Extracting and Labelling the Objects from an Image by Using the Fuzzy Clustering Algorithm and a New Cluster Validity

Chien-Hsing Chou, Yi-Zeng Hsieh, Mu-Chun Su, and Yung-Long Chu

Abstract—**Many real-world and man-made objects are line symmetry. To detection the line-symmetry objects from an image, in this paper, a new cluster validity measure which adopts a non-metric distance measure based on the idea of "line symmetry" is presented. The thresholding technique is first applied to extract the objects from the original image; and the object pixels are transferred to be the data patterns. Then the fuzzy clustering algorithm is applied to label the object pixels; and the proposed validity measure is used in determining the number of objects. Simulation results are used to illustrate the performance of the proposed measure.**

Index Terms—**Extract object, cluster validity, clustering algorithm, line symmetry, similarity measure.**

I. Introduction

Many real-world and man-made objects are line symmetry. Base on this idea, we apply cluster analysis technique to detect the line-symmetry objects from an image. Cluster analysis is an important tool for exploring the underlying structure of a given data set and plays an important role in many applications [1]-[4]. In cluster analysis, two crucial problems required to be solved are (1) the determining of the similarity measure based on which patterns are assigned to the corresponding clusters and (2) the determining of the optimal number of clusters. While the determining of the similarity measure is the so-called data clustering problem, the estimation of the number of clusters in the data set is the cluster validity problem. In this paper, we focus on the research topic of cluster validity.

Many different cluster validity measures have been proposed [5]-[12], such as the Dunn's separation measure [5], the Bezdek's partition coefficient [6], the Xie-Beni's separation measure [7], Davies-Bouldin's measure [8], the Gath-Geva's measure [9], the CS measure [10] etc. Some of these validity measures assume a certain geometrical structure in cluster shapes. For example, the Gath-Geva's validity measure that uses the value of fuzzy hypervolume as a measure is a good choice for compact hyperellipsoidal clusters. However, it is a bad choice for shell clusters since the decision as to whether it is a well or badly recognized ellipsoidal shell should be independent of the radii or the volume of ellipses. A minimization of the fuzzy hypervolume makes no sense for the recognition of ellipsoidal shells. Hence, some special validity measures (such as Dave's fuzzy shell covariance matrix [11] and shell thickness) are proposed for shell clusters. Depending on the desired results, a particular validity measure should be chosen for the respective application.

The organization of the rest of the paper is as follows. In Section II, we introduced the idea of line symmetry distance measures. Then the proposed validity measure employing the line symmetry distance was fully discussed in Section III. Two examples were used to demonstrate the effectiveness of the new validity measure. Section IV presents the simulation results. Finally, Section V presents the conclusion.

II. The Line Symmetry Distance

In one of our previous work, a so-called "line symmetry" distance was proposed in [12]. Following the definition of a figure with line symmetry (see Fig. 1), we may point out that the line symmetrical data pattern relative to $\underline{x}_j$ with respect to a center $\underline{c}$ and a unit direction vector $\underline{e}$ is the data pattern $\underline{x}_{j*}^{ls}$, where the point symmetrical data pattern relative to $\underline{x}_j$ with respect to a center $\underline{c}$ is denoted as $\underline{x}_{j*}^{ps}$. The definition of the line symmetry distance is given as follows. Given a reference vector $\underline{c}$ and a unit direction vector $\underline{e}$, the "line symmetry distance" of a pattern $\underline{x}_j$ in the data set X with respective to a reference vector $\underline{c}$ and a unit direction vector $\underline{e}$ is defined as

$$d_{ls}(\underline{x}_j,\ \underline{c},\ \underline{e}) = \min_{\substack{i=1,\cdots,N \\ and\ \ i\neq j}} \frac{\|(\underline{x}_j-\underline{p})+(\underline{x}_i-\underline{p})\|}{(\|\underline{x}_j-\underline{p}\|+\|\underline{x}_{j*}^{ls}-\underline{p}\|+\|\underline{x}_i-\underline{x}_{j*}^{ls}\|)} \tag{1}$$

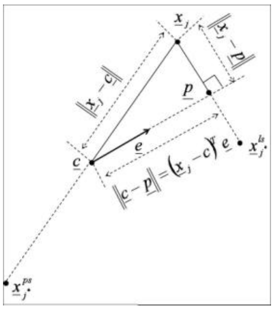

$\underline{x}_j$
$\|\underline{x}_j-\underline{p}\|$
$\|\underline{x}_j-\underline{c}\|$
$\underline{p}$
$\underline{e}$
$\underline{c}$
$\|\underline{c}-\underline{p}\|=(\underline{x}_j-\underline{c})^T\underline{e}$
$\underline{x}_{j*}^{ls}$
$\underline{x}_{j*}^{ps}$

Fig. 1. A geometrical explanation about the definitions of point symmetry and line symmetry.

Manuscript received November 20, 2012; revised January 28, 2013. This work was supported by the National Science Council, Taiwan, R.O.C., under the Grant NSC 101-2221-E-032-055.

Chien-Hsing Chou and Yung-Long Chu are with the Department of Electrical Engineering, Tamkang University, Taiwan (e-mail: chchou@mail.tku.edu.tw).

Yi-Zeng Hsieh and Mu-Chun Su are with the Department of Computer Science & Information Engineering, National Central University, Taiwan.

DOI: 10.7763/IJCCE.2013.V2.188

where the data pattern $\underline{p}$ is the normal projection of the data pattern $\underline{x}_j$ onto the line formed by the data pattern $\underline{c}$ and the unit direction vector $\underline{e}$. As for how to find the three vectors, $\underline{c}$, $\underline{p}$ and $\underline{e}$ from the data set *X*, the computational procedure will be explained as follows. First of all, the mean vector $\underline{c}$ and the covariance matrix *Cov* can be approximated from the *N* data patterns by

$$\underline{c} = \frac{1}{N}\sum_{i=1}^{N}\underline{x}_i \tag{2}$$

$$Cov = \frac{1}{N}\sum_{i=1}^{N}\underline{x}_i\underline{x}_i^T - \underline{c}\underline{c}^T \tag{3}$$

III. The Validity Measure Using Line Symmetry

The proposed validity measure is referred to as *LS* measure and is computed as follows. Consider a partition of the data set $X = \{\underline{x}_j;\ j = 1, 2, \ldots, N\}$ and each data pattern $\underline{x}_j$ is assigned to its corresponding cluster by a particular clustering algorithm. In order to calculate line symmetry distance, we need re-compute the cluster center $\underline{v}_i$ (i.e. mean vector) and the covariance matrix Cov_i by using the following equation:

$$\underline{v}_i = \frac{1}{N_i}\sum_{\underline{x}_j\in S_i}\underline{x}_j \tag{4}$$

$$Cov_i = \frac{1}{N_i}\sum_{\underline{x}_j\in S_i}\underline{x}_j\underline{x}_j^T - \underline{v}_i\underline{v}_i^T \tag{5}$$

where S_i is the set whose elements are the data patterns assi-gnned to the *i*th cluster and N_i is the number of elements in S_i. Note that we assign data patterns to the corresponding clusters using the maximum membership grade criterion if the clustering result is achieved by fuzzy clustering algorithms. Then we compute the degree of line symmetry of cluster *i* by

$$LS_i = \frac{1}{N_i}\sum_{\underline{x}_j\in S_i} d_c(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k) = \frac{1}{N_i}\sum_{\underline{x}_j\in S_i}(d_{ls}(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k) + d_0)d_e(\underline{x}_j, \underline{v}_i) \tag{6}$$

where the distance, $d_c(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k)$, represents the compo-site symmetry distance defined in Eq. (6), $d_e(\underline{x}, \underline{v}_i)$ re-presents the Euclidean distance between $\underline{x}_j$ and $\underline{v}_i$, and d_0 is a small valued positive constant. The reason why we use the composite symmetry distance, $d_c(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k)$, rather than the line symmetry distance itself, $d_{ls}(\underline{x}, \underline{v}_i, \underline{e}_{i*}^k)$, is as follows. The line symmetry distance itself may not work for situations where clusters themselves are line symmetric. A possible solution to overcome this limitation is to combine the line symmetric distance with the Euclidean distance in such a way that if data patterns are relatively close, then the line symmetry is more important. On the other hand, if the data patterns are very far, then the Euclidean distance is more important. The smaller the value of LS_i is the larger the de-gree of line symmetry of cluster *i* has. The separation of clus-ters is defined as the minimum distance between clusters

$$d_{\min} = \min_{\substack{m,n=1,\cdots,c \\ and\quad m\neq n}} d_e(\underline{v}_m, \underline{v}_n) \tag{7}$$

Finally, the *LS* measure is obtained by averaging the ratio of the degree of line symmetry of the cluster to the separation over all clusters, more explicitly.

$$\begin{aligned} LS(c) &= \frac{\frac{1}{c}\sum_{i=1}^{c}LS_i}{d_{\min}} \\ &= \frac{\frac{1}{c}\sum_{i=1}^{c}\left[\frac{1}{N_i}\sum_{\underline{x}_j\in S_i}d_c(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k)\right]}{d_{\min}} \\ &= \frac{\frac{1}{c}\sum_{i=1}^{c}\left[\frac{1}{N_i}\sum_{\underline{x}_j\in S_i}(d_{ls}(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k) + d_0)d_e(\underline{x}_j, \underline{v}_i)\right]}{d_{\min}} \end{aligned} \tag{8}$$

IV. Experimental Results

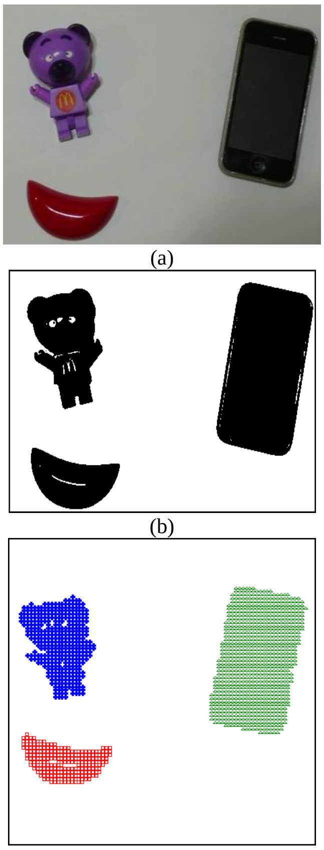
We illustrate the effectiveness of the proposed validity measure by testing two data sets with different geometrical structures. For the comparison purpose, these data sets were also tested by the three popular validity measures—the partition coefficient (PC) [6], the classification entropy (CE) [6] and the Xie-Beni's separation measure (S) [7]. The Gustafson-Kessel (GK) algorithm [7] is applied to cluster these data sets at each cluster number *c* from *c*=2 to *c*=10. The parameter d_o was chosen to be 0.005 for the modified version of line symmetry distance.

V. Example

This example demonstrates an application of the LS validity measure to detect the number of objects in an image. In image processing, it is very important to find objects in images. In this example, these objects have different geometric shapes. Fig. 2(a) shows a real image consisting of a mobile phone, a doll, and an object of crescent. First, we apply the thresholding technique to extract the objects from the original image (see Fig. 2(b)). Then we transfer the object pixels to be the data patterns. The GK algorithm is used to cluster the data set. Table I shows the performance of each validity measure. The LS validity measure finds that the optimal cluster number *c* is at *c*=3. However, the PC, CE and S validity measures find the optimal cluster number at *c*=2. Once again, this example demonstrates that the proposed LS validity measure can work well for a set of clusters of different geometrical shapes. The clustering result achieved by the GK algorithm at c=3 is shown in Fig. 2 (c). Three objects of line-symmetry structure are labeling by the proposed method.

TABLE I: NUMERICAL VALUES OF THE VALIDITY MEASURES FOR EXAMPLE 1.

C	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PC	**0.956**	0.846	0.786	0.728	0.682	0.638	0.605	0.588	0.570
CE	**0.101**	0.307	0.422	0.553	0.657	0.724	0.815	0.854	0.956
S	**0.071**	0.111	0.136	0.244	0.323	0.375	0.321	0.436	0.363
LS	0.034	**0.018**	0.027	0.043	0.052	0.039	0.064	0.062	0.056



(a)

(b)

(c)

Fig. 2. (a) The original image. (b) The binary image by applying thresholding. (c) Three objects are labeled by the GK algorithm.

VI. CONCLUSION

Based on the line symmetry distance, a new measure LS is then proposed for cluster validation. The simulation results reveal the interesting observations about the validity measures discussed in this paper. The proposed LS validity measure shows that consistency for the tested examples. Although these simulations show that the new measure outperforms the other three measures, we want to emphasize that the clusters should be assumed as line symmetrical structures. If the data set does not follow the assumption, the measure may not work well. In fact, a lot of future work can be done to improve not only the line symmetry distance but also the LS measure.

REFERENCES

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, New Jersey, NJ: Prentice Hall, 1988.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, New York: Wiley, 2001.

[3] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

[4] F. H öppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis-Methods for Classification, Data Analysis and Image Recognition*, John Wiley & Sons, Ltd, 1999.

[5] J. C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions," *Journal Cybern.*, vol. 4, pp. 95-104, 1974.

[6] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *J. Math. Biol.*, vol. 1, pp. 57-71, 1974.

[7] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847, 1991.

[8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, no. 4, pp. 224-227, 1979.

[9] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 11, pp. 773-781, 1989.

[10] C. H. Chou, M. C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 205-220, 2004.

[11] R. N. Dave, "New measures for evaluating fuzzy partitions induced through c-shells clustering," in *Proc. SPIE Conf. Intell. Robot Computer Vision X*, Boston, vol. 1670, pp. 406-414, 1991.

[12] Y. Z. Hsieh, M. C. Su, C. H. Chou, and P. C. Wang, "Detection of line-symmetry clusters," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 8, pp. 1-17, 2011.

Chien-Hsing Chou received the B.S. and M.S. degrees from the Department of Electrical Engineering, Tamkang University, Taiwan, in 1997 and 1999, respectively, and the Ph.D. degree at the Department of Electrical Engineering from Tamkang University, Taiwan, in 2003. He is currently an assistant professor of electrical engineering at Tamkang University, Taiwan. His research interests include image analysis and recognition, mobile phone programming, machine learning, document analysis and recognition, and clustering analysis.

Yi-Zeng Hsieh received the Ph.D. degree in computer science and information engineering from National Central University, Tao-yuan, Taiwan, respectively in 2012. His current research interests include neural networks, pattern recognition, image processing.

Mu-Chun Su received the B. S. degree in electronics engineering from National Chiao Tung University, Taiwan, in 1986, and the M. S. and Ph.D. degrees in electrical engineering from University of Maryland, College Park, in 1990 and 1993, respectively. He was the IEEE Franklin V. Taylor Award recipient for the most outstanding paper co-authored with Dr. N. DeClaris and presented to the 1991 IEEE SMC Conference. He is currently a professor of computer science and information engineering at National Central University, Taiwan. He is a senior member of the IEEE Computational Intelligence Society and Systems, Man, and Cybernetics Society. His current research interests include neural networks, fuzzy systems, assistive technologies, swarm intelligence, effective computing, pattern recognition, physiological signal processing, and image processing.

Yung-Long Chu received the B.S. degree from the Department of Electronic Engineering, Ming Chuan University, Taiwan, 2012. He is currently a master student at Tamkang University, Taiwan. His research interests include image analysis and recognition and mobile phone programming.