A New Measure of Cluster Validity Using Line Symmetry[*]

CHIEN-HSING CHOU[1], YI-ZENG HSIEH[2] AND MU-CHUN SU[2]
[1]Department of Electrical Engineering
Tamkang University
Tamhsui, 251 Taiwan
[2]Department of Computer Science and Information Engineering
National Central University
Chungli, 320 Taiwan
E-mail: muchun@csie.ncu.edu.tw

Many real-world and man-made objects are symmetry, therefore, it is reasonable to assume that some kind of symmetry may exist in data clusters. 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 proposed validity measure can be applied in finding the number of clusters of different geometrical structures. Several data sets are used to illustrate the performance of the proposed measure.

Keywords: cluster validity, clustering algorithm, line symmetry, cluster analysis, similarity measure, unsupervised learning

1. INTRODUCTION

Cluster analysis is one of the basic tools for exploring the underlying structure of a given data set and plays an important role in many applications [1-6]. 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 order to mathematically identify clusters in a data set, it is usually necessary to first define a measure of similarity or proximity which will establish a rule for assigning patterns to the domain of a particular cluster center. Recently, several different clustering algorithms have been proposed to deal with clusters with various geometric shapes. These algorithms can detect compact clusters [7], straight lines [8], shells [9-11], contours with polygonal boundaries [12] or well-separated non-convex clusters [13]. One thing that should be emphasized is that there is no cluster algorithm which can tackle all kinds of clusters. Some comprehensive overview of clustering algorithms can be found in the literature [1-4].

In fact, if cluster analysis is to make a significant contribution to engineering applications, much more attention must be paid to cluster validity issues that are concerned with determining the optimal number of clusters and checking the quality of clustering results. For the partition-based clustering algorithm [7-13], the cluster number should be decided in prior. Basically, there are three different approaches to the determination of

the cluster number of a data set. The first approach is to use a certain global validity measure to validate clustering results for a range of cluster numbers [14-19]. The second approach is based on the idea of performing progressive clustering [20-22]. The third approach is the projection-based approach [23-29]. Projection algorithms allow us to visualize high-dimensional data as a two-dimensional or three-dimensional scatter plot. The focus of the paper is not to argue which approach is the best one. If one decides to choose the first approach then we try to provide a new measure to validate clustering results more effectively than existing measures. The goal of this paper is to propose a new validity measure that is able to deal with clusters with line symmetry structure.

Many different cluster validity measures have been proposed [14-19, 30-38], such as the Dunn's separation measure [14], the Bezdek's partition coefficient [15], the Xie-Beni's separation measure [16], Davies-Bouldin's measure [17], the Gath-Geva's measure [18], the CS measure [19] *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 [30] and shell thickness [31]) are proposed for shell clusters. Depending on the desired results, a particular validity measure should be chosen for the respective application.

In many cases, several different geometrical structures (*e.g.* compact clusters and shell clusters) may simultaneously exist in a data set. Furthermore, clusters can be of arbitrary shapes. Examples can be found in various chromosome images and images of objects with different geometrical structures. Most of the common criteria for cluster validity cannot work well in complex real data with a large variability of cluster shapes. Since clusters can be of arbitrary shapes and sizes, we have to seek a more flexible validity measure to deal with the problem of geometrical shapes. The laws of nature give symmetry or near-symmetry to their products. Looking around us, we get the immediate impression that almost every interesting real-world objects (*e.g.* flowers, starfish, human faces, etc) or man-made objects (*e.g.* architectures, rose windows, 3C products, *etc.*) own some kind of generalized form of symmetry (*e.g.* point symmetry, radial symmetry, reflective symmetry or line symmetry, *etc.*). In addition to the shapes of objects, symmetry is also an important parameter in physical and chemical processes.

Since symmetry is so common in nature, it is reasonable to assume some kind of symmetry exist in the structures of data clusters. Therefore, we may assign data to a cluster if they present a symmetrical structure with respect to the corresponding cluster center. Those aforementioned symmetry operators [39-43] are efficient and affective in image processing for the detection of objects with some kind of symmetry; however, it seems that there is no simple way to generalize them to cluster high-dimensional data. Although objects with point symmetry is very widespread, line symmetry is the most common type of symmetry around us.

In this paper, we try to explore the possibility of using line symmetrical properties as a new validity measure to deal with data with different geometrical structures. The proposed validity measure can be applied in finding the number of clusters of different

geometrical structures. Several data sets are used to illustrate the performance of the proposed measure. The organization of the rest of the paper is as follows. In Section 2, we briefly reviewed some popular validity measure. In Section 3, we first introduced the idea of line symmetry distance measures. Then the proposed validity measure employing the line symmetry distance was fully discussed. Several examples were used to demonstrate the effectiveness of the new validity measure. Section 4 presents the simulation results. Finally, Section 5 presents the conclusion.

2. CLUSTER VALIDITY MEASURES

Cluster validation refers to procedures that evaluate the clustering results in a quantitative and objective fashion. Some kinds of validity measures are usually adopted to measure the adequacy of a structure recovered through cluster analysis. Determining the correct number of clusters in a data set has been, by far, the most common application of cluster validity. Bensaid *et al.* [44] grouped validity measures into three categories. The first category consists of validity measures that evaluate the properties of the crisp structure imposed on the data by the clustering algorithm [14, 17]. The second category consists of measures that use the membership degrees produced by the corresponding fuzzy clustering algorithms (*e.g.* the FCM algorithm) [3, 15], where the membership degree represents the possibility of a data pattern belonging to the specified cluster. The third category consists of validity measures that take into account not only the membership degrees but also the data themselves [16, 18, 31]. Each has its own considerations and limitations. A comparative examination of thirty validity measures is presented in [33] and an overview of the various measures can be found in [45]. Since it is not feasible to attempt a comprehensive comparison of our proposed validity measure with many others, we just chose three of the popular measures for comparisons. These three measures have different rationales and properties. Consider a partition of the data set $X = \{\underline{x}_j, j = 1, 2, \ldots, N\}$ and the center of each cluster $\underline{v}_i (i = 1, 2, \ldots, c)$, where N is the data number and c represents the cluster number.

Partition coefficient (PC) [15]:

Bezdek designs the partition coefficient (PC) to measure the amount of "overlap" between clusters. He defines the partition coefficient (PC) as follows:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^2 \tag{1}$$

where $u_{ij}(i = 1, 2, \ldots, c; j = 1, 2, \ldots, N)$ is the membership of data pattern j in cluster i. The closer this value is to unity the better the data are classified. In the case of a hard partition, we obtain the maximum value $PC(c) = 1$. If we are looking for a good partition, we aim at a partition with a maximum partition coefficient. This kind of partition yields the "most unambiguous" assignment. The disadvantages of the partition coefficient are its monotonic decreasing with c and the lack of direct connection to some properties of the data themselves.

Classification entropy (CE) [3]:

The classification entropy measure strongly resembles the partition coefficient; however, it is related on Shannon's information theory. The classification entropy is defined as follows:

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij} \log(u_{ij}).$$

(2)

If we have a crisp partition we have the most information (*i.e.* the minimum entropy). Consequently, a partition with the minimum entropy is regarded as a good partition. Although this validity measure is based on Shannon's information theory, it can be viewed as a measure for the fuzziness of cluster partition, which is very similar to the partition coefficients. Bezdek proves that the relation $0 \leq 1 - PC(c) \leq CE(c)$ holds for all probabilistic cluster partitions c. The limitation of the classification entropy can be attributed to its apparent monotony and to an extent, to the heuristic nature of the rationale underlying its formulation.

Separation measure (S) [16]:

Xie and Beni introduce a validity measure, similar to the Dunn's separation measure [14]. The separation measure S is defined as follows:

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^{2} \left\| \underline{x}_j - \underline{v}_i \right\|^2}{N \min_{\substack{m,n=1,\dots,c \\ and\ m \neq n}} \{ \left\| \underline{v}_m - \underline{v}_n \right\|^2 \}} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^{2} \left\| \underline{x}_j - \underline{v}_i \right\|^2}{N * d_{min}^2}$$

(3)

where $\|\underline{x}_j - \underline{v}_i\|$ denotes the Euclidean distance between the pattern, $\underline{x}_j$, and the cluster center, $\underline{v}_i$, and d_{min} represents the minimum Euclidean distance between cluster centers. The separation measure S is based on fuzzy compactness and separation. While the numerator of Eq. (3) measures the total variance (or the compactness) of each fuzzy cluster, the denominator measures the separation defined as the minimum distance between cluster centers. The smallest value of $S(c)$ indicates a valid optimal partition because u_{ij} will be high when $\|\underline{x}_j - \underline{v}_i\|$ is low and well separated clusters will produce a high value of d_{min}.

Before introducing the proposed measure, we first use the example shown in Fig. 1 (a) to illustrate the motivation of the new measure. There are three clusters in the dataset containing a spherical cluster and two linear clusters, and two linear clusters are close to each other. First, we use the FCM algorithm [3] to group the data set into 2 clusters (shown in Fig. 1 (b)) and 3 clusters (shown in Fig. 1 (c)). For the case of 2 clusters, the membership degrees for the two patterns, $\underline{x}_1$ and $\underline{x}_2$ (shown in Fig. 1 (b)), are $u_{11} = 0.991$, $u_{21} = 0.009$, $u_{12} = 0.901$ and $u_{22} = 0.098$. As for the case of 3 clusters, the membership degrees become $u_{11} = 0.544$, $u_{21} = 0.033$, $u_{31} = 0.423$, $u_{12} = 0.909$, $u_{22} = 0.025$, and $u_{32} = 0.066$. By examining these clustering results, we have the following observations:
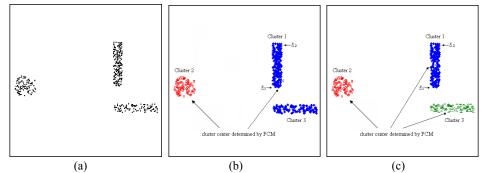
(a)   (b)   (c)

Fig. 1. (a) The data set containing a combination of a spherical cluster and two linear clusters. The clustering result achieved by the FCM algorithm at (b) $c = 2$; (c) $c = 3$.

1. While the membership degree, u_{12}, of the data pattern, $\underline{x}_2$, increases from 0.901 (for the case of 2 clusters) to 0.909 (for the case of 3 clusters); the membership degree, u_{11}, of the data pattern, $\underline{x}_1$, decreases from 0.991 (for the case of 2 clusters) to 0.544 (for the case of 3 clusters). For data patterns located at the upper region of cluster 1 (*e.g.* $\underline{x}_2$) or the right region of cluster 3, the increment of their membership degrees induced by partitioning the data set into 3 clusters indicates that it is appropriate to partition the data set into 3 clusters. On the contrary, the decrement of the membership degrees for data patterns located at the lower region of cluster 1 (*e.g.* $\underline{x}_1$) or the left region of cluster 3 indicates that the partition of 3 clusters is not a good choice. Because the total decrement of membership degrees is larger than the total increment of membership degrees due to the increases of the number of clusters, the value of the PC measure decreases from 0.923 to 0.877. Understandably, the PC measure favors the partition of two clusters in this example. A similar reason can also be applied to explain why the classification entropy measure CE can not find the right number of clusters for this data set either.

2. Not only dose the value of the numerator in Eq. (3) changes from 86.59 to 46.87 when the number of clusters changes from 2 to 3, but also the minimum Euclidean distance between cluster centers, d_{min}, changes from 4.51 to 2.17. While the decrement of the numerator indicates that the degree of compactness increases, the decrement of the denominator indicates the degree of separation decreases. Finally, the value of $S(c)$ changes from 0.048 to 0.054. Therefore, the $S(c)$ measure prefers the partition of 2 clusters.

3. The two measures, PC and CE, use the basic heuristic rule that good clusters are not fuzzy. Consequently, they rely solely upon the memberships to determine the fuzziness of the partition. This example confirms what one already knows: validity measures that use only the fuzzy membership grades usually lack direct connection to some geometrical properties of the data themselves.

4. The separation measure S uses $\sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^{2} \left\| \underline{x}_j - \underline{v}_i \right\|^2$ to measure the compactness. For compact spherical clusters, this kind of measure may be effective; however, for some clusters with different geometrical structures (*e.g.* linear clusters, shells, *etc*.), it may not work well. In this example, the data set includes not only a spherical

cluster but also two linear clusters so that it is not very surprising when we find out that the separation measure S can not find the correct partition.

5. If we move two linear clusters closer to the spherical cluster (*e.g.* the distances between each other do not vary too much), the measures, PC, CE, and S, can find the correct number of clusters.

Each validity measure implicitly imposes a certain structure on the data set, and if the data set happens to exhibit the structures, the validity measure can find the correct partition. No matter how good a validity measure is, there always are limitations associated with it since clusters can be of arbitrary shapes and sizes. Hence, there is no single "best" validity measure which can work well for any kind of clustering results.

These observations motivate us to explore whether there is a validity measure which can deal with the validation of clusters with different geometric shapes under some certain assumptions. In our previous work [46], we find that line symmetry distance is very effective in clustering data with different structures. Consequently, we try to propose a new validity measure which is based on the idea of line symmetry. This new measure considers not only the distributions of the data set but also the degree of symmetry in each cluster.

3. THE LINE SYMMETRY DISTANCE AND THE VALIDITY MEASURE USING LINE SYMMETRY

What is line symmetry? For a 2-dimenstional figure, if it can be folded in such a way that one-half of it lies exactly on the other half is said to have line symmetry. The idea of line symmetry is very clear and simple but an immediate problem is how to find a metric to measure line symmetry. A kind of line symmetry distance was proposed in [47-49]. In this approach, the symmetrical line of a data set is defined by a center vector and an angle between the major axis of the data set and the x axis. The information of the major axis of the data points belonging to a class or a cluster is computed by the moment of order $(p + q)$ method. Then the major axis is treated as the symmetrical line of that class or cluster. Furthermore, the Kd-tree based nearest neighbor search is used to reduce the complexity of computing the line distance. It is not clear how the proposed line symmetry distance can be generalized to high-dimensional space. In their approach, the symmetrical line is defined by a point and an angle between the major axis and the x axis. As we know, a point combined with an angle can define one and only one line in 2-dimensional space but will produce an infinite number of lines in high-dimensional space. In addition, the authors themselves stated that the performance of their line symmetry distance depends on the selection of the number of nearest neighbors used in the computation of line symmetry distance. The authors in [47-49] proposed another similar definition of line symmetry in [50]. They assumed that a data set is symmetrical then it should be also be symmetric with respect to the first principal axis of the data set [50]. This assumption may not be hold for many situations. There are objects which is symmetrical with respective to their second principal axis but not symmetrical to their first principal axis. Therefore, a new line symmetry distance which can deal with high-dimensional data points is developed by our previous work [46].

3.1 The Line Symmetry Distance

In one of our previous work, a so-called "point symmetry" distance (symmetry about a cluster center) was proposed in [51]. In [52], Bandyopadhyay and Saha proposed a new point symmetry-based distance measure. Their algorithm applies GA and Kd-tree based nearest neighbor search to reduce the complexity of finding the closest symmetric point. Although object with point symmetry is very widespread, line symmetry is the most common type of symmetry around us. Based on the point symmetry distance, a new line symmetry distance is developed in our another work [46]. Before we present the definition of the proposed line symmetry distance, we briefly review the definition of the point symmetry distance.

A 2-dimensional figure is with point symmetry if it can be rotated 180 degrees about a point onto itself. To generalize this idea of point symmetry to measure the degree of point symmetry for a set of high-dimensional data patterns, a point symmetry distance is defined as follows [51]. Given a data set X containing N patterns, $\underline{x}_j$, $j = 1, \ldots, N$, and a reference vector $\underline{c}$ (*e.g.* a cluster center), the "point symmetry distance" between a pattern $\underline{x}_j$ and the reference vector $\underline{c}$ is defined as

$$d_{ps}(\underline{x}_j, \underline{c}) = \min_{\substack{i=1,\cdots,N \\ and \; i \neq j}} \frac{\| (\underline{x}_j - \underline{c}) + (\underline{x}_i - \underline{c}) \|}{(\| \underline{x}_j - \underline{c} \| + \| \underline{x}_i - \underline{c} \|)}. \tag{4}$$

Note that Eq. (4) is minimized (*i.e.*, $d_{ps}(\underline{x}_j, \underline{c}) = 0$) if there is a pattern $\underline{x}_{j*}^{ps} = (2\underline{c} - \underline{x}_j)$ exists in the data set (see Fig. 2). The pattern $\underline{x}_{j*}^{ps}$ is then denoted as the point symmetrical data pattern relative to $\underline{x}_j$ with respect to $\underline{c}$ as shown in Fig. 2.

Fig. 2. An example for illustrating the bias problem incurred by the original version of the point symmetry distance.

By further analyzing the point symmetry distance defined in Eq. (4), we find that the distance measure has a bias for data patterns with a larger distance from the center $\underline{c}$ even if they are with the same distance to the point symmetrical data pattern $\underline{x}_{j*}^{ps}$. For example, two data patterns, $\underline{x}_1$ and $\underline{x}_2$, are with the same distance to the symmetrical data pattern $\underline{x}_{j*}^{ps}$ as shown in Fig. 2. According to the definition of Eq. (4), we find that

$$\frac{\| (\underline{x}_j - \underline{c}) + (\underline{x}_1 - \underline{c}) \|}{(\| \underline{x}_j - \underline{c} \| + \| \underline{x}_1 - \underline{c} \|)} > \frac{\| (\underline{x}_j - \underline{c}) + (\underline{x}_2 - \underline{c}) \|}{(\| \underline{x}_j - \underline{c} \| + \| \underline{x}_2 - \underline{c} \|)}$$ even if these two patterns are with the same

distances to the point symmetrical data pattern $\underline{x}_{j*}^{ps}$. To amend this flaw, we modify the point symmetry distance as follows:
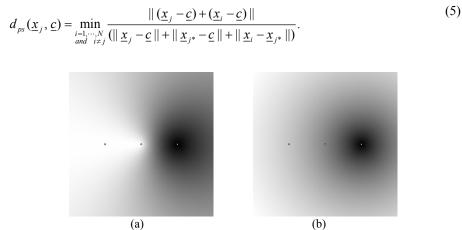
$$d_{ps}(\underline{x}_j, \underline{c}) = \min_{\substack{i=1,\cdots,N \\ and \ i \neq j}} \frac{\|(\underline{x}_j - \underline{c}) + (\underline{x}_i - \underline{c})\|}{(\|\underline{x}_j - \underline{c}\| + \|\underline{x}_{j*} - \underline{c}\| + \|\underline{x}_i - \underline{x}_{j*}\|)}. \tag{5}$$

(a)	(b)

Fig. 3. The distance contours generated by the two versions of the point symmetry distance; (a) By Eq. (4). (b) By Eq. (5).

Fig. 3 shows examples of the distance functions defined in Eqs. (4) and (5) for the case of $\underline{x}_j = (-1, 0)$, $\underline{c} = (0, 0)^T$, and $\underline{x}_{j*}^{ps} = (1, 0)^T$ (corresponding to three small circle points in the figure). For each point in the picture plane, the distance is used for the intensity. Here, small distance gets a dark color and large distance gets a light grey color. From Fig. 3, we find that Fig. 3 (b) presents concentric contours with center at the point symmetrical data pattern $\underline{x}_{j*}^{ps}$ but Fig. 3 (a) presents non-concentric contours. Therefore, the bias problem has been fixed by modifying the denominator of Eq. (4).

Following the definition of a figure with point symmetry, 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 $\underline{x}_{j*}^{ls}$ as shown in Fig. 4 where the point symmetrical data pattern relative to $\underline{x}_j$ with respect to a center $\underline{c}$ is denoted as $\underline{x}_{j*}^{ps}$. Similar to the definition of the modified version of the point symmetry distance defined in Eq. (5), 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

$$d_{ls}(\underline{x}_j, \underline{c}, \underline{e}) = \min_{\substack{i=1,\cdots,N \\ and \ i \neq j}} \frac{\|(\underline{x}_j - \underline{p}) + (\underline{x}_i - \underline{p})\|}{(\|\underline{x}_j - \underline{p}\| + \|\underline{x}_{j*}^{ls} - \underline{p}\| + \|\underline{x}_i - \underline{x}_{j*}^{ls}\|)} \tag{6}$$

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

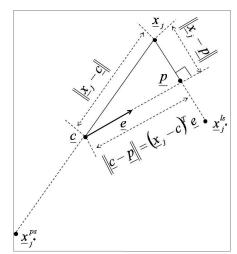$$\underline{c} = \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i, \tag{7}$$

Fig. 4. A geometrical explanation about the definitions of point symmetry and line symmetry.

$$Cov = \frac{1}{N}\sum_{i=1}^{N}\underline{x}_i\,\underline{x}_i^{T} - \underline{c}\underline{c}^{T}. \tag{8}$$

Since the covariance matrix Cov is real and symmetric, we can find a set of n orthonormal eigenvectors from the matrix. One of the n orthonormal eigenvector will be chosen to be the unit direction vector $\underline{e}$. The vector $\underline{e}$ is then regarded as the symmetric line of the data set. For each cluster, the eigenvector with the smallest amount of line symmetry distances (*e.g.*, the i^*th eigenvector in this case) is chosen to be the symmetrical line of that cluster based on the following minimum-value criterion:

$$i^* = Arg \min_{i=1,\cdots n} \sum_{j\in S_k} d_{ls}(\underline{x}_j,\,\underline{c}_k,\underline{e}_i^{k}) \tag{9}$$

where S_k represents the data set consisting of data points belonging to cluster k and $\underline{e}_i^{k}$ represents the ith eigenvector computed from the covariance matrix corresponding to the data set S_k. The normal projected data pattern $\underline{p}$ can be computed by

$$\underline{p} = \underline{c} + \|\underline{c} - \underline{p}\| \times \underline{e} = \underline{c} + \lfloor(\underline{x}_j - \underline{c})^{T}\underline{e}\rfloor \times \underline{e}. \tag{10}$$

After we have computed the normal projected data pattern $\underline{p}$, we can find the line symmetrical data pattern, $\underline{x}_{j*}^{ls}$, relative to $\underline{x}_j$ with respect to the center $\underline{c}$ and the unit direction vector $\underline{e}$ by the following equation:

$$\begin{aligned}\underline{x}_{j*}^{ls} &= \underline{x}_j + 2\|\underline{p} - \underline{x}_j\| \times (\underline{p} - \underline{x}_j)\\ &= \underline{x}_j + 2\sqrt{\|\underline{x}_j - \underline{c}\|^{2} - \|\underline{p} - \underline{c}\|^{2}} \times (\underline{p} - \underline{x}_j)\\ &= \underline{x}_j + 2\sqrt{\|\underline{x}_j - \underline{c}\|^{2} - [(\underline{x}_j - \underline{c})^{T}\underline{e}]^{2}} \times (\underline{p} - \underline{x}_j)\cdot\end{aligned} \tag{11}$$

3.2 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:

$$\underline{v}_i = \frac{1}{N_i} \sum_{\underline{x}_j \in S_i} \underline{x}_j \tag{12}$$

$$Cov_i = \frac{1}{N_i} \sum_{\underline{x}_j \in S_i} \underline{x}_j \underline{x}_j^T - \underline{v}_i \underline{v}_i^T \tag{13}$$

where S_i is the set whose elements are the data patterns assigned to the ith 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

$$LS_i = \frac{1}{N_i} \sum_{\underline{x}_j \in S_i} d_c(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k) = \frac{1}{N_i} \sum_{\underline{x}_j \in S_i} (d_{ls}(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k) + d_0) d_e(\underline{x}_j, \underline{v}_i) \tag{14}$$

where the distance, $d_c(\underline{x}_j, \underline{v}_j, \underline{e}_{i*}^k)$, represents the composite line symmetry distance defined in Eq. (14), $d_e(\underline{x}_j, \underline{v}_i)$ represents 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 line symmetry distance, $d_c(\underline{x}_j, \underline{v}_j, \underline{e}_{i*}^k)$, 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 degree of line symmetry of cluster i has. The separation of clusters is defined as the minimum distance between clusters

$$d_{\min} = \min_{\substack{m,n=1,\cdots,c \\ and \quad m \neq n}} d_e(\underline{v}_m, \underline{v}_n). \tag{15}$$

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

$$LS(c) = \frac{\frac{1}{c} \sum_{i=1}^{c} LS_i}{d_{\min}}$$

$$= \frac{\dfrac{1}{c}\sum_{i=1}^{c}\left[\dfrac{1}{N_i}\sum_{\underline{x}_j \in S_i} d_c(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k)\right]}{d_{\min}}$$

$$= \frac{\dfrac{1}{c}\sum_{i=1}^{c}\left[\dfrac{1}{N_i}\sum_{\underline{x}_j \in S_i} (d_{ls}(\underline{x}_j, \underline{v}_i, \underline{e}_{i*}^k) + d_0)d_e(\underline{x}_j, \underline{v}_i)\right]}{d_{\min}}.$$

(16)

While the numerator of Eq. (16) measures the average degree of line symmetry of entire clusters, the denominator measures the separation defined as the minimum distance between cluster centers. This measure favors the creation of a set of clusters that are maximally separated from one another and each cluster has the degree of line symmetry as large as possible. Therefore, a partition with the minimum $LS(c)$ is regarded as a good partition.

4. EXPERIMENTAL RESULTS

We illustrate the effectiveness of the proposed validity measure by testing some 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), the classification entropy (CE) and the Xie-Beni's separation measure (S). The FCM algorithm or the Gustafson-Kessel (GK) algorithm [7] is applied to cluster these data sets at each cluster number c from $c = 2$ to $c = 10$. According the experimental results, we notice that the FCM algorithm is not suitable to be applied for ellipsoidal clusters, but the GK algorithm can be used to cluster spherical and ellipsoidal clusters. Therefore, in examples 2 to 5, the GK algorithm is applied to cluster the data patterns. A value of the fuzzy exponent $m = 2$ was chosen for the FCM algorithm and the GK algorithm. The parameter d_0 was chosen to be 0.005 for the composite line symmetry distance.

Example 1: The data set shown in Fig. 1 (a) consists of a combination of a spherical cluster and two linear clusters. We generated the data patterns randomly with uniform distribution. The total number of data patterns is 400. In this example, the FCM algorithm is applied to cluster the data patterns. The total number of data patterns is 400. The performance of each validity measure is tabulated in Table 1. In Table 1 and others to follow, the highlighted (**bold and shaded**) entries correspond to optimal values of the measures. Note that only the LS validity measure finds the optimal cluster is three, but the PC, CE and S validity measures choose two clusters as the optimal partition. The clustering result achieved by the FCM algorithm at $c = 3$ is shown in Fig. 1 (c).

Table 1. Numerical values of the validity measures for Example 1.

C	2	3	4	5	6	7	8	9	10
PC	**0.923**	0.877	0.790	0.829	0.792	0.723	0.738	0.674	0.661
CE	**0.153**	0.242	0.397	0.358	0.440	0.546	0.573	0.666	0.709
S	**0.048**	0.054	1.372	0.075	0.142	0.297	0.363	0.257	0.218
LS	0.050	**0.017**	0.037	0.096	0.238	0.170	0.160	0.184	0.158

Example 2: We randomly generated a mixture of spherical and ellipsoidal clusters with Gaussian distribution. This data set consists of 850 data patterns distributed on five clusters, as shown in Fig. 5 (a). The clustering results achieved by the GK and FCM algorithms at $c = 5$ are shown in Figs. 5 (b) and (c). We notice that the FCM algorithm is not suitable to be applied for ellipsoidal clusters; the two non-overlapped compact clusters are not clustered correctly. Therefore, in this example, the GK algorithm is applied to cluster the data patterns. The performance of each validity measure is given in Table 2. Both the S and LS validity measures find that the optimal cluster number c is at $c = 5$, but both the PC and CE validity measures indicate two clusters as the optimal partition.

Table 2. Numerical values of the validity measures for Example 2.

C	2	3	4	5	6	7	8	9	10
PC	**0.836**	0.743	0.738	0.780	0.731	0.681	0.654	0.634	0.591
CE	**0.287**	0.469	0.524	0.484	0.592	0.698	0.771	0.840	0.911
S	0.398	0.180	0.186	**0.082**	0.383	0.392	0.296	0.274	0.796
LS	0.143	0.083	0.045	**0.041**	0.091	0.152	0.126	0.180	0.137

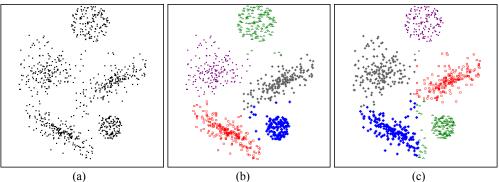|		
(a)	(b)	(c)

Fig. 5. (a) The data set used in example 2; (b) the clustering result achieved by the GK algorithm at $c = 5$; (c) the clustering result achieved by the FCM algorithm at $c = 5$.

Example 3: This data set consists of a combination of a ring-shaped cluster, a rectangular compact cluster and a linear cluster as shown in Fig. 6 (a). We generated the data patterns randomly with uniform distribution. The total number of data patterns is 400. In this example, the GK algorithm is applied to cluster the data patterns. This example is considered to show that the validity measures, PC, CE and S, are not indicative of a good detection of clusters with different geometric shapes. The performance of each validity measure is given in Table 3. Only the LS validity measure finds the optimal cluster c at $c = 3$. The PC, CE and S validity measures fail to choose the correct number of clusters, and select $c = 2$ as the optimal partition. The clustering results achieved by the GK and FCM algorithms at $c = 3$ are shown in Figs. 6 (b) and (c).

Table 3. Numerical values of the validity measures for Example 3.

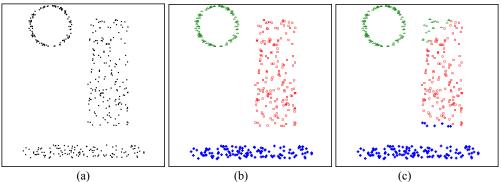| C	2	3	4	5	6	7	8	9	10
PC	**0.837**	0.834	0.772	0.708	0.669	0.658	0.657	0.645	0.629
CE	**0.279**	0.312	0.438	0.574	0.660	0.724	0.720	0.762	0.814
S	**0.116**	0.124	0.710	1.062	0.631	0.439	0.409	0.652	0.551
LS	0.076	**0.035**	0.041	0.215	0.185	0.164	0.161	0.160	0.167

|(a)|(b)|(c)|

Fig. 6. (a) The data set used in example 3; (b) the clustering result achieved by the GK algorithm at $c = 3$; (c) the clustering result achieved by the FCM algorithm at $c = 3$.

Example 4: 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. 7 (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. 7 (b)). Then we transfer the object pixels to be the data patterns. The GK algorithm is used to cluster the data set. Table 4 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 results achieved by the GK and FCM algorithms at $c = 3$ are shown in Figs. 7 (c) and (d).

Table 4. Numerical values of the validity measures for Example 4.

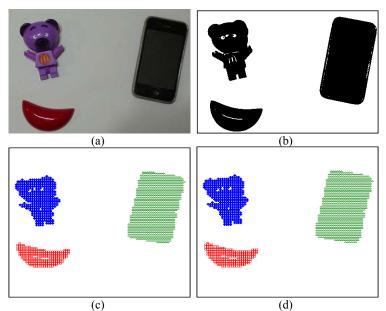| C	2	3	4	5	6	7	8	9	10
PC	**0.956**	0.846	0.786	0.728	0.682	0.638	0.605	0.588	0.570
CE	**0.101**	0.307	0.422	0.553	0.657	0.724	0.815	0.854	0.956
S	**0.071**	0.111	0.136	0.244	0.323	0.375	0.321	0.436	0.363
LS	0.034	**0.018**	0.027	0.043	0.052	0.039	0.064	0.062	0.056

Fig. 7. (a) The original image containing a mobile phone, a toll, and an object of crescent; (b) the binary image by applying threshold to the original image; (c) the clustering result achieved by the GK algorithm at $c = 3$; (d) the clustering result achieved by the FCM algorithm at $c = 3$.

Example 5: In this example, there are four objects with different geometric shapes in Fig. 8 (a), which consists of an adhesive tape, a knife, a note and a cover of a teacup. The threshold technique is applied to extract the objects from the original image (see Fig. 8 (b)), and then the object pixels are transferred to be the data patterns. The GK algorithm is used to cluster the data set. Table 5 shows the performance of each validity measure. The LS validity measure finds that the optimal cluster number c is at $c = 4$. However, the PC and CE validity measures find the optimal cluster number at $c = 2$, and the S validity measure finds the optimal cluster number at $c = 3$. The clustering results achieved by the GK and FCM algorithms at $c = 4$ are shown in Figs. 8 (c) and (d).

Table 5. Numerical values of the validity measures for example 5.

C	2	3	4	5	6	7	8	9	10
PC	**0.878**	0.775	0.740	0.686	0.665	0.645	0.611	0.605	0.586
CE	**0.221**	0.431	0.517	0.631	0.702	0.734	0.842	0.896	0.923
S	0.151	**0.110**	0.113	0.287	0.225	0.723	0.419	0.908	0.302
LS	0.023	0.044	**0.012**	0.114	0.083	0.127	0.096	0.086	0.120

5. DISCUSSION AND 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 S validity measure fails if not in one example, at
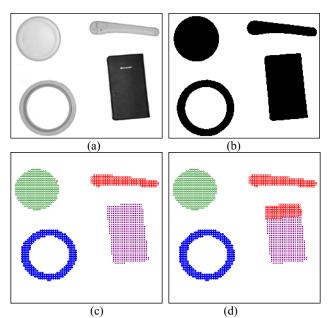
Fig. 8. (a) The original image containing an adhesive tape, a knife, a note and a cover of a teacup;
(b) the binary image by applying threshold to the original image; (c) the clustering result
achieved by the GK algorithm at $c = 4$; (d) the clustering result achieved by the FCM algo-
rithm at $c = 4$.

least in the other, and the PC and CE validity measures fail in all the examples. The pro-
posed LS validity measure shows that consistency for these and several other examples
that are tried. Because the line symmetry distance is extended from the point symmetry
distance, one may interest the performance if the cluster validity measure is based on
point symmetry. According our simulations, if the clusters with point symmetry structure
(*e.g.* examples 2 and 3), the validity measure based on point symmetry should also per-
form well for these cases. However, if the geometrical structure of clusters is line sym-
metry (*e.g.* example 4), it not a good choice to use point symmetry as validity measure.

Although these simulations show that the new measure outperforms the other three
measures, we want to emphasize that there are also limitations associated with this new
measure. First, we need to assume that clusters are line symmetrical structures. If the data
set does not follow the assumption, the measure may not work well. Second, for large
data sets, the determination of the measure is very computationally expensive. The price
paid for the flexibility in detecting cluster number is the increase of computational com-
plexity. The computing complexity is $O(N^2)$, where N indicates the data number. 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, Pren-
   tice Hall, NJ, 1988.

2. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, NY, 2001.
3. J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, NY, 1981.
4. F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis-Methods for Classification*, *Data Analysis and Image Recognition*, John Wiley & Sons, Ltd., 1999.
5. C.-S. Fahn and Y.-C. Lo, "On the clustering of head-related transfer functions used for 3-D sound localization," *Journal of Information Science and Engineering*, Vol. 19, 2003, pp. 141-157.
6. P.-C. Wang and J.-J. Leou, "New fuzzy hierarchical clustering algorithms," *Journal of Information Science and Engineering*, Vol. 9, 1993, pp. 461-489.
7. D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proceedings of IEEE Conference on Decision and Control*, 1979, pp. 761-766.
8. R. N. Dave, "Use of the adaptive fuzzy clustering algorithm to detect lines in digital images," *Intelligent Robots and Computer Vision VIII*, Vol. 1192, 1989, pp. 600-611.
9. R. N. Dave, "Fuzzy shell-clustering and application to circle detection in digital images," *International Journal General Systems*, Vol. 16, 1990, pp. 343-355.
10. Y. Man and I. Gath, "Detection and separation of ring-shaped clusters using fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, 1994, pp. 855-861.
11. R. N. Dave and K. Bhaswan, "Adaptive fuzzy c-shells clustering and detection of ellipses," *IEEE Transactions on Neural Networks,* Vol. 3, No. 5, 1992, pp. 643-662.
12. F. Höppner, "Fuzzy shell clustering algorithms in image processing: fuzzy *c*-rectangular and 2-rectangular shells," *IEEE Transactions on Fuzzy Systems*, Vol. 5, 1997, pp. 599-613.
13. M. C. Su and Y. C. Liu, "A new approach to clustering data with arbitrary shapes," *Pattern Recognition*, Vol. 38, 2005, pp. 1887-1901.
14. J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal Cybern*, Vol. 4, 1974, pp. 95-104.
15. J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *Journal of Mathematical Biology*, Vol. 1, 1974, pp. 57-71.
16. X. L. Xie and G. Beni, "A validity measure for fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, 1991, pp. 841-847.
17. D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 1, 1979, pp. 224-227.
18. I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, 1989, pp. 773-781.
19. 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, 2004, pp. 205-220.
20. T. L. Huntsherger, C. L. Jacobs, and R. L. Cannon, "Iterative fuzzy image segmentation," *Pattern Recognition*, Vol. 18, 1985, pp. 131-138.
21. R. N. Dave and K. J. Patel, "Progressive fuzzy clustering algorithms for characteristic shape recognition," in *Proceedings of North American Fuzzy Information Processing Society*, Vol. 1, 1990, pp. 121-124.
22. J. M. Jolion, P. Meer, and S. Bataouche, "Robust clustering with applications in computer vision," *IEEE Transactions on Pattern Analysis and Machine*, Vol. 13,

1991, pp. 791-802.

23. J. W. Shammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, Vol. 18, 1969, pp. 491-509.

24. M. A. Kraaijveld, J. Mao, and A. K. Jain, "A nonlinear projection method based on Kohonen's topology preserving maps," *IEEE Transactions on Neural Networks*, Vol. 6, 1995, pp. 548-559.

25. J. Mao and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Transactions on Neural Networks*, Vol. 6, 1995, pp. 296-317.

26. M. C. Su, N. DeClaris, and T. K. Liu, "Application of neural networks in cluster analysis," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 1997, pp. 1-6.

27. M. C. Su and H. C. Chang, "A new model of self-organizing neural networks and its application in data projection," *IEEE Transactions on Neural Networks*, Vol. 12, 2001, pp. 153-158.

28. H. Yin, "Data visualization and manifold mapping using the ViSOM," *Neural Networks*, Vol. 15, 2002, pp. 1005-1016.

29. M. C. Su, S. Y. Su, and Y. X. Zhao, "A swarm-inspired projection algorithm," *Pattern Recognition*, Vol. 42, 2009, pp. 2764-2786.

30. R. N. Dave, "New measures for evaluating fuzzy partitions induced through c-shells clustering," in *Proceedings of SPIE Conference on Intelligent Robot Computer Vision X*, Vol. 1670, 1991, pp. 406-414.

31. R. N. Dave, "Validating fuzzy partitions obtained through c-shells clustering," *Pattern Recognition Letters*, Vol. 17, 1996, pp. 613-623.

32. R. Krishnapuram, H. Frogui, and O. Nasraoui, "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation – Part 1&2," *IEEE Transactions on Fuzzy Systems*, Vol. 3, 1995, pp. 29-61.

33. M. P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets and Systems*, Vol. 5, 1981, pp. 177-185.

34. J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on System, Man, and Cybernetics*, Vol. 28, 1998, pp. 301-315.

35. V. S. Tseng and C.-P. Kao, "An efficient approach to identifying and validating clusters in multivariate datasets with applications in gene expression analysis," *Journal of Information Science and Engineering*, Vol. 20, 2004, pp. 665-677.

36. J.-S. Wang and J.-C. Chiang, "A cluster validity measure with outlier detection for support vector clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, Vol. 38, 2008, pp. 78-89.

37. R. Jain and A. Koronios, "Innovation in the cluster validating techniques," *Fuzzy Optimization and Decision Making*, Vol. 7, 2008, pp. 233-241.

38. A. K. Jain, M. N. Murthy, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, Vol. 31, 1999, pp. 265-323.

39. D. Reisfeld, H. Wolfsow, and Y. Yeshurun, "Context-free attentional operators: the generalized symmetry transform," *International Journal of Computer Vision*, Vol. 14, 1995, pp. 119-130.

40. D. Reisfeld and Y. Yeshurun, "Preprocessing of face images: Detection of features and pose normalisation," *Computer Vision and Image Understanding*, Vol. 71, 1998,

pp. 413-430.

41. R. K. K. Yip, "A hough transform technique for the detection of reflectional symmetry and skew-symmetry," *Pattern Recognition Letters*, Vol. 21, 2000, pp. 117-130.

42. G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Transactions on Pattern Analysis and Machine*, Vol. 25, 2003, pp. 959-973.

43. G. Loy and J. Eklundh, "Detecting symmetry and symmetric constellations of features," in *Proceedings of European Conference on Computer Vision*, LNCS 3952, 2006, pp. 508-521.

44. A. M. Bensaid, L.O. Hall, J. C. Bezdek, C. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh, "Validity-guided (Re) clustering with applications to image segmentation," *IEEE Transactions on Fuzzy Systems*, Vol. 4, 1996, pp. 112-123.

45. R. Dubes and A. K. Jain, "Validity studies in clustering methodologies," *Pattern Recognition*, Vol. 11, 1979, pp. 235-253.

46. 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, 2011, pp. 1-17.

47. S. Saha, S. Bandyopadhyay, and C. T. Singh, "A new line symmetry distance based pattern classifier," in *Proceedings of International Joint Conference on Neural Network*, 2008, pp. 1425-1432.

48. S. Saha and S. Bandyopadhyay, "A new line symmetry distance and its application in data clustering," *Journal of Computer Science and Technology*, Vol. 24, 2009, pp. 544-556.

49. S. Saha and S. Bandyopadhyay, "On principle axis based line symmetry clustering techniques," *Memetic Computing*, Vol. 3, 2011, pp. 129-144.

50. S. Bandyopadhyay and S. Saha, "A new principal axis based line symmetry measurement and its application to clustering," *The Proceedings of ICONIP*, LNCS 5507, 2009, pp. 543-550.

51. M. C. Su and C. H. Chou, "A modified version of the K-means algorithm with a distance based on cluster symmetry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 674-680.

52. S. Bandyopadhyay and S. Saha, "GAPS: A new symmetry based genetic clustering technique," *Pattern Recognition*, Vol. 40, 2007, pp. 3430-3451.

Chien-Hsing Chou (周建興) 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 (謝易錚) received the Ph.D. degree in Computer Science and Information Engineering from National Central University, Taoyuan, Taiwan, respectively in 2012. His current research interests include neural networks, pattern recognition, image processing.

Mu-Chun Su (蘇木春) 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 Unirsity, 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.