[3] T. Kohonen, Self-organizing maps, 2nd ed., New York: Springer, 1997.

[4] H. Yin, "ViSOM: A novel method for multivariate data projection and structure visualization," IEEE Trans. on Neural Network, vol. 13(1), 2002, pp. 237–243.

[5] S. Wu and T. S. W. Chow, "PRSOM: A new visualization method by hybriding multidimensional scaling and self-organizing map," IEEE Trans. on Neural Network, vol. 16(6), 2005, pp. 1362–1380.

[6] D. Merkl, and A. Rauber, "Alternative ways for cluster visualization in self-organizing maps," In Proceeding Workshop on SOM, Espoo, Finland, 1997, pp. 106-111.

[7] D. Merkl, and A. Rauber, "The similarity of eagles, hawks, and cows: visualization of semantic similarity in self-organizing maps," Proc. Int'l Workshop on Fuzzy-Neuro-Systems, Soest, Germany, 1997.

[8] Z. T. Sarwar and C. S. Teh, "Hybridization of Learning Vector Quantization (LVQ) and Adaptive Coordinate (AC) for data visualization and classification," Int'l Conf. on Intelligent & Advance Systems, Malaysia, 2007.

[9] Z. T. Sarwar and C. S. Teh, "AC-ViSOM: Hybridising the Modified Adaptive Coordinate (AC) and ViSOM for Data Visualization," Int'l Symp. on Information Technology, Malaysia, 2008.

[10] D. J. Newman, S. Hettich, C. L. Blake and C. J. Merz, UCI Repository of machine learning databases, Dept. of ICS, Univ. of California at Irvine, 1998.

[11] A. E. Oprea, R. Strungaru, and G. M. Ungureanu, "A self organizing map approach to breast cancer detection," Int'l Conf. of IEEE in Engineering in Medicine and Biology Society. Canada, 2008, pp. 3032-3035.

[12] S. Wu, and T. W. S. Chow, "Self-organizing and self-evolving neurons: a new neural network for optimization," IEEE Trans. on Neural Netw., vol. 18(2), 2007, pp. 385-394.

[13] A. B. Far, F. Flitti, B. Guo, and A. Bermak, "A bio-inspired pattern recognition system for tin-oxide gas sensor applications," IEEE Sensor Journal, vol. 9(6),2009, pp. 713-722.

[14] D. E. Ilea and P.F. Whelan, "CTex – An Adaptive Unsupervised Segmentation Algorithm Based on Color-Texture Coherence," IEEE Trans. on Image Processsing, vol. 17(10), 2008, pp. 1926-1939.