Journal of Engineering Science and Technology Review 6 (3) (2013) 10-15

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Research Article

A Novel Spectral Clustering and its Application in Image Processing

Gu Ruijun*, Chen Shenglei and Wang Jiacai

School of Information Science, Nanjing Audit University,Nanjing, 211815, China

Received 15 June 2012; Accepted 23 January 2013

Abstract

This paper proposes an improved spectral clustering algorithm based on neighbour adaptive scale, who fully considers the local structure of dataset using neighbour adaptive scale, which simplifies the selection of parameters and makes the improved algorithm insensitive to both density and outliers. This paper illustrates the proposed algorithm not only has inhibition for certain outliers but is able to cluster the data sets with different scales. Experiments on UCI data sets show that the proposed method is effective. Some experiments were also performed in image clustering and image segmentation to demonstrate its excellent features in application.

Keywords: —Spectral graph theory, Spectral clustering, Neighbour adaptive scale, Image segmentation

1. Introduction

Clustering is one popular data analysis method and has been widely used in pattern recognition, image processing and data mining. Based on variable density of dataset, DBSCAN[1] can deal with datasets with arbitrary shape, but an inappropriate choice of parameters may yield poor result. Spectral clustering algorithms have seen an explosive development over the past years and been successfully used in image clustering and image segmentation[2]. They can deal with arbitrary distribution dataset and easy to implement. However, spectral clustering also has some weaknesses[3][4], for example, it is sensitive to the datasets with distinctly different densities and the parameters of algorithm must be selected cautiously.

This paper proposes an improved spectral clustering algorithm who fully considers the local structure of dataset like DBSCAN. Neighbour adaptive scale simplifies the selection of parameters and makes the improved algorithm insensitive to both density and outliers. To extend its application domain, a robust image clustering using image distance was proposed. Experiments on images returned by search engine show that the proposed method is effective.

The rest of the paper is organized as follows. In Section 2, Spectral clustering algorithm is summarized briefly. In Section 3, the proposed adaptive spectral clustering algorithm is described. In Section 4, experiments are presented and the results are discussed. Finally, a conclusion is provided in Section 5.

2. Spectral clustering

Spectral clustering can be interpreted in multi-view, such as graph cut theory [2][5], random walks point of view [6] and perturbation theory [7]. Spectral clustering in the end is the problem of finding eigenvectors of Laplacian matrix, and then clustering eigenvectors into clusters. Normalized Cut based-on spectral clustering algorithm [8] is a representative one [9](i.e. standard spectral clustering). Given a set of n points $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ in $\Re^l$, cluster them into c clusters as follows:

1) computer affinity matrix $\boldsymbol{A}\in\Re^{n\times n}$, in which $\boldsymbol{A}_{ii}=0$ and
$$\boldsymbol{A}_{ij}=\exp(-\left\|x_i-x_j\right\|^2/\sigma^2)\ (i\neq j) \quad (1)$$
2) Construct Laplacian matrix $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, in which $\boldsymbol{D}$ is diagonal matrix defined as $\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{A}_{ij}$.
3) Find $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_c$, the c largest eigenvectors of matrix $\boldsymbol{L}$ and form the matrix $\boldsymbol{F}=[\boldsymbol{f}_1,\ldots\boldsymbol{f}_c]\in\Re^{n\times c}$ (normalization when required).
4) Normalize the rows of $\boldsymbol{F}$ to be unit length, i.e. $\bar{\boldsymbol{F}}_{ij}=\boldsymbol{F}_{ij}/(\sum_{j=1}^{c}\boldsymbol{F}_{ij}^2)^{1/2}$.
5) Treating each row of $\bar{\boldsymbol{F}}$ as a point in $\Re^k$, cluster into clusters using *k-means* algorithm or any other sensible clustering algorithm.
6) Assign the original point $\boldsymbol{x}_i$ to cluster j if and only if row i of $\bar{\boldsymbol{F}}$ was assigned to cluster j.

We can see that two parameters influence the final clustering result and they are scale parameter σ in affinity function and clusters number c. In [8] the parameter σ is chosen based on that value of σ that gives the least distorted clusters by repeated experiments. It is obviously that the value of σ depends on specific problem and

* E-mail address: grj79@hotmail.com
ISSN: 1791-2377

detecting the range of σ is difficult. Moreover, repeated experiments will increase the compute complexity.

In conclusion, standard spectral clustering is sensitive to σ and specially not robust to dataset with variable densities. Besides, spectral clustering can find clusters in non-spherical-shape distribution with appropriate σ, however, k-means etc can not do.

3. Adaptive Spectral clustering algorithm based on neighbour adaptive scale

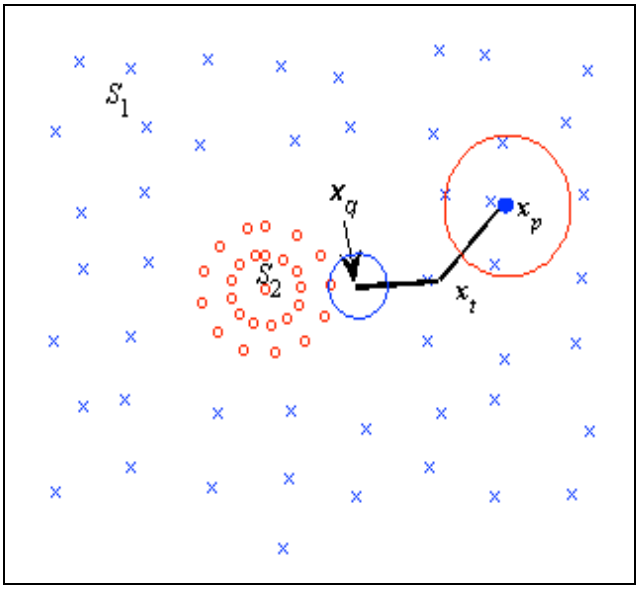


Fig.1 Theory of spectral clustering

From above analysis we know that maybe a varying σ is more appropriate for spectral clustering. Inspired by DBSCAN, for each point $\boldsymbol{x}_i$ an adaptive scale σ_i is set, which considers the local distribution of neighbours of xi. Let

$$\sigma_i = \frac{1}{k}\sum_{m=1}^{k}\|x_i - x_m\| \tag{2}$$

where σ_i is the average distance between $\boldsymbol{x}_i$ and its k nearest neighbours, called the Neighborhood Adaptive Scale (ASC) of point $\boldsymbol{x}_i$. However, self-tuning spectral clustering in [3] only considers the kth neighbour (usually k=7) and tends to be affected by outliers. Similar to [3], affinity between point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as

$$\bar{A}_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma_i\sigma_j) \tag{3}$$

because scale σ_i varies with neighbour distribution and is adaptive to local structure. This trait enlarges the affinity between two points in the same cluster and reduces that in different clusters as Fig.1 shows.

For convenience, assume that $\|\boldsymbol{x}_t\text{-}\boldsymbol{x}_p\| = \|\boldsymbol{x}_t\text{-}\boldsymbol{x}_q\|$. If constant scale σ is used , according formula (1), we can obtain

$$A_{tp} = \exp(-\|x_t - x_p\|^2 / \sigma^2) = \exp(-\|x_t - x_q\|^2 / \sigma^2) = A_{tq},$$

which shows that if arbitrary pair points have the same distances they have the same affinities. In fact, $\boldsymbol{x}_t$ and $\boldsymbol{x}_p$ are located in one cluster whereas $\boldsymbol{x}_t$, and $\boldsymbol{x}_q$ in different clusters, so if $A_{tp} > A_{tq}$, spectral clustering will benefit from this. If neighbour adaptive scale is adopted, then $\sigma_p > \sigma_q$ and

$$\begin{aligned}\bar{A}_{tp} &= \exp(-\|x_t - x_p\|^2 / \sigma_t\sigma_p) = \exp(-\|x_t - x_q\|^2 / \sigma_t\sigma_p) \\ &> \exp(-\|x_t - x_q\|^2 / \sigma_t\sigma_q) = \bar{A}_{tq}.\end{aligned}$$

It is obvious that neighbour adaptive scale makes points closer in one cluster and farther in different clusters. Replace the first step of standard spectral clustering with above algorithm, the improved spectral algorithm is named Adaptive Spectral Clustering (ASC). ASC can distinguish clusters with different densities. ASC utilizes average distance to tune affinity by means of the idea of DBSCAN. ASC is resistant to outliers, because σ_i is determined by the average distance, which is more reliable than the nearest neighbour distance or *kth* neighbour [3] distance. ASC does not have a bias towards a particular cluster shape or size compared with *k*-means.

4. Experimental results and analysis

To validate the efficiency of the ASC, we conducted extensive experiments using two publicly available datasets, and then compared ASC with *k*-means, standard spectral clustering and self-tuning spectral clustering.

Table I. Clustering Comparison on UCL Database

Datasets				*k*-means	Standard spectral clustering		Self-tuning spectral clustering	ASC	
Dataset	Sample size	Categories	Attributes		σ=0.1d_r	σ=0.2d_r		*k*=3	*k*=5
Glass	214	7	9	0.6808	0.6282	0.7061	0.6927	0.6973	**0.7084**
WDBC	569	2	30	0.7504	0.7554	0.7612	0.7707	**0.7973**	0.7965
Image segment	210	7	19	0.8448	0.8346	0.8120	0.8374	0.8505	**0.8525**
Iris	150	3	4	0.8797	0.8797	0.8737	0.8823	**0.8859**	**0.8859**
Ionosphere	351	2	34	0.5191	0.5191	0.5191	0.5004	**0.5938**	**0.5938**
average				0.7350	0.7234	0.7344	0.6139	0.7650	**0.7674**

4.1 Experiment on the UCI dataset

Next, five real datasets selected from UCI[11]database were used to evaluate ASC. To evaluate clustering performance quantitatively, we adopt popular Rand index as evaluation, The Rand index has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

The value of σ is typically set to 10 to 20 percent of the total range of the feature distance function d_r (d_r=max(***D***)- min(***D***)) according to [2]. Without loss of generality, the parameters of these algorithms were set as follows:

(1) standard spectral clustering: σ =0.1d_r, σ =0.2d_r;
(2) self-tuning spectral clustering: k=7;
(3) ASC: k=3, k=5.

TABLE 1 shows the attributes of UCI database and the experimental results of four clustering algorithms. As can be seen, ASC almost achieved all optimal *RIs,* moreover, the values are very near when k=3 or k=5. We can conclude that ASC is the most stable algorithm of them and insensitive to k. Although the *RIs* of standard spectral clustering varied widely, at most one *RI* is larger than that of k-means clustering. That is to say standard spectral clustering is certain to outperform k-means clustering if a suitable σ is set. Self-tuning spectral clustering is also not stable, because it just considers the seventh neighbour and is sensible to outliers.

4.2 Experiment on image clustering

The goal of image clustering is to find a mapping of the archive images into clusters such that the set of clusters provide nearly the same information about the entire image collection[13][14]. To improve the accuracy of image clustering, we introduce a novel image distance to measure the similarity between the images. Then neighbour spectral clustering is performed on the image similarity matrix.

Unlike the traditional Euclidean distance, IMD [15] takes into account the spatial relationships of pixels. Therefore, it is robust to small perturbation of images. Let vetor $\mathbf{x}$, $\mathbf{y}$ be two m by n images, $\mathbf{x}$ =(x_1, x_2, …, x_{mn}), $\mathbf{y}$ =(y_1, y_2, …, y_{mn}), the Euclidean distance $d^2_E(\mathbf{x}, \mathbf{y})$ is given by

$$d^2{}_E(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{mn}(x_i - y_i)^2 \tag{4}$$

Suppose $\mathbf{e}_1$, $\mathbf{e}_2$, …, $\mathbf{e}_{mn}$ is a base of mn-dimensional image space, the metric coefficients g_{ij} is defined as

$$g_{ij} = <\mathbf{e}_i, \mathbf{e}_j> = \sqrt{<\mathbf{e}_i, \mathbf{e}_i>}\sqrt{<\mathbf{e}_j, \mathbf{e}_j>}\cdot\cos\theta_{ij} \tag{5}$$

where <,> is the scalar product and θ_{ij} is the angle between $\mathbf{e}_i$ and $\mathbf{e}_j$. Then, the IMD of two images $\mathbf{x}$, $\mathbf{y}$ is written by

$$d^2{}_{IMD}(\mathbf{x},\mathbf{y}) = \sum_{i,j=1}^{mn} g_{ij}(x_i - y_i)(x_j - y_i) = (\mathbf{x}-\mathbf{y})^T\mathbf{G}(\mathbf{x}-\mathbf{y}) \quad ,$$

where $\mathbf{G} = (g_{ij})_{mn\times mn}$ is a symmetric positive definite matrix. g_{ij} can be represented as a Gaussian function

$$g_{ij} = \frac{1}{2\pi\sigma^2}\sum_{i,j=1}^{mn}\exp\{-|P_i - P_j|^2/2\sigma^2\} \tag{6}$$

where $|P_i$-$P_j|$ is the pixel distance between P_i and P_j. Let $\sigma = 1$, then IMD between image $\mathbf{x}$ and image $\mathbf{y}$ is defined as

$$d^2{}_{IMD}(\mathbf{x},\mathbf{y}) = \frac{1}{2\pi}\sum_{i,j=1}^{mn}\exp\{-|P_i - P_j|^2/2\}(x_i - y_i)(x_j - y_i)$$

To reduce computation complexity, $\mathbf{G}$ is decomposed by

$$\mathbf{G} = \mathbf{\Gamma\Lambda\Gamma}^T = (\mathbf{\Gamma\Lambda}^{1/2}\mathbf{\Gamma}^T)(\mathbf{\Gamma\Lambda}^{1/2}\mathbf{\Gamma}) = \mathbf{G}^{1/2}\mathbf{G}^{1/2} \tag{8}$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose elements are eigenvalues of $\mathbf{G}$ and $\mathbf{\Gamma}$ is an orthogonal matrix whose column vectors are eigenvectors of $\mathbf{G}$. Let $\mathbf{u} = \mathbf{G}^{1/2}\mathbf{x}$, $\mathbf{v} = \mathbf{G}^{1/2}\mathbf{y}$, so

$$d^2{}_{IMD}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{G}^{1/2}\mathbf{G}^{1/2}(\mathbf{x}-\mathbf{y}) = (\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})$$

which has the form of Euclidean distance. Because $\mathbf{G}$ is only associated with the size of an image, so can be computed before transform is performed.

Utilizing the property of IMD and spectral clustering, we propose a robust image clustering method named IMD-ASC , by replacing Euclidean distance with IMD. IMD- ASC can be descried as follows [16].

Step1: For N images (m by n), compute $\mathbf{G}^{1/2}$ according to equation (7) and equation (8);

Step2: For image $\mathbf{x}_i$, $\mathbf{x}_j$ (i,j=1,2,…,N), let $\mathbf{u}_i$=$\mathbf{G}^{1/2}\mathbf{x}_i$, $\mathbf{u}_j$=$\mathbf{G}^{1/2}\mathbf{x}_j$ then

$$d^2{}_{IMD}(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{G}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{u}_i - \mathbf{u}_j)^T(\mathbf{u}_i - \mathbf{u}_j)$$

Step3: Replace the Euclidean distance in ASC by IMD to perform spectral clustering.

Utilizing Google search engine, images were collected by inputting keywords like apple, Africa, Nanjing separately. Then the former 500 images were selected as the input of IMD-ASC. Clustering results were *partially* figured out in Fig.2-3.

In Fig.2, images were clustered into three main clusters identified by apples for eating, Apple logo and Apple products. In Fig.3, elephants, landscape and eagles are the typical Africa elements.

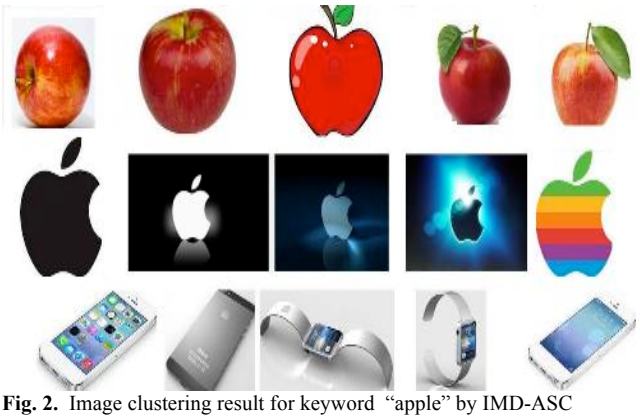


Fig. 2. Image clustering result for keyword "apple" by IMD-ASC

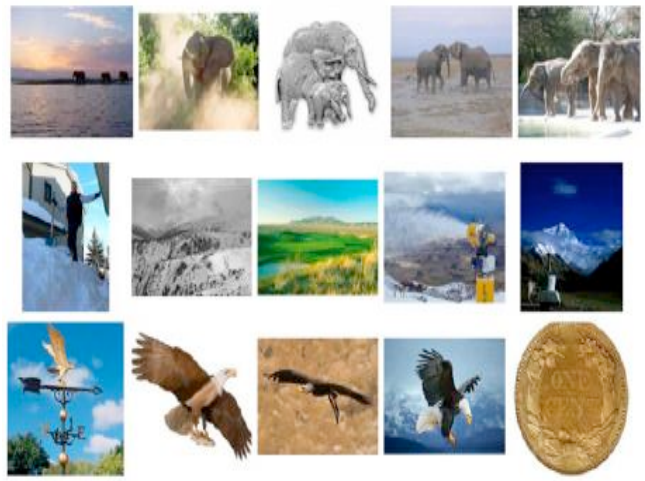
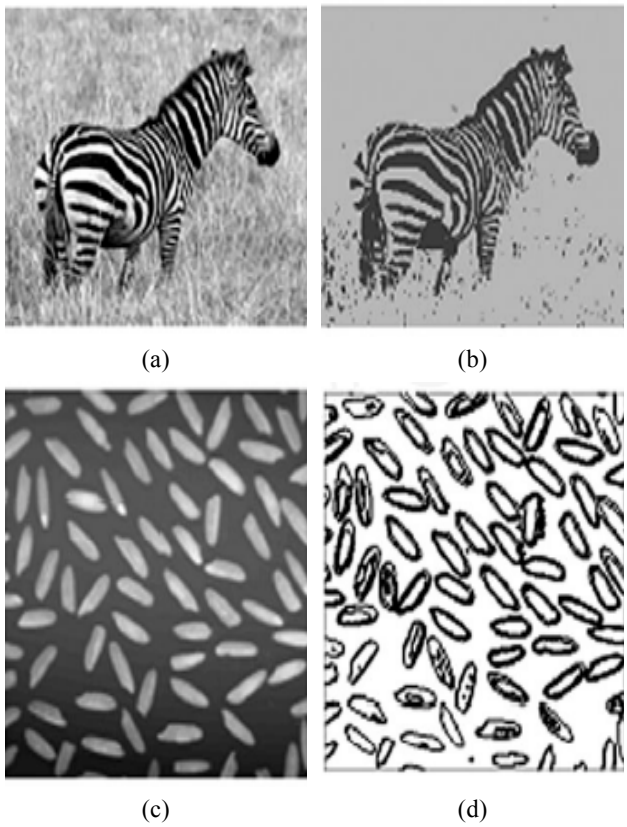


Fig. 3. Image clustering result for keyword "Affrica" by IMD-ASC

4.3 Experiment on image segmentation

Image segmentation [17] is the process of dividing an image into multiple parts, which is a difficult problem in computer vision. Image segmentation is typically used to identify objects or other relevant information in digital images. ASC can also be used on image segmentation. Fig.4 shows the result using ASC(k=5) on two typical gray images. Notice that for both images, our method can clearly segmented the object from the background.



(a) (b)

(c) (d)

Fig. 4 Gray image segmentaion result by ASC(k=5)

5. Conlusions and future work

ASC algorithm makes each point with a different neighbour adaptive scale computed according to the mean distance to its k nearest neighbors. Neighbour adaptive scale simplifies the selection of parameters and makes the improved algorithm insensitive to both density and outliers. Experimental results show that, compared with k-means and standard spectral clustering, our algorithm can achieve better clustering effect on artificial datasets and UCI public databases. However, ASC is sensible to noises and the number of clusters is difficult to select. Besides, computational complexity is increased. For NAS, some quantitative comparison has been performed. But for image clustering and segmentation, only qualitative results were concluded. It is expected that using evaluation to improve the performance further.

Acknowledgements

This research was funded by a grant (No.71271117/ G0112) from the National Natural Science Foundation of China and the Priority Academic Program Development of Jiangsu Higher Education Institutions (Auditing Science and Technology).

References

1. M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", in Proc. of the 2nd International Conference on Knowledge Discovery and Data mining. Portland, Oregon: AAAI Press, pp.226–231, 1996.
2. J. Shi, and J. Malik, "Normalized Cuts and Image Segmentation", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, pp. 888–905, Aug. 2000. Z. M. Lihi, and P. Pietro, "Self-tuning Spectral Clustering", in Advances in Neural Information Processing Systems 17, Cambridge, MA: MIT Press, pp.1601–1608, 2005.

3. U. von Luxburg, O. Bousquet, and M. Belkin, "Limits of Spectral Clustering", in Advances in Neural Information Processing Systems 17. Cambridge, MA: MIT Press, pp.857–864, 2005.
4. F. R. K. Chung, Spectral Graph Theory, Providence, R.I.: American Mathematical Society, 1997.
5. M. Meila, and J. Shi, "A Random Walks View of Spectral Segmentation", in Proc. of 8th International Conference on Artificial Intelligence and Statistics. Key West, FL, 2001.
6. U. von Luxburg, "A Tutorial on Spectral Clustering", Journa Statistics and Computing, vol.17,pp. 395 – 416, Dec. 2007.
7. A. Ng, M. Jordan, and Y. Weiss. "On Spectral Clustering: Analysis and an Algorithm", in Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002.
8. D. Verma, and M. Meila, A Comparison of Spectral Clustering Algorithms, University of Washington, Tech Rep: UW-CSE-03-05-01, 2003. [Online]. Available: ftp://ftp.cs.washington.edu/tr/2003/05/ UW-CSE-03-05-01.PS.Z
9. B. Feil, and J.Abonyi, "Geodesic Distance Based Fuzzy Clustering", in Proc. of 11th Online World Conference on Soft Computing in Industrial Applications, 2006. [Online]. Available: http://www.cs. Armstrong.edu /wsc11/index.html.
10. C. Blake E. Keogh, and C. J. Merz, UCI Repository of Machine Learning Databases, Dept. Inf. Comp. Sci., Univ. California, Irvine, 1998. [Online]. Available: http://www.ics.uci.edu/~mlearn/ MLRepository . html
11. W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods", Journal of the American Statistical Association, pp.846–850, 1971.
12. S. Gordon, H.Greenspan,"Unsupervised Image-set Clustering Using an Information Theoretic Framework", IEEE Trans. on Image Processing,vol 15,pp.449-458,Feb. 2006.
13. H. Li ,X.Huang, "A Hierarchical Image Clustering Cosegmentation Framework", in Proc. of the IEEE CVPR, pp. 686 - 693, 2012.
14. L. Wang, Y. Zhang and J. Feng, "On the Euclidean Distance of Images", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, pp. 1334-1339, Aug.2005.
15. C. Pughineanu, I. Balan. "Parallel Algorithm Evaluation in the Image and Clustering Processing", Electronics and Electrical Engineering. Vol110,No 4,pp.89-92, 2011.
16. D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. "Icoseg: Interactive Co-Segmentation with Iintelligent Scribble Guidance", in Proc. of the IEEE CVPR, pp.3169–3176, 2010.
17. Y, Fang, L. Tian and B. Han."An Improved Watermarking Algorithm to Colour Image Based on Wavelet Domain",Journal of Engineering Science and Technology Review. vol.6,pp.139-144,2012.