Clustering by Scale-Space Filtering

Yee Leung, Jiang-She Zhang, and Zong-Ben Xu

Abstract—In pattern recognition and image processing, the major application areas of cluster analysis, human eyes seem to possess a singular aptitude to group objects and find important structures in an efficient and effective way. Thus, a clustering algorithm simulating a visual system may solve some basic problems in these areas of research. From this point of view, we propose a new approach to data clustering by modeling the blurring effect of lateral retinal interconnections based on scale space theory. In this approach, a data set is considered as an image with each light point located at a datum position. As we blur this image, smaller light blobs merge into larger ones until the whole image becomes one light blob at a low enough level of resolution. By identifying each blob with a cluster, the blurring process generates a family of clusterings along the hierarchy. The advantages of the proposed approach are: 1) The derived algorithms are computationally stable and insensitive to initialization and they are totally free from solving difficult global optimization problems. 2) It facilitates the construction of new checks on cluster validity and provides the final clustering a significant degree of robustness to noise in data and change in scale. 3) It is more robust in cases where hyperellipsoidal partitions may not be assumed. 4) It is suitable for the task of preserving the structure and integrity of the outliers in the clustering process. 5) The clustering is highly consistent with that perceived by human eyes. 6) The new approach provides a unified framework for scale-related clustering algorithms recently derived from many different fields such as estimation theory, recurrent signal processing on self-organization feature maps, information theory and statistical mechanics, and radial basis function neural networks.

Index Terms—Hierarchical clustering, scale space theory, cluster validity.

1 Introduction

Data clustering aims at the partitioning of a given data set with known or unknown distribution into homogeneous subgroups. Such a problem is rampant in various applications such as pattern recognition, image processing, data transmission, and data storage in physical and biological systems. Literature on clustering techniques and their applications, especially in pattern recognition, is voluminous (see, for example, [1], [2], [3]).

Clustering algorithms in the literature can generally be classified into two types: hierarchical clustering and partitional clustering. The output of a hierarchical clustering algorithm is a dendrogram, which is a tree showing a sequence of clusterings with each clustering being a partition of the data set [4], [5], [6], [7], [8]. According to the structure adopted, hierarchical clusterings may be further categorized into *nested hierarchical clustering* and *nonnested hierarchical clustering*. In nested hierarchical clustering, each small cluster fits itself in whole inside a larger cluster at the merging scale (or threshold) and every datum is not permitted to change cluster membership once assignment has been made. In nonnested hierarchical clustering, a cluster obtained at small scale may divide itself into several parts and fit these parts into different clusters at the merging scale and, therefore, each datum is allowed to change its cluster membership as scale varies. The algorithms proposed in [3], [9], [10], [11], [12], [13] all generate nonnested hierarchical clusterings, while clusterings generated by SLINK [4], [5], COMLINK [4], [5], [6], [7], and MSTCLUS [4], [8], as well as those in [14], [15], are all nested hierarchical clusterings.

The partitional clustering techniques partition the data set into a small number of clusters. Unlike results obtained by hierarchical techniques, output of a partitional clustering algorithm is only a single partition of the data set. The majority of partitional algorithms obtain the partition though the minimization of some suitable measures such as the cost functions. K-means clustering [4], [16], FORGY [4], [16], ISODATA [4], [16], [17], WISH [4], [16], and fuzzy ISODATA [18], for example, are essentially based on the minimization of a square-error function. Since the minimization problems involved are, in general, NP-hard and combinatorial in nature, many techniques, such as simulated annealing [19], deterministic annealing [20], and EM algorithm [21], are often required to solve them with lower computation overhead.

There are several commonly recognized deficiencies in the existing clustering methods:

1. Clustering results are sensitive to initialization. Different initial configurations may lead to different partitions due to multimodality of the cost function.
2. Global optimum is not guaranteed when global minimization problems are involved.
3. Perhaps most importantly, the algorithms do not provide formal cluster validity checks (i.e., they do not entertain questions such as:
   a. Do the data exhibit a predisposition to cluster?
   b. How many clusters are present in the data?
   c. Are the clusters yielded real or merely artifacts of algorithms?

- *Y. Leung is with the Department of Geography, Center for Environmental Policy and Resource Management and Joint Laboratory for Geoinformation Science, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: yeeleung@cuhk.edu.hk.*
- *J.-S. Zhang and Z.-B. Xu are with the Institute for Information and System Sciences, Faculty of Sciences Xi'an, Jiaotong University, Xi'an, 710049, P.R. China. Email: {jszhang, zbxu}@xjtu.edu.cn.*

Manuscript received 9 Nov. 1998; revised 19 Nov. 1999; accepted 13 Apr. 2000.
Recommended for acceptance by K. Bowyer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 108214.

d. Which partition or which individual cluster is valid?).

Cluster validity is a vexing but very important problem in cluster analysis because each clustering algorithm always finds clusters even if the data set is entirely random. While many clustering algorithms can be applied to a given problem, there is in general no guarantee that any two algorithms will produce consistent answers. This should make cluster validity check an essential requirement of any algorithm. One widely used strategy is to employ visual processing to examine distributions on each separate variable by ways such as histograms and nonparametric density estimates and plots of each pair of variables using scattergram. However, visual processing is intuitively employed in these techniques without any theoretical basis. Another developed strategy out of this difficulty is to produce clustering algorithms based directly on the laws of psychology of *form perception*. Zahn [8] has proposed a clustering algorithm based on the laws of Gestalt psychology of form perception. The algorithm is a graphical one which is based on the minimal spanning tree and attempts to mechanize the Gestalt law of *proximity*, which says that perceptual organization favors groupings representing smaller interpoint distance. Zahn's algorithm has a strong influence on cluster analysis. Many algorithms have been developed on the basis of similar ideas. However, Zahn's algorithm is derived from Gestalt psychology laws in a heuristic way since Gestalt laws cannot be represented in an accurate computational model. This inaccuracy makes it difficult to establish a formal and efficient cluster validity check.

In recent years, physiological discoveries and researches in computer-aided neuroanatomy have advanced several quite accurate computational models of primary visual system, each modeling some parts of the human visual system at a particular level of details. Among them is scale space theory, which models the blurring effect of lateral retinal interconnection by applying Gaussian filtering to a digital image [22], [23], [24], [25], [26], [27], [28]. In fact, the theory sheds light on the way we cluster data, regardless of whether they are digital images or raw data. It also renders a biological perspective on data clustering. Through evolution and training, our visual system has become optimal in the clustering of images. Therefore, clustering of nonimage or high-dimensional data should more or less bear the blueprint of the visual system and is directly or indirectly influenced by the way we cluster image data. Hence, we may expect that psychovisual criteria provide relevant guides to our clustering of nonimages or high-dimensional data.

The purpose of this paper is to develop a new approach to data clustering based on scale space theory. In our approach, a data set can be considered as an image with each datum being a light point attached with a uniform luminous flux. As we blur this image, each datum first becomes a light blob. Throughout the blurring process, smaller blobs merge into larger ones until the whole image contains only one light blob at a low enough level of resolution. If we equate each blob with a cluster, the above blurring process will generate a hierarchical clustering with resolution being the height of a dendrogram. The blurring process is described by scale-space filtering.

The proposed approach has several advantages:

1. The algorithms thus derived are computationally stable and insensitive to initialization. They are totally free from solving difficult global optimization problems.
2. It facilitates the formulation of new cluster validity checks and gives the final clustering a significant degree of robustness to noise in the data and change in scale.
3. It is more robust where hyperellipsoidal partitions may not be assumed.
4. It is suitable for the task of preserving the structure and integrity of the outliers in the clustering process.
5. The patterns of clustering are highly consistent with the perception of human eyes.
6. It provides a unified generalization of the diversely derived algorithms in [11], [12], [13], [14], [15].

In Section 2, we describe briefly the scale space theory and show how we can relate this theory to data clustering. Extended on the scale space theory, we construct in Section 3 the theory and algorithms of hierarchical clusterings. Cluster validity checks and procedures for the selection of "good" clusterings are presented in Section 4. In Section 5, a numerical simulation and an application in multidimensional clustering are used to illustrate the performance of the algorithms. Relationships between the proposed algorithms and other scale-based algorithms are discussed in Section 6. To substantiate our theoretical arguments, some applications of the theory are presented in Section 7. A summary is given in the final section to conclude the paper.

2 Scale Space Theory

Let us first consider a two-dimensional image given by a continuous mapping $p(x) : R^2 \mapsto R$. In scale space theory, $p(x)$ is embedded into a continuous family $P(x, \sigma)$ of gradually smoother versions of it. The original image corresponds to the scale $\sigma = 0$ and increasing the scale should simplify the image without creating spurious structures. If there are no prior assumptions that are specific to the scene, then it is proven that one can blur the image in a unique and sensible way in which $P(x, \sigma)$ is the convolution of $p(x)$ with the Gaussian kernel, i.e.,

$$P(x,\sigma) = p(x) * g(x,\sigma) = \int p(x-y)\frac{1}{(\sigma^2 2\pi)} e^{-\frac{\|y\|^2}{2\sigma^2}} dy, \quad (1)$$

where $g(x, \sigma)$ is the Gaussian function $g(x,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x\|^2}{2\sigma^2}}$, σ is the scale parameter, (x, σ)-plane is the scale space, and $P(x, \sigma)$ is the scale space image. It should be noted that there is a direct relation with neurophysiological findings in animals and psychophysics in man supporting this theory [29].

For each maxima $y \in R^2$ of $p(x)$, we define the corresponding light blob as being a region specified as follows:

$$B_y = \{x_0 \in R^2 : \lim_{t\to\infty} x(t, x_0) = y\}, \quad (2)$$

where $x(t, x_0)$ is the solution of the gradient dynamic system

$$\begin{cases} \frac{dx}{dt} = \bigtriangledown_x p(x) \\ x(0) = x_0. \end{cases} \tag{3}$$

In what follows, y is referred to as the *blob center* of B_y. All blobs in an image produce a partition of R^2 with each point belonging to a unique blob except the boundary points.

Remark 1. For a given function $p(x)$, its magnitude scaling is defined as $f(p(x))$ with f being a strictly increasing function. By the fact that $f(p(x_1)) < f(p(x_2))$ if and only if $p(x_1) < p(x_2)$, we know that the blob and blob center defined by (2) and (3) is invariant to this transformation. This is consistent with the contrast invariance assumption in visual processing [30]. In what follows, the logarithmic scaling is often used in the implementation so that the gradient vector can be computed more stably.

Let $p(x) = g(x, \sigma)$, which contains only one blob for $\sigma > 0$. As $\sigma \to 0$, this blob concentrates on a light point defined as

$$\delta(x) = \lim_{\sigma \to 0} g(x, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x\|^2}{2\sigma^2}}. \tag{4}$$

Mathematically, such a function is called a δ function or a generalized function.

A light point at $x_0 \in R^2$ in an image is defined as a δ function situated at x_0, i.e., $\delta(x - x_0)$, and $\delta(x - x_0)$ satisfies

$$g(x, \sigma) * \delta(x - x_0) = g(x - x_0, \sigma), \tag{5}$$

where g is the Gaussian function. From (5), we can see that if we blur a light point, it becomes a light blob again.

In our everyday visual experience, blurring of an image leads to the erosion of structure: Small blobs always merge into large ones and new ones are never created. Therefore, the blobs obtained for images $P(x, \sigma)$ at different scales form a hierarchical structure: Each blob has its own survival range of scale; large blobs are made up of small blobs. The survival range for a blob is characterized by the scale at which the blob is formed and the scale at which the blob merges with others. Each blob manifests itself purely as a simple blob within its survival range of scale.

We now relate such a blurring process with the process of clustering.

If $p(x)$ is a probability density function from which the data set is generated, then each blob is a connected region containing a relatively high density probability, separated from other blobs by a boundary with relatively low-density probability. Therefore, each blob is a cluster as defined in [31]. All blobs together produce a classification of a data space which provides a clustering for the data set with known distribution $p(x)$.

For a given data set $X = \{x_i \in R^2 : i = 1, \ldots, N\}$, the empirical distribution for the data set X can be expressed as

$$\widehat{p}_{emp}(x) = \frac{1}{N} \sum_{i=0}^{N} \delta(x - x_i). \tag{6}$$

The image corresponding to $\widehat{p}_{emp}(x)$ consists of a set of light points situated at the data set, just like a scattergram of the data set. When we blur this image, we get a family of smooth images $P(x, \sigma)$ represented as follows:

$$P(x, \sigma) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x - x_i\|^2}{2\sigma^2}}. \tag{7}$$

The family $P(x, \sigma)$ can be considered as the Parzen estimation with Gaussian window function. At each given scale σ, the scale space image $P(x, \sigma)$ is a smooth distribution function so that the blobs and their centers can be determined by analyzing the limit of the solution $x(t, x_0)$ of the following differential equation:

$$\begin{cases} \frac{dx}{dt} = \bigtriangledown_x P(x, \sigma) = \frac{1}{\sigma^2 N} \sum_{i=1}^{N} \frac{(x_i - x)}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x - x_i\|^2}{2\sigma^2}} \\ x(0) = x_0. \end{cases} \tag{8}$$

When a distribution $p(x)$ is known, but contains noise or is indifferentiable, we can also use scale space filtering method to erase the spurious maxima generated by the noise. In this case, the scale-space image is

$$P(x, \sigma) = p(x) * g(x, \sigma) = \int \frac{p(y)}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x - y\|^2}{2\sigma^2}} dy \tag{9}$$

and the corresponding gradient dynamical system is given by:

$$\begin{cases} \frac{dx}{dt} = \bigtriangledown_x P(x, \sigma) = \int \frac{p(y)(y - x)}{(\sigma\sqrt{2\pi})^2 \sigma^2} e^{-\frac{\|x - y\|^2}{2\sigma^2}} dy \\ x(0) = x_0. \end{cases} \tag{10}$$

When the noise in $p(x)$ is an independent white noise process, (9) provides an optimal estimate of the real distribution [13].

By considering the data points falling into the same blob as a cluster, the blobs of $P(x, \sigma)$ at a given scale produce a pattern of clustering. In this way, each data point is deterministically assigned to a cluster via the differential gradient dynamical equation in (8) or (10) and, thus, our proposed scheme is a hard clustering method. As we change the scale, we get a hierarchical clustering. In what follows, we give a detailed description of the clustering procedure and the corresponding numerical implementations.

3 Hierarchical Clustering in Scale Space

In scale-space clustering, we use the maxima of $P(x, \sigma)$ with respect to x as the description primitives. Our discussion is based on the following theorem whose proof is omitted here due to space limitation (proof can be obtained from the authors).

Theorem 1. *For almost all data sets, we have: 1) zero is a regular value of $\bigtriangledown_x P(x, \sigma)$, 2) as $\sigma \to 0$, the clustering obtained for $P(x, \sigma)$ with $\sigma > 0$ induces a clustering at $\sigma = 0$ in which each datum is a cluster and the corresponding partition is a Voronoi tessellation, i.e., each point in the scale space belongs to its nearest-neighbor datum, and 3) as σ increases from $\sigma = 0$, there are N maximal curves in the scale space with each of them starting from a datum of the data set.*

We know that the maxima of $P(x,\sigma)$ are the points satisfying:

$$\bigtriangledown_x P(x,\sigma) = 0. \tag{11}$$

Therefore, 0 being a regular value of $\bigtriangledown_x P(x,\sigma)$ means that: 1) All maxima form simple curves in the scale space. 2) We can follow these curves by numerical continuational method [32]. 3) In terms of the criterion for cluster centers (i.e., maximizing $P(x,\sigma)$), there is a unique solution at small scale with N centers and, hence, the method is independent of initialization. In the following discussion, we always assume that 0 is a regular value of $\bigtriangledown_x P(x,\sigma)$.

3.1 Nested Hierarchical Clustering

For convenience purposes, we call each maximum the blob center of the corresponding cluster (or blob) in the following discussion.

The construction procedure of a nested hierarchical clustering based on the scale-space image is as follows:

1. At scale $\sigma = 0$, each datum is considered as a blob center whose associated data point is itself.
2. As σ increases continuously, if the blob center of a cluster moves continuously along the maximal curve and no other blob center is siphoned into its blob, then we consider that the cluster has not changed and only its blob center moves along the maximal curve. If an existing blob center disappears at a singular scale and falls into another blob, then the two blobs merge into one blob and a new cluster is formed with the associated data points being the union of those of the original clusters.
3. Increase the scale until the whole data set becomes one single cluster. This stopping rule is well-defined because we have only one blob in the data space when scale is large enough.

In this way, a hierarchical clustering dendrogram is constructed with scale as height. Such a hierarchical clustering dendrogram may be viewed as a regional tree with each of its nodes being a region so that data falling within the same region form a cluster. Therefore, the nested hierarchical clustering thus constructed provides a classification of the data space. In the one-dimensional case, such a regional tree is in fact an interval tree, as is shown in Fig. 4c.

3.2 Nonnested Hierarchical Clustering

Nested hierarchical clustering has been criticized for the fact that, once a cluster is formed, its members cannot be separated subsequently. Nevertheless, we can construct a nonnested hierarchical clustering which removes such a problem. In a nonnested hierarchical clustering, we partition the data set $X = \{x\}$ at a given scale by assigning a membership to each datum $x_0 \in X$ according to (2). This process is similar to how we perceive the data set at a given distance or a given resolution. Clusterings obtained at different scales are related to each other by the cluster center lines. As σ changes, a nonnested hierarchical clustering is obtained since each datum may change its membership under such a scheme. The evolution of the cluster centers in the scale-space image may be considered as a form of dendrogram. By Theorem, 1 we know that 0 is a regular value of $\bigtriangledown_x P(x,\sigma)$ for almost all data sets. This means that cluster centers form simple curves in the scale space which can be computed through the path that follows the solutions of the equation $\bigtriangledown_x P(x,\sigma) = 0$ by the continuational method [32].

Nonnested hierarchical clustering is more consistent with that obtained by the human eye at different distance or different resolution, while nested hierarchical clustering has a more elegant hierarchical structure.

3.3 Numerical Solution for Gradient Dynamic System

In the proposed clustering method, clusters are characterized by the maxima of $P(x,\sigma)$ and the membership of each datum is determined by the gradient dynamical system in (8) or (10). Since the solution of the initial value problem of (8) or (10) cannot be found analytically, some numerical methods must be used. If the Euler difference method is used, the solution of (8) or (10), $x(t,x_0)$, is then approximated by the sequence $\{x(n)\}$ generated in one of the following difference equations:

$$\begin{cases} x(n+1) = x(n) + h \bigtriangledown_x p(x(n),\sigma) \\ \qquad\quad = x(n) + \dfrac{h}{\sigma^2 N} \displaystyle\sum_{i=1}^{N} \frac{(x_i - x(n))}{\left(\sigma\sqrt{2\pi}\right)^2} e^{-\frac{\|x(n)-x_i\|^2}{2\sigma^2}} \\ x(0) = x_0, \end{cases} \tag{12}$$

or

$$\begin{cases} x(n+1) = x(n) + \frac{h}{\sigma^2} \int p(y)(y - x(n)) e^{\frac{-\|x(n)-y\|^2}{2\sigma^2}} dy \\ x(0) = x_0, \end{cases} \tag{13}$$

where h is the step length.

If the magnitude of P is scaled by the logarithmic function, i.e., $lg(P)$, the corresponding gradient dynamical system of (8) and (10) becomes:

$$\frac{dx}{dt} = \frac{1}{\sigma^2} \frac{\displaystyle\sum_{i=1}^{N} (x_i - x) e^{-\frac{\|x-x_i\|^2}{2\sigma^2}}}{\displaystyle\sum_{i=1}^{N} e^{-\frac{\|x-x_i\|^2}{2\sigma^2}}}, \tag{14}$$

and

$$\frac{dx}{dt} = \frac{1}{\sigma^2} \frac{\int p(y)(y - x) e^{\frac{-\|x-y\|^2}{2\sigma^2}} dy}{\int p(y) e^{-\frac{\|x-y\|^2}{2\sigma^2}} dy}, \tag{15}$$

and the discrete approximations to (14) and (15) then become:

$$x(n+1) = x(n) + \frac{h}{\sigma^2} \frac{\displaystyle\sum_{i=1}^{N} (x_i - x(n)) e^{-\frac{\|x(n)-x_i\|^2}{2\sigma^2}}}{\displaystyle\sum_{i=1}^{N} e^{-\frac{\|x(n)-x_i\|^2}{2\sigma^2}}}, \tag{16}$$

or

$$x(n+1) = x(n) + \frac{h}{\sigma^2} \frac{\int p(y)(y - x(n)) e^{-\frac{\|x(n)-y\|^2}{2\sigma^2}} dy}{\int p(y) e^{-\frac{\|x(n)-y\|^2}{2\sigma^2}} dy}. \tag{17}$$

Setting the step length $h = \sigma^2$ in (17), we get

$$x(n+1) = \frac{\sum_{i=1}^{N} x_i e^{-\frac{\|x(n)-x_i\|^2}{2\sigma^2}}}{\sum_{i=1}^{N} e^{-\frac{\|x(n)-x_i\|^2}{2\sigma^2}}}. \quad (18)$$

Such an iteration can be interpreted as an iterative local centroid estimation [13], [33], [34], [35], [36].

When the size of the data set is large or the data are given in a serial form, we can use the stochastic gradient descent algorithm to search the blob center and determine the memberships of the data. In fact, our aim is to find the maximum of $P(x, \sigma)$ which can be represented as (see (9)):

$$P(x, \sigma) = E\{e^{\frac{-\|x-y\|^2}{2\sigma^2}}\}, \quad (19)$$

where $E\{\cdot\}$ is the expectation of the density of the data set y. By the theory of stochastic gradient descent algorithm, the blob center of a datum x_0 can be obtained by the following iteration initialized at x_0:

$$x(n+1) = x(n) + h^{(n)}(x^{(n)} - x(n))e^{-\frac{\|x(n)-x^{(n)}\|^2}{2\sigma^2}}, \quad (20)$$

where $x^{(n)}$ is the nth randomly chosen member of X or the nth datum generated according to the distribution $p(x)$ to be presented to the algorithm and $h^{(n)}$ is the adaptive step length chosen as

$$h^{(n)} = \frac{1}{1+n}. \quad (21)$$

Finally, we associate the datum x_0 with a center x^* if $x(n)$ initialized from x_0 converges to x^*. In practice, we define $x(n+1)$ as a blob center if $\|x(n+1) - x(n)\| < \epsilon$ or $\| \bigtriangledown_x p(x(n+1)\| < \epsilon$, where ϵ is a small positive value which may vary with problems. If two centers x_1 and x_2 satisfy $\|x_1 - x_2\| < \epsilon$, we consider them as one blob center.

3.4 Implementation of Hierarchical Clustering

There are several ways to implement the proposed hierarchical clustering. The first one uses the path-following algorithm to trace the blob centers along the maximal curves. When a singular scale at which a blob center disappears is encountered, we find the new blob center by solving the differential equation (8) or (10) with initial value $x_0 = x^*$ and follow the new blob center by the path-following algorithm again.

Parallel to existing algorithms in [14], [15], the second method uses the discretization of scale and an iterative scheme which works as follows:

ALGORITHM I—Nested Hierarchical Algorithm

1. Given a sequence of scales $\sigma_0, \sigma_1, \cdots$ with $\sigma_0 = 0$. At $\sigma_0 = 0$, each datum is a cluster and its blob center is itself. Let $i = 1$.
2. Find the new blob center at σ_i for each blob center obtained at scale σ_{i-1} by one of the iterative schemes in (12), (13), (14), (15), (16), (17), (18). Merge the clusters whose blob centers arrive at the same blob center into a new cluster.
3. If there are more than two clusters, let $i := i + 1$, go to 2.
4. Stop when there is only one cluster.

ALGORITHM II—Nonested Hierarchical Algorithm

1. Given a sequence of scales $\sigma_0, \sigma_1, \cdots$ with $\sigma_0 = 0$.
2. At $\sigma_0 = 0$, each datum is a cluster and its blob center is itself. Let $i = 1$.
3. Cluster the data at σ_i. Find the new blob center at σ_i for each blob center obtained at scale σ_{i-1} by one of the iterative schemes in (12), (13), (14), (15), (16), (17), (18). If two new blob centers arrive at the same point, we consider that the old clusters disappear and a new cluster is formed.
4. If there are more than two clusters, let $i := i + 1$, go to 2.
5. Stop when there is only one cluster.

When the size of the data set is very large, we can substitute each datum in the iterative scheme in (12), (13), (14), (15), (16), (17), (18) with its blob center and σ_i with $\sigma_i - \sigma_{i-1}$ in Step 2 to reduce the computational cost of the above algorithm. For example, (18) becomes

$$x(n+1) = \frac{\sum_{j=1}^{N_i} k_j p_j e^{-\frac{\|x(n)-p_j\|^2}{2\sigma^2}}}{\sum_{j=1}^{N_i} k_j e^{-\frac{\|x(n)-p_j\|^2}{2\sigma^2}}}, \quad (22)$$

where p_j is blob center j obtained at scale σ_i, N_i is the number of p_j, k_j is the number of data points in the blob whose center is p_j and $\sigma = \sigma_i - \sigma_{i-1}$. Since N_i is usually much smaller than N, the computational cost can be reduced significantly.

In practical applications, σ_i should increase according to

$$\sigma_i - \sigma_{i-1} = k\sigma_{i-1}. \quad (23)$$

This comes from the requirement of accuracy and stability of the representation, as proven in [24]. In psychophysics, Weber's law says that the minimal size of the difference $\triangle I$ in stimulus intensity which can be sensed is related to the magnitude of standard stimulus intensity I by $\triangle I = kI$, where k is a constant called Weber fraction. Therefore, psychophysical experimental results may be used to propose a low bound for k in the algorithms since we cannot sense the difference between two images $p(x, \sigma_{i-1})$ and $p(x, \sigma_i)$ when k is less than its Weber fraction. For instance, $k = 0.029$ in (23) is enough in one-dimensional applications because scale σ is the window length in the scale space and the Weber fraction for line length is 0.029 [30].

Other implementations of our proposed hierarchical clustering may include the algorithms proposed in [11], [12], [13], [14], [15] and are not elaborated here.

4 Cluster Validity

For any given data set X, we can always construct hierarchical clusterings by the algorithms previously proposed, even though there is no structure inside the data. Therefore, if we wish to successfully apply these algorithms to practical problems, we should first answer the

cluster validity questions raised in the introduction. Literature trying to answer the validity questions for various clustering algorithms is voluminous. In this paper, we will tackle these questions on the basis of human visual experience: The real cluster should be perceivable over a wide range of scales. This leads us to adopt the notion of "lifetime" of a cluster as its validity criterion: A cluster with longer lifetime is preferred to a cluster with shorter lifetime. Such a point of view is also supported by Witkin's empirical observation "that survive over a broad range of scale tend to leap out at the eye,..." in [22], [23].

In what follows, we first define the notion of lifetime of a cluster and lifetime of a clustering in the more general sense by including nonnested hierarchical clustering. Then, the lifetime of a cluster is used to test the "goodness" of a cluster and the lifetime of a clustering is used to determine the number of clusters in a specific pattern of clustering.

4.1 Lifetime, Compactness, Isolation, and Outlierness

We first define the lifetime of a cluster and a clustering and then show why logarithmic scale is used to measure lifetime.

Definition 2. Lifetime of a cluster *is defined as the range of logarithmic scales over which the cluster survives, i.e., the logarithmic difference between the point when the cluster is formed and the point when the cluster is absorbed into or merged with other clusters.*

Each pattern of clustering in a nonnested hierarchical clustering only consists of clusters which are formed at the same scale. A pattern of clustering in a nested hierarchical clustering, however, is a partition of the data set X which may consist of clusters obtained at the same scale or at different scales. In what follows, we define the lifetime for these two kinds of clusterings, respectively.

Definition 3. *Let $\pi(\sigma)$ be the number of clusters in a clustering achieved at a given scale σ. Suppose C_σ is a clustering obtained at σ with $\pi(\sigma) = m$. The* σ-lifetime of C_σ *is defined as the supremum of the logarithmic difference between two scales within which $\pi(\sigma) = m$.*

Definition 4. *Suppose a clustering C in a hierarchical clustering contains K clusters $\{C_1, \cdots, C_K\}$. Denote the number of data points in C_i by $|C_i|$ and the lifetime of C_i by l_i. Then, the* mean lifetime of all clusters in clustering C *is defined as*

$$\sum_{i=1}^{K} l_i \frac{|C_i|}{|X|}. \tag{24}$$

The lifetime of clustering C *is the mean lifetime of all its clusters. If a cluster C_i is further divided into K_i subclusters $\{C_{i_1}, \cdots, C_{i_K}\}$ and the lifetime of C_{i_j} is denoted by l_{i_j}, then* the mean lifetime of all its subclusters *is defined as*

$$\sum_{j=1}^{K_i} l_{i_j} \frac{|C_{i_j}|}{|C_i|}. \tag{25}$$

Now, we interpret why logarithmic scale is used in the above definitions.

The experimental tests in [11] show that $\pi(\sigma)$ decays with scale σ according to:

$$\pi(\sigma) = ce^{-\beta\sigma} \tag{26}$$

if the data are uniformly distributed, where β is a positive constant related to the dimensionality of the data space. If a data structure exists, then $\pi(\sigma)$ is a constant over a range of scales. So, the stability of $\pi(\sigma)$ can be used as a criterion to test whether the data tend to cluster, i.e., have a structure. However, β is unknown and $\pi(\sigma)$ is only allowed to take integers and, from (26), we can see that, even for a uniformly distributed data set, if β is small, $\pi(\sigma)$ will be a constant over a wide range of scales for a small σ; if β is large, $\pi(\sigma)$ will also be a constant over a wide range of scales for a large σ. This makes it difficult to find the structure in the $\pi(\sigma)$ plot. However, if the data are uniformly distributed and we rescale σ by a new parameter k such that the number of clusters in the clustering obtained at the new parameter k, denoted by $\pi(k)$, decays linearly with respect to k, i.e.,

$$\pi(k) = \pi(0) - k, \tag{27}$$

we can easily find the structure in the plot of $\pi(k)$. The reason is that it is much simpler to test whether $\pi(k)$ decays linearly with respect to k than to test whether $\pi(\sigma)$ decays according to (26), in which an unknown parameter β is involved.

Now, we derive the relationship of k and σ under the assumption that $\pi(k)$ decays linearly with respect to k.

Suppose σ relates to k through a function $\sigma(k)$. Then, we have

$$\pi(k) = \pi(\sigma(k)) = ce^{-\beta\sigma(k)}. \tag{28}$$

Under the assumption that $\pi(k)$ decays linearly with respect to k, see (27), we know that

$$\frac{d\pi(k)}{dk} = -1. \tag{29}$$

From (26), we get

$$\frac{d\pi(k)}{dk} = -c\beta e^{-\beta\sigma}\frac{d\sigma}{dk}. \tag{30}$$

Equations (29) and (30) imply that the new parameter k should satisfy

$$\frac{d\sigma}{dk} = \frac{1}{c\beta}e^{\beta\sigma}. \tag{31}$$

Solving this differential equation, we get

$$k = c(1 - e^{-\beta\sigma}). \tag{32}$$

Such a scaling is an ideal one, but it contains a parameter β which is usually unknown. In practice, we take the approximation $\frac{\beta}{e^{\beta\sigma}} = \frac{\beta}{1+\beta\sigma+\cdots} \approx \frac{1}{\sigma}$ in (30), which does not contain the unknown parameter β, and this leads to the logarithmic scale

$$k = clog\frac{\sigma}{\varepsilon}, \tag{33}$$

where ε is a positive constant.

The term k defined in (33) is called the *sensation intensity* under Fechner's Law [30].

In terms of the new parameter k, lifetime should be measured by the logarithmic scale of σ. While such a scaling is used in [14], [15], no explanation such as the one furnished above is provided in these papers.

Once a partition has been established to be valid, a natural question that follows is how good are the individual clusters.

The first suggested measure of "goodness" of a cluster is naturally its lifetime: A good cluster should have a long lifetime. The other suggested measures are compactness and isolation.

Intuitively, a cluster is good if the distance between the data inside the cluster are small and those outside are large. To make this idea operational, we define two measures for the identification of good clusters. They are the *compactness* and *isolation* of a cluster, parallel to similar notions in [14], [15]. For a cluster C_i, they are defined as follows:

$$isolation = \frac{\sum_{x \in C_i} e^{-\|x-p_i\|^2/2\sigma^2}}{\sum_x e^{-\|x-p_i\|^2/2\sigma^2}}, \tag{34}$$

$$compactness = \frac{\sum_{x \in C_i} e^{-\|x-p_i\|^2/2\sigma^2}}{\sum_{x \in C_i} \sum_j e^{-\|x-p_j\|^2/2\sigma^2}}. \tag{35}$$

where p_i is the blob center of cluster C_i. For a good cluster, the compactness and isolation are close to one. This measure is dependent on the scale and will be used to find the optimal scale at which the clustering achieved by nonnested hierarchical clustering is good.

A data set invariably contains noisy data points or outliers. How to detect them is an important problem in many diagnostic or monitoring systems. In the proposed scale-based clustering algorithms, we can use the number of data points in a cluster C_i and the lifetime of C_i to decide whether C_i is an outlier. If C_i contains a small number of data and survives a long time, then we say that C_i is an outlier, otherwise, C_i is a normal cluster. Therefore, we can use

$$outlierness_i = \frac{lifetime\ of\ C_i}{number\ of\ data\ in\ C_i} \tag{36}$$

as a test criterion for outliers, which means that an outlier is a well-isolated group with a small number of data in a large scale range. Since the method treats the data point as a light point, each outlier (usually with small number of data) should be a stable cluster in quite a large scale range. That is to say, an outlier in general exhibits a high degree of "outlierness" (whose threshold usually depends on the applications) and this fact may be used to exclude the outlier from estimated partition.

4.2 Clustering Selection Rules

Hierarchical clustering provides us with a sequence of clusterings. The problem is which clustering is really good? Now, we give several selection rules to choose a good clustering from the sequence of clusterings in the hierarchy.

Our first rule is based on the σ-lifetime of clustering and try to find a scale at which the clustering achieved has long lifetime and high degree of compactness or isolation.

Rule I

1. Find the integer m such that the clustering obtained at σ with $\pi(\sigma) = m$ has the longest σ-lifetime.
2. a) In nested hierarchical clustering, clusterings which satisfy $\pi(\sigma) = m$ are identical to each other, so we can get a unique clustering when m is obtained.
   b) In nonnested hierarchical clusterings, clusterings obtained at two scales σ_1 and σ_2 are usually different from each other, even if $\pi(\sigma_1) = \pi(\sigma_2) = m$. Therefore, we still need a method to find the right scale at which a good clustering can be achieved when m is fixed. In the present paper, we propose a method based on the maximization of overall isolation and overall compactness, which are defined for a clustering achieved at σ with $\pi(\sigma) = m$ as follows:

$$F^{(i)}(\sigma) = \left(\sum_i^m i\text{th } isolation - m \right) \tag{37}$$

$$F^{(c)}(\sigma) = \left(\sum_i^m i\text{th } compactness - m \right), \tag{38}$$

where the ith $isolation$ and ith $compactness$ are the isolation and compactness of the ith cluster, respectively. By maximizing $F^{(i)}$ or $F^{(c)}$ under the condition that $\pi(\sigma) = m$, we can get a σ at which a partition with maximal isolation or maximal compactness is achieved. In the general case, $\pi(\sigma) = m$ is held in an interval $[\sigma_1, \sigma_2]$, therefore, we can use the gradient descent method to optimize $F^{(i)}$ or $F^{(c)}$. The gradient is given by

$$dF/d\sigma = \sum_{i=1}^m \bigtriangledown_{x_i} F dx_i/d\sigma, \tag{39}$$

where F is $F^{(i)}$ or $F^{(c)}$, x_i is the center of the ith cluster. The term $dx_i/d\sigma$ can be obtained as follows: We know that each cluster center x_i is a maximal point of $P(x, \sigma)$ which satisfies

$$\bigtriangledown_x P(x, \sigma) = 0. \tag{40}$$

Differentiating the above equation, we get

$$\bigtriangledown_{xx} P(x, \sigma) \frac{dx}{d\sigma} + \bigtriangledown_{x\sigma} P(x, \sigma) = 0. \tag{41}$$

By the definition that each cluster center x is a maximal point and under the condition that 0 is a regular value of $\bigtriangledown_x P(x, \sigma)$, we can prove that the matrix $\bigtriangledown_{xx} P(x, \sigma)$ is nonsingular. Solving (41), we obtain

$$\frac{dx}{d\sigma} = -[\bigtriangledown_{xx} P(x, \sigma)]^{-1} \bigtriangledown_{x\sigma} P(x, \sigma). \tag{42}$$

Finally, we obtain a σ which is a minimal point of $F^{(i)}$ or $F^{(c)}$ and we consider that the clustering obtained at this scale is good.

We have indicated that a clustering in a nested hierarchical clustering may consist of clusters obtained at

different scales. Our second selection rule is used to search a clustering with the longest lifetime in the nested clusterings. In what follows, we use Ω to denote the set of all clusterings in a nested hierarchical clustering. For each clustering $P_i \in \Omega$, its lifetime is denoted by l_{P_i}. With these notations, the aim of our second rule can be stated as finding a clustering P_j such that

$$l_{P_j} = max_{P_i \in \Omega} l_{P_i}. \tag{43}$$

Since such a problem is usually difficult to solve, several heuristic procedures may be used to solve it. Here, we propose two greedy methods, Rule II.1 and Rule II.2: One is to find the local maxima by a "depth-first search" and the other is by a "breadth-first search."

The first procedure is similar to Witkin's "top-level description." It works as follows:

Rule II.1 (maximization with depth-first search).

1. Initially, let P be a clustering with the whole data set as a cluster. Assign 0 as the lifetime of this unique cluster.
2. Find a cluster C_k in P whose lifetime is shorter than the mean lifetime of its children and delete the cluster C_k from P and add all children clusters of C_k into P, i.e., the new clustering P consists of the children clusters of C_k and other clusters except C_k. Repeat this process until the lifetime of each cluster in P is longer than the mean lifetime of its own children.

Clustering obtained by this procedure is usually less complex, i.e., with small number of clusters.

The second procedure can also considered as a "longest-lifetime-first" procedure [14]. It works as follows:

Rule II.2 (maximization with breadth-first search).

1. Initialize U to be an empty set. Let $C = \{C_1, C_2, \cdots, C_K\}$ be the set of all clusters in the hierarchical clustering.
2. Pick the element C_k in C with the longest lifetime and put it into U. Remove C_k and the clusters in C that are either contained in or contain C_k until C is empty. The number of elements in U is the number of clusters and U is the corresponding clustering.

To recapitulate, in scale-space clustering, we can tackle the cluster validity issues as follows:

1. If $\pi(\sigma)$ takes a constant over a wide range of the scale, we say that a valid structure exists in the data, otherwise, no structure exists in the data.
2. If the data do have a predisposition to cluster, the cluster lifetime can then be used to determine the number of clusters present in the data and the corresponding clustering by selection Rule I, Rule II.1, or Rule II.2.
3. As suggested in Rule I, we could determine the scale σ at which the clustering achieved is "real" or "good" by finding a clustering with maximal σ-lifetime and maximal overall compactness or isolation.
4. The validity of an individual cluster may be measured by its lifetime, compactness, and isolation.
5. The measure of outlierness could be used to delete noisy data points or detect outliers in a data set.

5 Some Experimental Illustrations

A large number of experiments have been performed and, due to the limitation of space, we only use two examples to illustrate the performance of the proposed algorithms.

The first is a two-dimensional data set with 150 data points which was generated using the five cluster Gaussian mixture model with different shapes. Fig. 1a is the data plot and Fig. 1b is the $\pi(k)$ plot. From Fig. 1b, we can see that $\pi(k)$ has an approximately linear decrease, with scale k between $0 < k < 60$, where $k = clog(\sigma/\epsilon)$ with $\epsilon = 0.1$ and $c = 1/log(1.05)$. For $k > 60$, the hidden data structure appears and $\pi(k) = 5$ has the longest σ-lifetime. Fig. 1c and Fig. 1d are, respectively, the overall isolation and overall compactness plots. $F^{(i)}$ and $F^{(c)}$ achieve their maxima at about $k = 67(\sigma = 2.628)$. At this scale, the clustering obtained by the nonnested hierarchical clustering algorithm is consistent with that obtained by the nested-hierarchical clustering algorithm (the corresponding clustering is shown in Fig. 2b).

Fig. 2a is the evolutionary plot of the blob centers obtained by Algorithm I. Fig. 2b is the data partition obtained at different scales. In this example, the clusterings obtained by Rule I and Rule II.1 and Rule II.2 are all identical. In the general cases, such a consistency may not hold and we should select a rule in accordance with the application.

In the second example, we apply the scale-space clustering algorithm to an actual Landsat TM image with bands to show that this algorithm is capable of effective clustering of multidimensional data.

It should be noted that if the data set $X = \{x_i \in R^n : i = 1, \cdots, N\}$ is in the space R^n, its empirical distribution is expressed as $\widehat{p}_{emp}(x) = \frac{1}{N}\sum_{i=0}^{N} \delta(x - x_i)$.The scale space image of $\widehat{p}_{emp}(x)$, $P(x, \sigma)$, can be written as

$$P_x(x, \sigma) = \frac{1}{N}\sum_{i=1}^{k}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\|x-x_i\|^2}{2\sigma^2}},$$

which is the convolution of $\widehat{p}_{emp}(x)$ with the Gaussian kernel

$$G(x, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\|x\|^2}{2\sigma^2}}.$$

Each maxima of $P(x, \sigma)$ is considered as a cluster center and a point in X is assigned to a cluster via gradient dynamic equation for $P(x, \sigma)$. Since Theorem 1 holds in any dimension, then our algorithms can straightforwardly be extended to n-dimensional data with slight adaptation.

The study area is located in the northwest of Hong Kong, Yeun Long, corresponding to an area of $230KM^2$ on the Hong Kong topographic maps with geographical coordinates ($113^0 58'E - 114^0 07'E$ to $22^0 21'N - 22^0 31'N$). The main land covers include forest, grass, rock, water, built-up area, trees, marshland, shoals, etc. They are distributed in a

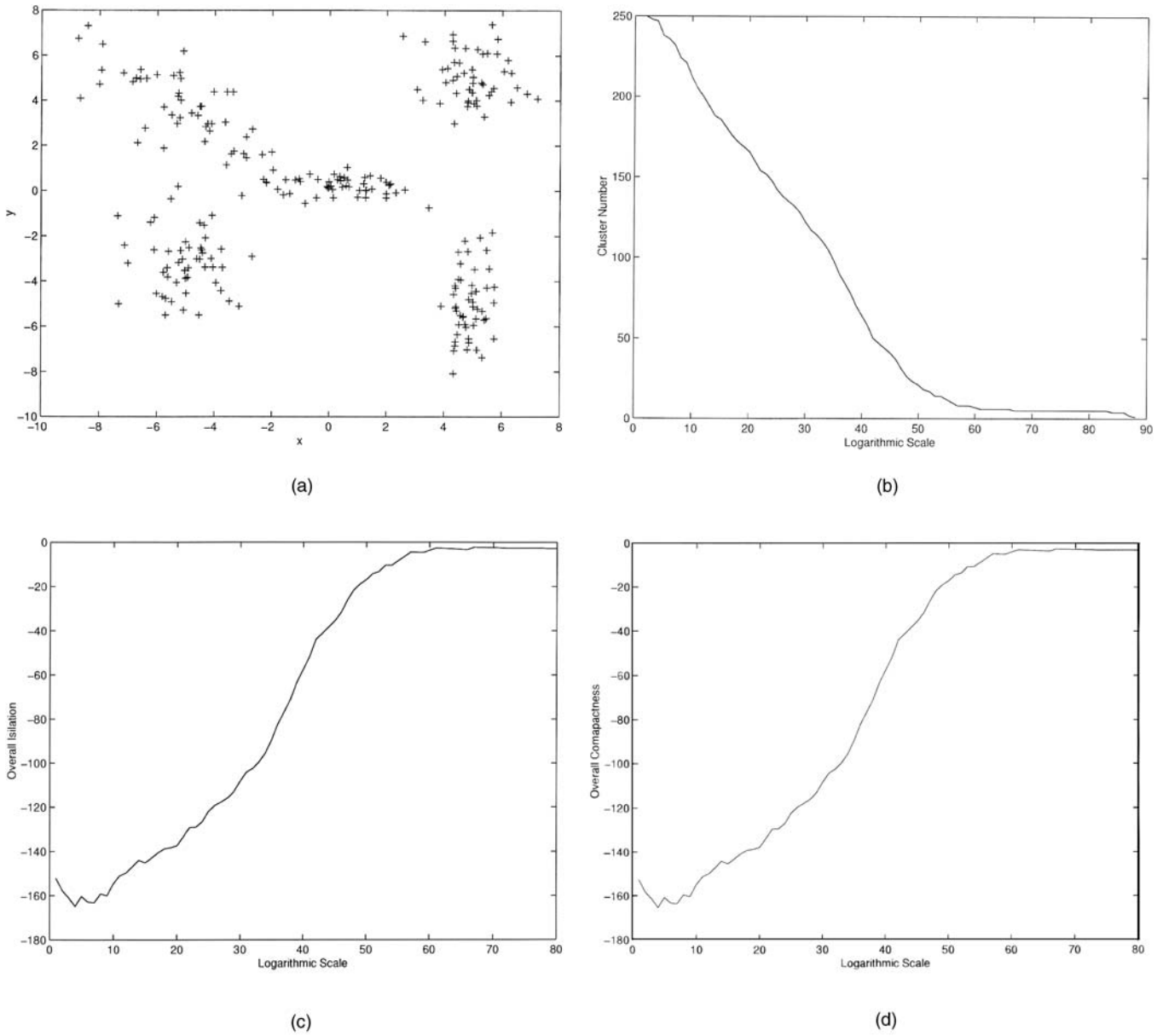


Fig. 1. (a) Plot of the data set. (b) Logarithmic-scale plot of the cluster number $\pi(k)$. (c) Logarithmic-scale plot of overall isolation. (d) Logarithmic-scale plot of overall compactness.

complex way. The Landsat TM10 image used is from 3 March 1996 with fine weather. The image size is 455×568 pixels. In our experiment, six bands, TM1, 2, 3, 4, 5, and 7, are utilized, i.e., the clustering is done in six dimensions.

In the test, we first cluster a data set consisting of 800 pixels randomly sampled from the image and then assign each pixel to its nearest cluster center. Fig. 3a is the Landsat image of Yuen Long, Hong Kong, and Fig. 3b shows the 15-cluster solution obtained by applying the scale space clustering algorithm to this image. The 15 clusters are obtained from Rule II.2 and the outliers are deleted according to their outlierness defined in (36). Compared with the ground truth, we find that the scale space clustering is capable of finding the fine land covers. For example, three classes of water bodies corresponding to deep sea water, shallow seawater, and freshwater of the studied area have respectively been identified (Fig. 3b), while they cannot be distinguished by ISODATA method. In our experiments, we also find that 150 to 1,000 sample points are usually large enough to find the land covers contained in the image.

6 The Relationships with Other Scale-Based Algorithms

Several scale-based clustering algorithms have been proposed in recent years [3], [9], [10], [11], [12], [13], [14], [15]. The scale-based algorithms in [11], [12], [13], [14], [15] are derived from very different approaches, such as estimation theory, self-organization feature mapping, information theory, and statistical mechanics, as well as radial basis function networks. Based on the algorithms developed in the present paper, we can show that these algorithms are

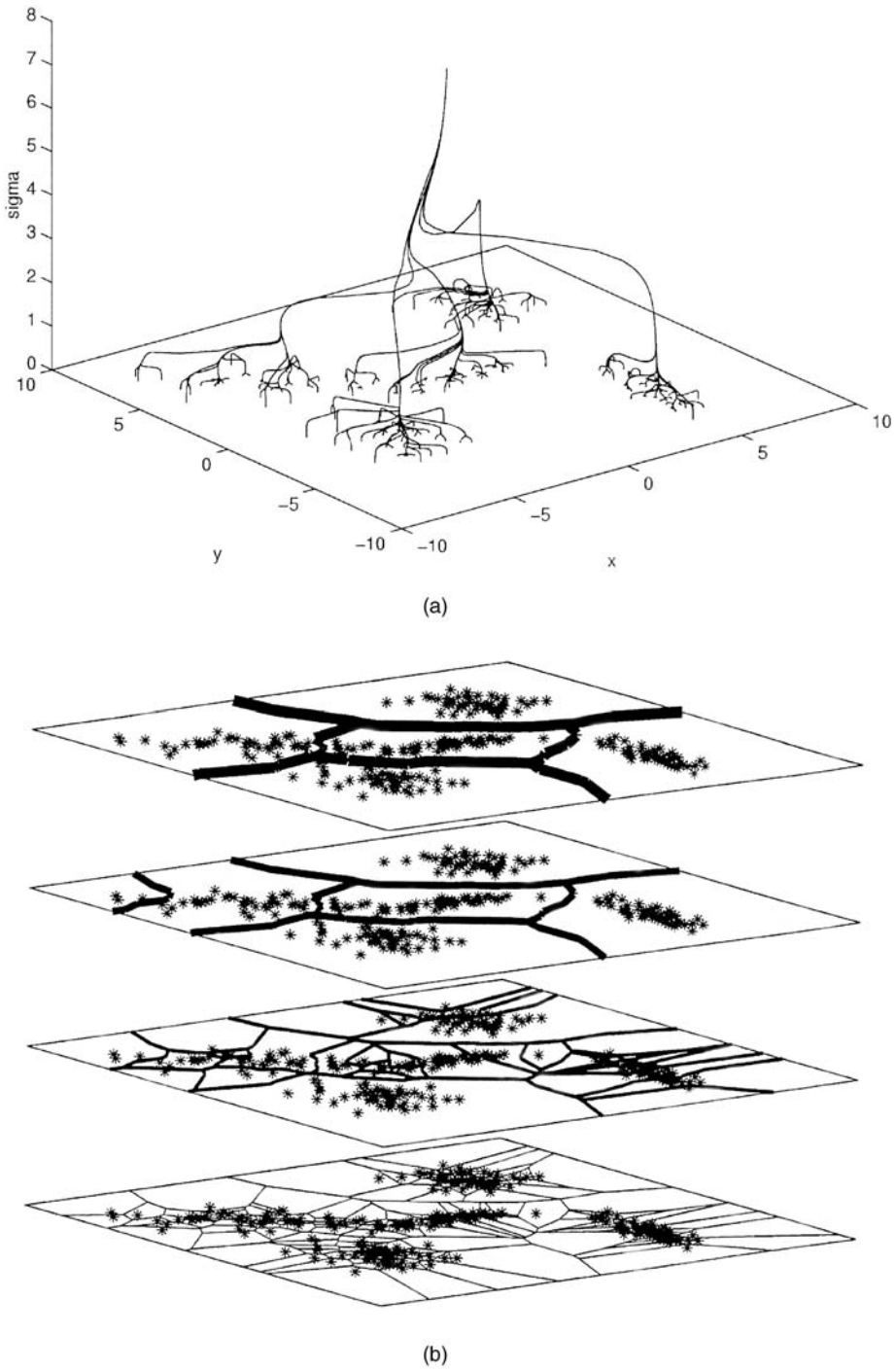


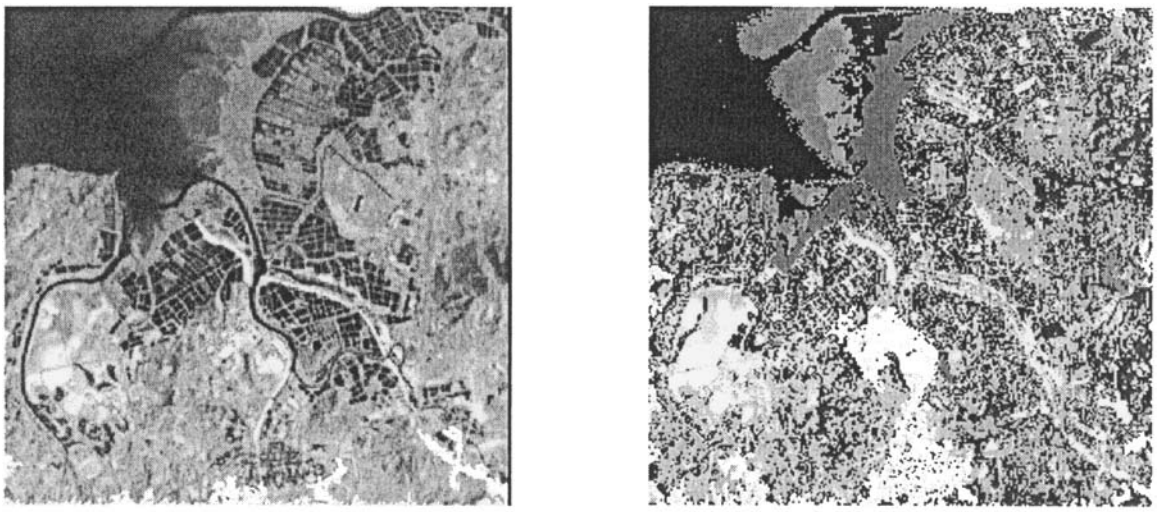
Fig. 2. (a) Evolutionary tree of cluster centers obtained by algorithm. (b) The partition of the data space obtained by the nested hierarchical clustering algorithm at scales $\sigma_0 = 0$, $\sigma_1 = 0.99$, $\sigma_2 = 2.38$, and $\sigma_3 = 2.628$ (from bottom to top).

closely related to each other and, in fact, each of these algorithms is equivalent to a special implementation of our proposed algorithm.

The iterative algorithm proposed by Wilson and Spann [13] is based on the estimation theory. This algorithm is equivalent to using the iterative procedure in (13) with $h = \sigma^2$ to find the cluster center and assign membership to the data. The idea in [13] was further developed by Roberts [11], which is based directly on the computation of maxima of $P(x, \sigma)$ and the stability of $\pi(\sigma)$ is used to check the cluster validity [11]. Therefore, both of these algorithms are the implementations of the nonnested hierarchical clustering algorithm discussed in Section 3.1.

The algorithm proposed Taven et al. [12] is derived from self-organization feature mapping and is equivalent to using (22) in the algorithm with $k_j = 1$. Since each blob may contain different number of data points, then (22) should be more reasonable for clustering.

Wong's algorithm [14], on the other hand, is based on information theory and statistical mechanics. This algorithm



(a) (b)

Fig. 3. (a) 455×568 Landsat image of Yuen Long, Hong Kong. (b) Clustering result of multispectral Ladsat image shown by the scale-space clustering algorithm.

constructs a nested hierarchical clustering through the use of a special iterative scheme, (18), in Algorithm I and selects the number of clusters by Rule II.2. Therefore, our method can be considered as a generalization of Wong's algorithm.

The algorithm proposed by Chakravarthy and Ghosh [15] is derived from the radial basis function neural network. This algorithm constructs a hierarchical clustering in a way very similar to Wong's method, but the stochastic gradient descent procedure in (20) is used.

The above discussion shows that the proposed algorithms are the natural extensions of these algorithms, but they at the same time provide a unified framework. It appears that scale-space clustering can be applied to diverse fields of research. Several other scale-based algorithms have also been proposed recently. In the adaptive K-means algorithm introduced in [10], a scale parameter r is used in the K-means algorithm as a limit to when a new cluster should start. Another scale-based algorithm called tree-structured deterministic annealing method is proposed in [9]. This probabilistic algorithm uses the minimum cross-entropy inference to solve the clustering problem subject to a tree structure. The third scale-based algorithm is based on the physical properties of an inhomogeneous ferromagnet. Both algorithms use temperature as the scale parameter. Except for the use of an explicit scale parameter, these works, in fact, have little relevance not only to each other, but also to the other scale-based algorithms in [11], [12], [13], [14], [15] and our proposed algorithms.

7 Several Theoretic Applications

In this paper, we have derived a clustering method directly from one of the computational vision models: the scale-space filtering theory. Thus, many theoretical results developed in this theory and visual systems can be used as tools to devise new algorithms and analyze related clustering algorithms. In this section, we will show, as a demonstration how, to

1. construct an interval tree for one-dimensional data set based on the simplicity of maximal curves which can be derived from Theorem 1 (Section 7.1);
2. interpret why a pitchfork merging is seldom observed in scale-space clustering [14], [15] based on the conclusion of Theorem 1 (Section 7.2);
3. correct a theoretical result obtained by Roberts [11] through a counter example given in [36], [37] (Section 7.3);
4. determine the discrete schedule of scale parameter based on a psychophysical law (Section 7.4).

7.1 Construction of Interval Clustering Tree in One-Dimensional Case

Parallel to Witkin's work [22], [23], we can construct an interval clustering tree for a one-dimensional data set based on scale space theory. Without loss of generality, we assume that the data set is the whole x-axis and the scale-space image is $p(x, \sigma)$. From the theory of scale space filtering, the minima of $p(x, \sigma)$ form simple curves in the $(x, \sigma\text{-plane})$. Each curve is rooted at $\sigma = 0$, grows monotonically to infinity, or disappears at some scale. This result can be obtained under the assumption that 0 is a regular value of dp/dx. At a given scale, there is an unique maximum to which every datum between two neighboring minima converges. This allow us to form a clustering at each given scale by one of the following rules: 1) The points fall into the interval bounded by the roots of two neighboring minima is a cluster and 2) the points fall into the interval bounded by two neighboring minima is a cluster.

Based on the simplicity of the minimal curves, the first rule results in a nested interval hierarchical clustering (see Fig. 4c), while the second rule generates a nonnested hierarchical clustering (see Fig. 4d). We call such a hierarchical clustering a clustering interval tree. The clustering interval tree differs from the interval tree proposed by Witkin in purpose and substance. While Witkin's interval tree is used to describe the signal, the

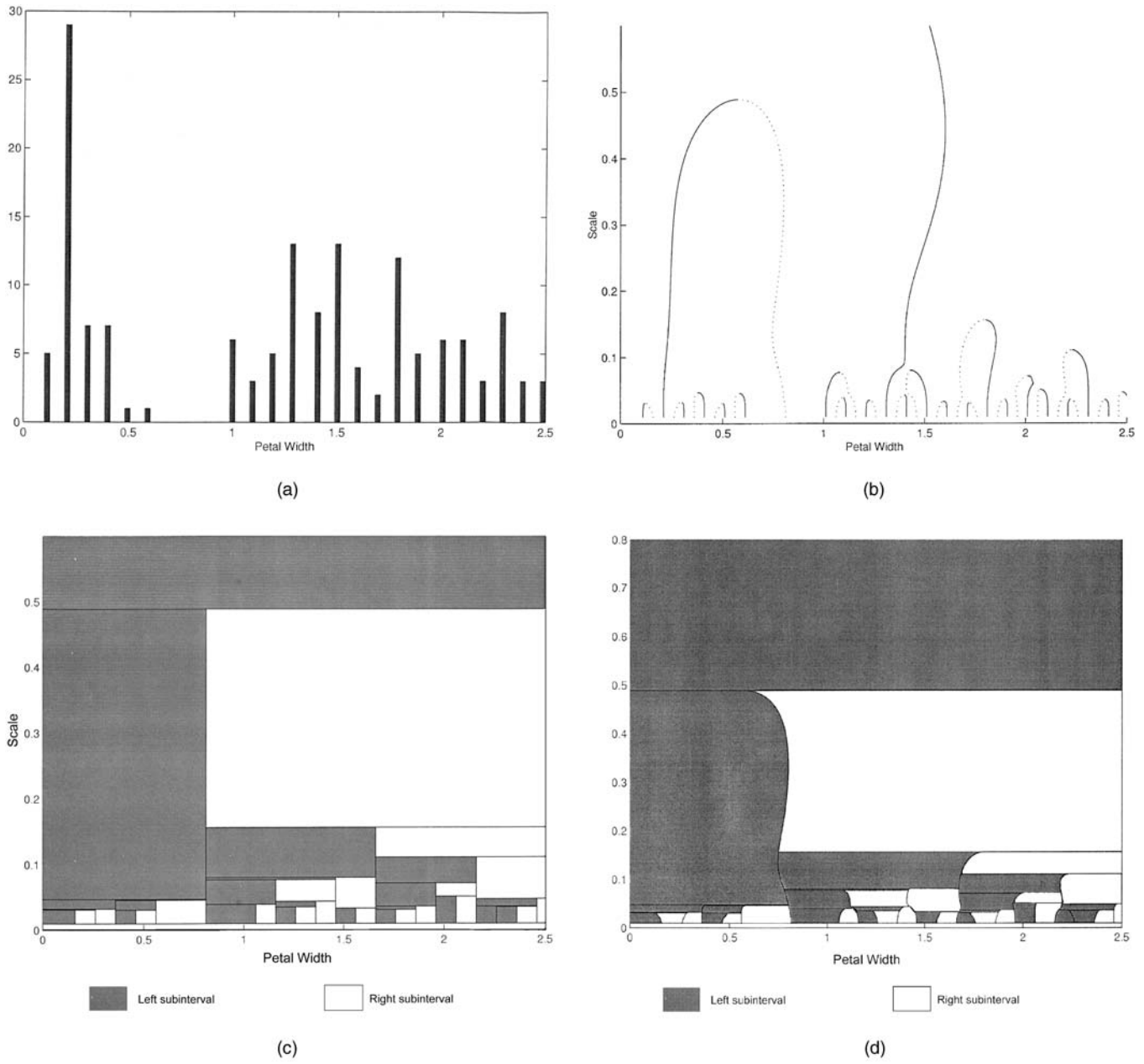


Fig. 4. (a) Histogram of Fisher's Iris petal width data. (b) Zero-crossings of dp/dx with p being the scale space image generated from Fisher's Iris petal width data. The dotted lines are the minimal curves and the solid lines are the maximal curves, i.e., blob centers at different scales. (c) Nested hierarchical clustering interval tree for Fisher's Iris petal width data. (d) Nonnested hierarchical clustering interval tree for Fisher's Iris petal width data.

purpose of the proposed clustering interval tree is to partition the data space according to the probability distribution. We use the minimal curve of distribution $P(x, \sigma)$ to construct the binary clustering tree; Witkin, on the other hand, uses the contour of extrema in the signal to construct the ternary interval tree.

As an illustration, we construct a clustering interval tree for the petal width data coming from the well-known Fisher Iris data. Fig. 4a is the histogram of this data set, and Fig. 4b is the plot of zero-crossings of dp/dx. The dotted curves in Fig. 4b are the minimal curves and the solid curves are the maximal curves. Fig. 4c is the clustering interval tree of Fisher's Iris pedal width data. Fig. 4d is a plot of the nonnested hierarchical clustering interval tree for Fisher's Iris petal width data.

The original Fisher Iris data contains measurements of three species of iris with four features (petal length, petal width, sepal length, and sepal width) in each pattern. From Fig. 4c, we can see that, at $\sigma = 0.12$, three clusters are obtained and the resulting clustering only commits six mistakes, but the standard clustering algorithms usually commit 16 to 17 mistakes [15]. A perfect classification can be obtained by the scale-space clustering algorithms (see [15]) if four features are considered and one rescaling scheme is used. The algorithms using Iris data as a benchmark example can be found in [38], [39].

7.2 Interpretation of Why Pitchfork Merging Is Seldom Observed

In hierarchical clustering based on scale-space filtering, there are two possible types of mergings as σ increases: 1) pitchfork merging and 2) saddle-node merging. In a pitchfork merging, two cluster centers smoothly merge into one supercluster center, while, in a saddle-node merging, a cluster center suddenly disappears and is siphoned into another cluster center. It has been observed that saddle-node merging is most frequent, but so far no theoretical result has been proposed to interpret this phenomenon [14], [15]. From Theorem 1, we know that, for almost all data sets, 0 is a regular value of $\bigtriangledown_x P(x,\sigma)$. This implies that, for almost all data sets, we can only observe saddle-node merging since a pitchfork merging means that 0 is not a regular value of $\bigtriangledown_x P(x,\sigma)$. Therefore, Theorem 1 provides a theoretical interpretation of this observation.

7.3 Decrease of the Number of Clusters

From the theoretical point of view, in order to guarantee that we can obtain a meaningful hierarchy, we should require that the number of cluster centers (i.e., the maxima of $P(x,\sigma)$), $\pi(\sigma)$, decreases as σ increases. Roberts [11] has recently attempted to prove that $\pi(\sigma_1) \leq \pi(\sigma_2)$ for all $\sigma_2 < \sigma_1$ in any dimensions. However, one simple example has been given in [36], also see [37], which shows that this does not hold, even in the two-dimensional case. Based on Theorem 1, we can prove that, in the one-dimensional case and for almost all data sets, $\pi(\sigma)$ decreases as σ increases. Nevertheless, we cannot extend our proof to higher dimensions.

It should be noted that such a problem does not exist in the nested hierarchical clustering algorithm.

7.4 The Increasing Schedule of the Scale Sequence Used in [14], [15]

In both algorithms proposed in [14] and [15], the scale sequence σ_i is given by

$$\sigma_{i+1} = h\sigma_i, \quad i = 1, 2, \cdots, \tag{44}$$

where $h > 1$ is a constant. However, no interpretation is provided to explain why such a schedule should be adopted. In scale-space theory, this can be explained as the requirement of accuracy and stability of the representation, as proven in [24]. It is because (44) is equivalent to

$$\frac{\sigma_{i+1} - \sigma_i}{\sigma_i} = h - 1 \tag{45}$$

and this corresponds to the "natural scale" in the resolution axis [24]. Based on Weber's law, we can suggest a low-bound for h in practical applications, as detailed in Section 3.5.

8 Remarks and Conclusion

We have proposed in this paper a new approach to data clustering based on scale space theory. By mimicking how human eyes unravel intrinsic structures in images, clustering by scale-space filtering performs clustering through a blurring process which treats a data set as an image with each datum being a light point attached with a uniform luminous flux. Blobs (clusters) form throughout the blurring process, with smaller ones merging to larger ones along the merging scale until the whole image contains only one light blob (cluster) at a low enough resolution. A hierarchical tree of clustering which gives a family of realistic data clustering is thus obtained. The approach advances a method of clustering with a psychophysiological basis and interpretation. It also provides a way to solve the vexing problem of cluster validity checks and establishes a unifying framework for other scale-based algorithms.

From the proposed approach, we can get a family of clustering algorithms by employing different numerical difference methods to solve the gradient differential equation in (8) or (10). We can also use other high order optimization methods, such as conjugate gradient and quasi-Newton-like methods, to construct new clustering algorithms. If the data consist of long and thin clusters, we can make use of Mahalanobis distance instead of Euclidean distance in the algorithms and the covariance matrices can be estimated iteratively with a particular regulation technique if too few data is contained in a given cluster. In this paper, several illustrative examples and a test on remote sensing images have been given with convincing results. We have also tested different implementations of scale-space clustering on the generated data sets and the remotely sensed images and the results will be reported in another paper. These results in brief show that:

1. Lifetime is a suitable cluster-validity criterion. This can also be observed in Fig. 2.
2. The algorithms are robust to the variation of cluster shapes, it can even be non-Gaussian. This is mainly because the objective function in (7) is the density distribution estimate and the algorithm is a "mode-seeking" one which tries to find the dense regions. This phenomenon can also be seen in Fig. 1, where data are of different shapes.
3. The algorithms are insensitive to outliers because outliers can easily be detected in these algorithms. From (7) and (8), we can see that the influence of one point on a given cluster center is proportional to $O(de^{-\frac{d^2}{\sigma^2}})$, with d being the distance between them. When d is large, $O(de^{-\frac{d^2}{\sigma^2}})$ is very small. An outlier is usually very far from the cluster centers, so it has little influence on the estimation of the cluster center. On the other hand, the normal data points are usually far away from the outlier, so they have little influence on an outlier. That is to say, an outlier can survive for a long time as a cluster, therefore, it has large outlierness (see (36)) and can easily be detected. This technique has been successfully used to process data related to Fig. 3. Since outlier detection is theoretically predictable, we then need not to provide an empirical study for illustration.
4. Since the proposed algorithm allows cluster in a partition to be obtained at different scales, more subtle clustering, such as the classification of land covers, can be obtained.

5. The algorithms work equally well in small and large data sets with low and high dimensions.

Some main concepts and ideas discussed in this paper are not entirely new. For example, the scale-space theory is developed in image and signal processing [22], [23], [24], [25], [26], [27], [28]. The "mode-seeking" or "peak-seeing" idea has been widely used to define clusters and construct clustering algorithms in pattern recognition and image processing for quite a long time and it has been used in scale-related clustering in [11], [15]. The stability of $\pi(\sigma)$ as a measure of clustering validity is also used in [11], [12], [13], [14], [15], the lifetime for a cluster in a nested hierarchical clustering is suggested in [14] in the terminology of "robustness of a good cluster," and the lifetime of a partition at a given scale in a nested hierarchical clustering is introduced in [12]. Furthermore, the logarithmic discretization scheme for the scale parameter is also used in [12], [14], and [15].

The main contribution of this paper is that we bring all these ideas together into a unified whole, provide a thorough consolidation of such related works, and then formulate a generalized framework for scale-space clustering algorithms by showing how this extends to or differs from those in [11], [12], [13], [14], [15]. We have derived the algorithms directly from scale-space filtering theory and this allows us to use the theory developed for scale-space filtering to analyze scale-space clustering algorithms and explain the clustering process and results from the psychophysiological perspective. Based on this point of view, we manage to explain: 1) Why scale parameter should be increased by a constant factor in practice. 2) Why lifetime should be measured on a logarithmic scale. 3) Why pitchfork merging are seldom. The proposed clustering method can also be applied to the classification of data with known distribution containing noise or being indifferentiable.

For further research, mechanism should be devised to separate clusters which are close to each other. Furthermore, since Gaussian scale space theory is designed to be totally noncommittal, then it cannot take into account any a priori information on structures which are worthy of preserving. Such a deficiency may be improved by employing more sophisticated nonlinear scale space filters.

It should also be noted that scale-space filtering theory is concerned with the simple conscious experience associated with a stimulus (low-level processing), i.e., the first eye contact between an organism and the environment. Therefore, it is a theoretical model of front-end visual system and is not suitable for processing directly clustering problems related to high-level perception, such as clusters in an image with texture background or clusters with meaningful shapes. If we want to solve these problems in a biological setting, high-level theoretical and computational perception models might be employed. This does not mean that scale-space clustering algorithms cannot be applied to these problem indirectly. They may be quite useful to cluster feature vectors retrieved from images, as shown in [11], [13].

ACKNOWLEDGMENTS

The authors would like to thank the referees and editors for their valuable comments and suggestions and Dr. J.C. Luo for his assistance in performing the remote sensing experiment. This work was supported by Grant CUHK4136/99H of the Hong Kong Research Grants Council.

REFERENCES

[1] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley-Interscience, 1974.
[2] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs: N.J., Prentice Hall, 1988.
[3] M. Blatt, S. Wiseman, and E. Domany, "Data Clustering Using a Model Granular Magnet," *Neural Computation,* vol. 9, pp. 1,805-1,847, 1997.
[4] R. Dubes and A.K. Jain, "Clustering Techniques: The User's Dilemma," *Pattern Recognition,* vol. 8, pp. 247-260, 1976.
[5] L. Hubert, "Approximate Evaluation Technique for the Single-Link and Complete-Link Hierarchical Clustering Procedure," *J. Am. Statistical Assoc.,* vol. 69, p. 968, 1974.
[6] H.P. Friedman and J. Robin, "On Some Invariant Criteria for Grouping Data," *J. Am. Statistical Assoc.,* vol. 62, p. 1,159 1967.
[7] S.C. Johnson, "Hierarchical Clustering Scheme," *Psychometrika,* vol. 32, p. 241, 1967.
[8] C.T. Zahn, "Graphic-Theoretic Methods for Detecting and Describing Gestalt Clusters," *IEEE Trans. Computers,* vol. 20, pp. 68-86, 1971.
[9] D. Miller and K. Rose, "Hierarchical, Unsupervised Learning with Growing via Phase Transitions," *Neural Computation,* vol. 8, pp. 425-450, 1996.
[10] J. Waldemark, "An Automated Procedure for Cluster Analysis of Multivariate Satellite Data," *Int'l J. Neural Systems,* vol. 8, no. 1, pp. 3-15, 1997.
[11] S.J. Roberts, "Parametric and Nonparametric Unsupervised Clustering Analysis," *Pattern Recognition,* vol. 30, no. 2, pp. 261-272, 1997.
[12] P. Taven, H. Grubmuller, and H. Huhnel, "Self-Organization of Associative Memory and Pattern Classification: Recurrent Signal Processing on Topological Feature Maps," *Biological Cybernetics,* vol. 64, pp. 95-105, 1990.
[13] R. Wilson and M. Spann, "A New Approach to Clustering," *Pattern Recognition,* vol. 23, no. 12, pp. 1,413-1,425, 1990.
[14] Y.F. Wong, "Clustering Data by Melting," *Neural Computation,* vol. 5, no. 1, pp. 89-104, 1993.
[15] S.V. Chakravarthy and J. Ghosh, "Scale-Based Clustering Using the Radial Basis Function Network," *IEEE Trans. Neural Networks,* vol. 7, no. 5, pp. 1,250-1,261, 1996.
[16] M.R. Anderberg, *Cluster Analysis for Applications.* New York: Academic Press, 1973.
[17] G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," *Behavioral Science,* vol. 12, pp. 153-155, 1976.
[18] J.C. Bezdek, "A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 2, pp. 1-8, 1980.
[19] S. Kirpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, pp. 671-680, 1983.
[20] K. Rose, E. Gurewitz, and G. Fox, "A Deterministic Annealing Approach to Clustering," *Pattern Recognition Letters,* vol. 11, pp. 589-594, 1990.
[21] G. Celeux and G. Govaert, "A Classification EM Algorithm for Clustering and Two Stochastic Versions," *Computational Statistics and Data Analysis,* vol. 14, pp. 315-332, 1992.
[22] A.P. Witkin, "Scale Space Filtering," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 1,019-1,022, 1983.
[23] A.P. Witkin, "Scale Space Filtering: A New Approach to Multi-Scale Description," *Image Understanding,* S. Ullman and W. Richards, eds., Norwood, N.J.: Ablex, 1984.
[24] J.J. Koenderink, "The Structure of Images," *Biological Cybernetics,* vol. 50, pp. 363-370, 1984.
[25] J. Babaud, A. Witkin, M. Baudin, and R. Duda, "Uniqueness of the Gaussian Kernel for Scale-Space Filtering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 8, pp. 26-33, 1986.
[26] A.L. Yuille and T. Poggio, "Scaling Theorem for Zero Crossings," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 8, pp. 15-25, 1986.
[27] D. Marr, *Vision, A Computational Investigation into the Human Representation.* San Francisco: W.H. Freeman, 1982.

[28] R. Hummel and R. Moniot, "Reconstructions from Zero Crossings in Scale Space," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 37, no. 12, pp. 245-295, 1989.
[29] D.H. Hubel, *Eye, Brain, and Vision.* New York: Scientific Am. Library, 1995.
[30] S. Coren, L.M. Ward, and J.T. Enns, *Sensation and Perception.* Harcourt Brace College Publishers, 1994.
[31] B. Everitt, *Cluster Analysis.* New York: Wiley, 1974.
[32] E.L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction.* New York: Springer 1990.
[33] F. Mulier and V. Cherkassky, "Self-Organization as an Iterative Kernel Smoothing Process," *Neural Computation,* vol. 7, pp. 1,165-1,177, 1995.
[34] E.A. Nadaraya, "On Estimating Regression," *Theory Probability Application,* vol. 74, pp. 743-750, 1964.
[35] G.S. Watson, "Smooth Regression Analysis," *Sankhya,* series A, vol. 26, pp. 359-372, 1964.
[36] T. Linderberg, "Scale-Space for Discrete Signals," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, pp. 234-254, 1990.
[37] L.M. Lifshitz and S.M. Pizer, "A Multiresolution Hierarchical Approach to Image Segmentation Based on Intensity Extrema," internal report, Dept. of Computer Science and Radiology, Univ. of North Carolina, Chapel Hill, 1987.
[38] S.J. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian Approaches to Gaussian Mixture Modeling," *IEEE Trans. Pattern Analyis and Machine Intelligence,* vol. 20, no. 11, pp. 1,133-1,142, Nov. 1998.
[39] I. Gath and B. Geva, "Unsupervised Optimal Fuzzy Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 773-781, July 1989.

Yee Leung received the BSSc degree in geography from The Chinese University of Hong Kong in 1972, the MA and PhD degrees in geography and the MS degree in engineering from The University of Cororado, in 1974, 1977, and 1977, respectively. He is currently a professor of geography and chairman of the Department of Geography, research fellow of the Center for Environmental Policy and Resource Management, and deputy academic director of the Joint Laboratory for Geoinformation Science at The Chinese University of Hong Kong. He has published four monographs and more than 100 articles in international journals and book chapters. His areas of specialization cover the development and application of intelligent spatial decision support systems, spatial optimization, fuzzy sets and logic, neural networks, and evolutionary computation. Dr. Leung serves on the editorial boards of several international journals and is a council member of several Chinese professional organizations.

Jiang-She Zhang received the BS, MS, and PhD degrees in applied mathematics, all from Xi'an Jiaotong University in 1984, 1987, and 1993, respectively. From February 1995 to August 1995 and from May 1997 to September 1998, he worked as a research associate at The Center for Enviromental Studies, The Chinese University of Hong Kong. He is now an associate professor at The Institute for Information and System Sciences and The Research Center for Applied Mathematics, Xi'an Jiaotong University, China. He has published more than 15 academic papers on optimization algorithms, genetic algorithms, and computational geometry. His current research interests include global optimization, neural networks, evolutionary computation, and computational geometry.

Zong-Ben Xu received the MS degree in mathematics in 1981 and the PhD degree in applied mathematics in 1987 from Xi'an Jiaotong University, China. In 1988, he was a postdoctoral researcher in the Department of Mathematics, The University of Strathclyde, United Kingdom. He worked as a research fellow in the Information Engineering Department from February 1992 to March 1994, The Center for Environmental Studies from April 1995 to August 1995, and The Mechanical Engineering and Automation Department from September 1996 to October 1996, at The Chinese University of Hong Kong. From January 1995 to April 1995, He was a research fellow in The Department of Computing in The Hong Kong Polytechnic University. He has been with The Faculty of Science and Research Center for Applied Mathematics at Jiaotong University since 1982, where he was promoted to associate professor in 1987 and full professor in 1991, and now serves as an authorized PhD supervisor in mathematics and director of the Institute for Information and System Sciences. He has published two monographs and more than 70 academic papers on nonlinear functional analysis, numerical analysis, optimization techniques, neural networks, and genetic algorithms, most of which are in international journals. His current research interests include neural networks, evolutionary computation, and multiple objective decision making theory. Dr. Xu holds the title "Owner of Chinese PhD Degree Having Outstanding Achievements" awarded by the Chinese State Education Commission and the Academic Degree Commission of the Chinese Council in 1991. He is a member of the New York Academy of Sciences and International Mathematicians Union (IMU).