Exploiting Data Topology in Visualization and Clustering of Self-Organizing Maps

Kadim Taşdemir and Erzsébet Merényi, *Senior Member, IEEE*

***Abstract*—The self-organizing map (SOM) is a powerful method for visualization, cluster extraction, and data mining. It has been used successfully for data of high dimensionality and complexity where traditional methods may often be insufficient. In order to analyze data structure and capture cluster boundaries from the SOM, one common approach is to represent the SOM's knowledge by visualization methods. Different aspects of the information learned by the SOM are presented by existing methods, but data topology, which is present in the SOM's knowledge, is greatly underutilized. We show in this paper that data topology can be integrated into the visualization of the SOM and thereby provide a more elaborate view of the cluster structure than existing schemes. We achieve this by introducing a weighted Delaunay triangulation (a connectivity matrix) and draping it over the SOM. This new visualization, CONNvis, also shows both forward and backward topology violations along with the severity of forward ones, which indicate the quality of the SOM learning and the data complexity. CONNvis greatly assists in detailed identification of cluster boundaries. We demonstrate the capabilities on synthetic data sets and on a real 8-D remote sensing spectral image.**

***Index Terms*—Clustering, data mining, self-organizing map (SOM), topology preservation, visualization.**

I. Introduction

THE self-organizing map (SOM) [1] is a widely and effectively used neural paradigm for clustering and data mining of high-dimensional data due to its several advantageous properties such as topology preserving mapping and learning of the data distribution. By preserving the neighborhood relations on a rigid lattice, the SOM facilitates the visualization of the structure of a higher dimensional data space in lower (usually one or two) dimensions.

Informative representation of the SOM's knowledge can significantly assist accurate capture of cluster boundaries. Similarities of prototypes adjacent in the SOM, or the size of the receptive fields of neural units, are often used in various ways in existing visualization schemes (discussed in Section II). With this paper, we want to call attention to the power of a greatly underutilized component of the SOM's knowledge: data topology. We will show that the inclusion of data topology in the SOM visualization provides more sophisticated clues to cluster structure than existing SOM visualization approaches. This inclusion is especially important for high-dimensional, large, and intricate data sets with many meaningful clusters, which have interesting rare clusters to be discovered [2], [3].

A limited use of data topology in SOM representation is proposed by Pölzbauer *et al.* [4] to indicate topology violations and data distribution. In [4], lines are drawn between the prototypes on the SOM grid for those data vectors that are neighbors in data space according to some metric, but are mapped to different SOM prototypes. A large number of lines and their concentrations in the SOM show dense regions while the lengths of the lines express the range of topology violations. By using the neighborhood of the data vectors to determine topology violations (contrary to the common approach of using the neighborhood of the SOM prototypes), Venna and Kaski [5] construct two measures: "trustworthiness" and "neighborhood preservation" of the SOM. The visualization of Pölzbauer *et al.* [4] shows the set of violations that forms the measure of "neighborhood preservation" defined in [5]. The approach taken by [4] works well for estimating data distribution and finding topology violations when prototypes outnumber the data vectors. In contrast, when there are plenty of data, neighboring data vectors that are mapped to different prototypes are only the ones at the boundaries of the Voronoi polyhedra of the prototypes, in which case the method in [4] ignores a lot of helpful mapping information and estimates data distribution inadequately. That makes it a partial solution for the visualization of data topology on a SOM.

More generally, for a given set of data vectors and their corresponding prototypes obtained by any quantization process (including the SOM), a complementary approach for exploiting data topology is to use a graph model in the data space. One way to construct a topology representing graph is to use the induced Delaunay triangulation proposed by Martinetz and Schulten [6]. Several learning algorithms such as topology representing networks [6], Growing Neural Gas [7], and grow-when-required [8] use induced Delaunay triangulation in combination with parameters that depend on the occurrences of data samples for a better topographic mapping than with the Kohonen SOM. The induced Delaunay triangulation is binary: it reflects the adjacency relations of the quantization prototypes in data space, but it does not convey data distribution. Therefore, it may not be sufficient for detailed cluster analysis, especially in case of slightly overlapping clusters or noisy data. A more informative approach is to construct the graph by using statistical learning theory as proposed by Aupetit [9]. This method considers the statistics of

Manuscript received July 25, 2007; revised May 29, 2008; accepted August 24, 2008. First published February 18, 2009; current version published April 03, 2009. This work was supported in part by the Applied Information Systems Research Program of NASA, Science Mission Directorate, under Grant NNG05GA94G.

K. Taşdemir was with the Electrical and Computer Engineering Department, Rice University, Houston, TX 77005 USA. He is now with the Computer Engineering Department, Yasar University, Bornova, Izmir 35100, Turkey (e-mail: kadim.tasdemir@yasar.edu.tr).

E. Merényi is with the Electrical and Computer Engineering Department, Rice University, Houston, TX 77005 USA (e-mail: erzsebet@rice.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2008.2005409

the data distribution within the Voronoi polyhedra of the prototypes, which makes it useful for estimating data topology and robust to noise. However, its use for visualization is limited to low-dimensional, 1-D to 3-D, applications since it shows neighborhood relations in the data space.

The objective of this paper is to integrate the data topology, present in the SOM's knowledge, into the visualization of the SOM for improved capture of clusters. This objective will be accomplished through a new concept of the "connectivity matrix" and its specific rendering over the SOM. The term "connectivity matrix" exists in the literature, for pairwise distances of data points. Here we use it for quantization prototypes with a novel definition of connectivity. We define the connectivity matrix as a weighted version of the induced Delaunay triangulation where the weights of the edges signify the data distribution. The method we present is not limited by data dimensionality because the neighborhood relations in the data space are shown on the SOM grid. This new visualization also shows both forward and backward topology violations as a byproduct due to rendering data topology on the SOM.

Section II briefly reviews the SOM algorithm and discusses previous visualization schemes for the SOM. Section III introduces the "connectivity matrix," its visualization, and its use for assessing topology violations. Section IV gives a step-by-step procedure for the extraction of cluster boundaries from the SOM through the visualization of the connectivity matrix. It also presents a clustering example on a real 8-D data set. Section V discusses the advantages of this scheme, open and unresolved issues, and possible follow-up improvements.

II. Previous Work on Visualization of SOM Knowledge

The SOM is an unsupervised neural learning algorithm that maps a data manifold $\mathcal{M} \subset \mathbb{R}^d$ to a (lower dimensional) fixed lattice $\mathcal{G}$ of N neural units. Each neural unit i has a weight vector w_i assigned to it, which is adapted through a learning process as originally defined by Kohonen [1]. The process is based on finding the best matching unit w_i for a given data vector $v \in \mathcal{M}$, such that

$$\|v - w_i\| \leq \|v - w_j\| \qquad \forall j \in \mathcal{G} \tag{1}$$

and updating w_i and its neighbors according to

$$w_j(t+1) = w_j(t) + \alpha(t) h_{i,j}(t)(v - w_j(t)) \tag{2}$$

where t is time, $\alpha(t)$ is a learning parameter, and $h_{i,j}(t)$ is the neighborhood function, often defined by a Gaussian kernel around the best matching unit w_i. After the learning process, the weight vectors become the vector quantization prototypes of the input space $\mathcal{M}$. From now on, we will use the term "prototype" for SOM weight vectors. Ideally, the SOM is a topology preserving mapping, i.e., the prototypes that are neighbors in $\mathcal{G}$ are also neighbors (centroids of neighboring Voronoi polyhedra) in $\mathcal{M}$ and vice versa.

There is a variety of existing schemes for the representation of the SOM's knowledge including visualization of the (Euclidean) distances between prototypes that are immediate neighbors in $\mathcal{G}$. The most commonly used method, the U-matrix [10] and its variants (e.g., [11] and [12]) signify these distances by using proportional intensities of gray shades on grid cells. These work well for small data sets with a low number of clusters mapped to a relatively large SOM grid but, because of averaging of prototype distances over neighboring SOM grid cells, or thresholding, they tend to miss finer structure in complicated and large data sets [13]. Another method is the adaptation of the size or the shape of the grid cells according to the distances between neighboring prototypes [14], [15], which can help manual cluster extraction for simple data sets. The use of automated color assignments aims at qualitative exploration of the approximate cluster structure [3], [16]–[18]. Examination of individual component planes of the SOM is helpful in discovering information specific to the corresponding component, which may be hidden when all planes are examined together [17], [19].

Many researchers convey SOM knowledge through visualizing the receptive field sizes of prototypes (data histograms) by drawing vertical bars, curves, gray shades, etc. (e.g., [12], [17], and [19]). Pampalk *et al.* [20] propose smoothing data histograms by assigning a weighted membership of data vectors to the prototypes in order to get a precise visualization of density distribution. However, expression of the SOM's knowledge solely with data histograms conceals finer structure in complicated data. Approaches employing data histograms and distances between prototypes together in the same visualization, such as in [3] and [14], do not overcome the drawbacks of each individual method, which are discussed above.

In order to visualize the cluster structure during the training of the SOM, adaptive coordinates [21] and the double SOM [22] update not only the prototypes but also their positions in the SOM lattice. These methods expose the dissimilarities between the prototypes by the lattice distance of the prototypes, which in turn produces a visual separation of clusters. However, it is unclear how they would work for large data volumes. For the double SOM, finding the appropriate parameters for robust learning is difficult. Ressom *et al.* proposed an improved technique for the double SOM whereby the use of adaptive parameters produces more robust learning than the double SOM [23]. This technique worked demonstratively well for a data set of gene expression profiles consisting of a small number of vectors.

An innovative proposal to find structures in high-dimensional manifolds is a growing SOM [24], but it appears less robust than the Kohonen SOM because of the large number of parameters needing adjustment. Its performance for large data volumes is also undemonstrated. Another variant of the SOM that enables a direct and visually appealing measure of interpoint distances on the map is the visualization-induced SOM (ViSOM) [25]. The ViSOM produces a smooth and evenly graded mesh through the data points that reveals the discontinuities in the manifold. However, it requires a relatively large number of prototypes even for small data sets.

III. Topology Visualization Through Connectivity Matrix of SOM Prototypes

A. Induced Delaunay Triangulation and Connectivity Matrix

In order to faithfully characterize a data manifold $\mathcal{M}$ that can possibly be discontinuous or folded, Martinetz and Schulten [6]

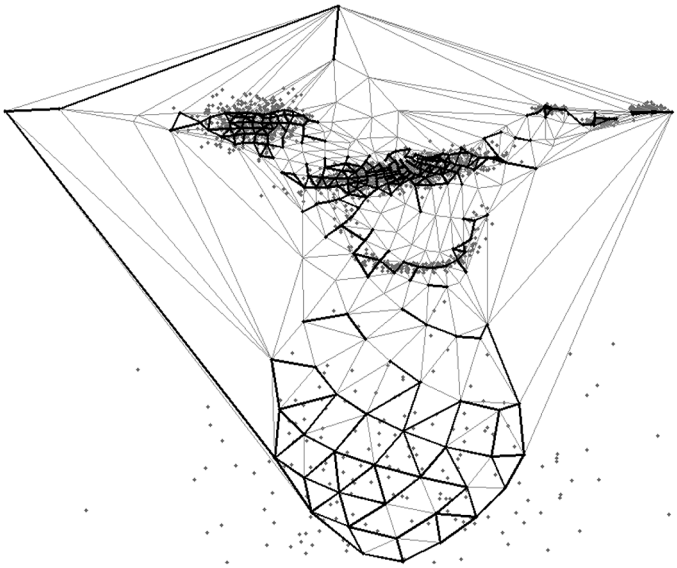


Fig. 1. Comparison of Delaunay triangulation (thin lines) and induced Delaunay triangulation (thick lines) for the 2-D Clown data using the SOM prototypes created by [26]. The "Clown" is indicated by the annotations in Fig. 2. The data manifold is shown by the dots. The induced Delaunay triangulation exposes the discontinuities in the data manifold, for example, the separations between the eyes, the nose, and the mouth, while the Delaunay triangulation does not highlight them.

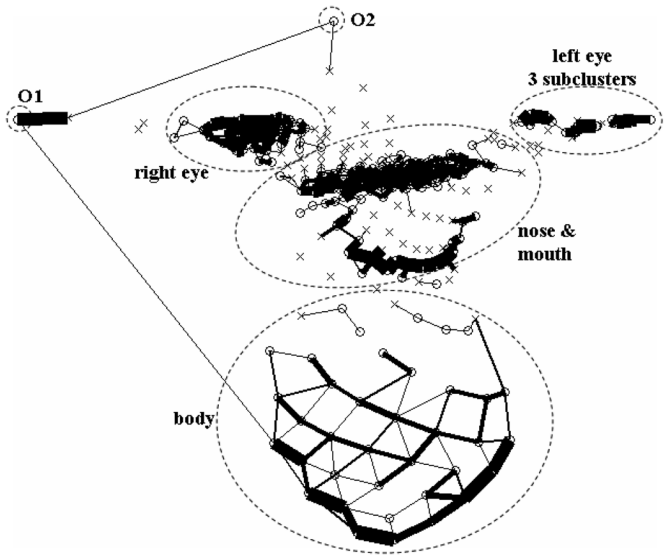


Fig. 2. Connectivity matrix (weighted induced Delaunay triangulation) for the same data and SOM prototypes as in Fig. 1. The prototypes with nonempty receptive fields are labeled by circles while interpolating prototypes are shown by "x." The width of the line connecting two prototypes w_i and w_j is proportional to the connectivity strength $\mathrm{CONN}(i,j)$, which is the number of data vectors for which one of w_i and w_j is the BMU and the other is the second BMU. This weighting makes the data distribution visible. Low-density regions, for example, the line connecting the outlier O2 and the interpolating prototype near O1, at the right end of the row of nonempty prototypes, and the lines at the cluster boundaries, are exposed by thin (or no) lines.

introduce the notion of induced Voronoi tessellation $\tilde{V}$ and its dual, the induced Delaunay triangulation $\tilde{D}$. These are the intersections of the regular Voronoi polyhedra V (or Delaunay triangulation) with $\mathcal{M}$. For prototype w_i, the Voronoi polyhedron V_i and its induced counterpart $\tilde{V}_i$ are

$$\begin{aligned} V_i &= \left\{v \in \mathbb{R}^d : \|v - w_i\| \leq \|v - w_k\| \forall k \in \mathcal{G}\right\} \\ \tilde{V}_i &= \left\{v \in \mathcal{M} : \|v - w_i\| \leq \|v - w_k\| \forall k \in \mathcal{G}\right\}. \end{aligned} \quad (3)$$

According to the definition in [6], w_i and w_j are adjacent in $\mathcal{M}$ if and only if their receptive fields RF_i and RF_j (their masked Voronoi polyhedra $\tilde{V}_i$ and $\tilde{V}_j$) are adjacent. An example of the Delaunay triangulation and its induced version is shown in Fig. 1 for a 2-D data set called "Clown" created by Vesanto and Alhoniemi [26]. The SOM prototypes were also computed and graciously provided to us along with the Clown data by these authors. This 2-D data set has several clusters with different shapes and sizes (eyes, nose, mouth, and body) and outliers. The induced Delaunay triangulation $\tilde{D}$ makes the disconnected parts of the manifold (such as the eyes, nose, and mouth) obvious, whereas the regular Delaunay triangulation D does not delineate the same separations.

As proposed by Martinetz and Schulten, the induced Delaunay triangulation can be determined from the relationships of the best matching units (BMUs) and the second BMUs, expressed in a so-called adjacency matrix A, provided that the SOM prototypes are "dense enough" in $\mathcal{M}$ [6]. Following that, we can build the matrix A, for a converged state, by sequentially presenting data vectors $v \in \mathcal{M}$ and each time setting $A(i,j)$ and $A(j,i)$ to 1 when one of w_i and w_j is the BMU and the other is the second BMU to v. A (the equivalent of $\tilde{D}$ under the above conditions) delineates the nonlinearities and the submanifolds in $\mathcal{M}$. However, A is a binary matrix that does not reflect the data distribution within the receptive fields, and two prototypes w_i and w_j are shown adjacent by A even if $A(i,j)$ was set to 1 by only one data vector. Thus, noise can easily obscure discontinuities in the manifold by showing otherwise obviously disjunct clusters as connected. An example in Fig. 1 is the uniform connectedness of the body or the mouth of the Clown by the thin lines representing the regular Delaunay triangulation. For a better characterization of the data topology and for discrimination of noisy connections from dense regions, we define a connectivity strength matrix, which we denote by CONN, and which is a weighted analog of A, where the weights indicate the density distribution of the input data among the prototypes adjacent in $\mathcal{M}$.

Let CONN be an $N \times N$ matrix where N is the number of neural units in $\mathcal{G}$. The connectivity strength $\mathrm{CONN}(i,j)$ between units i and j is the number of data vectors $v \in \mathcal{M}$ for which w_i or w_j is the BMU and the other is the second BMU

$$\mathrm{CONN}(i,j) = |\mathrm{RF}_{ij}| + |\mathrm{RF}_{ji}| \quad (4)$$

where RF_{ij} is the part of the receptive field of w_i where w_j is the second BMU, and $|\mathrm{RF}_{ij}|$ is the number of data vectors in RF_{ij}. Obviously, $|\mathrm{RF}_i| = \sum_{j=1}^{N} |\mathrm{RF}_{ij}|$ because $\mathrm{RF}_i = \cup_{j=1}^{N} \mathrm{RF}_{ij}$. CONN thus shows how the data is distributed within the receptive fields with respect to neighbor prototypes. This provides a finer density distribution than other existing density representations, which show the distribution only on the receptive field level. We define the similarity of two prototypes w_i and w_j based on their connectivity strength $\mathrm{CONN}(i,j)$.

Fig. 2 shows CONN visualized in the data space for the case given in Fig. 1. Compared to Fig. 1, all connections remain,

but now the strength of each connection is signified by the line width, which is proportional to $\mathrm{CONN}(i,j)$. This makes poorly connected (low density) regions obvious, such as the connections for the outliers O1 and O2 (encircled prototypes in Fig. 2) and the thin (or missing) lines at the cluster boundaries. Clusters not obvious in Fig. 1 clearly emerge here.

B. CONNvis: Visualization of the Connectivity Matrix on the SOM

We visualize CONN on the SOM lattice by connecting the neural units i and j whose prototypes w_i and w_j are adjacent in $\mathcal{M}$. Lines of various widths and colors are used for $\mathrm{CONN}(i,j) > 0$ (Fig. 3). The line width is proportional to the strength of the connection and therefore reflects the density distribution among the connected units. It also shows the *global importance* of the connection since it displays the number of data vectors in $\mathrm{RF}_{ij} \cup \mathrm{RF}_{ji}$ relative to the number of all data vectors. The connectivity strengths of w_i indicate how often w_i and each of its neighbors in $\mathcal{M}$ are selected together (are BMU and second BMU pairs for data vectors). This shows the local data distribution among its neighbors. Hence, a ranking of the connectivity strengths of w_i reveals the most-to-least dense regions local to w_i in data space. We show the ranking of neighbors of w_i by line colors, red, blue, green, yellow, and dark to light gray levels, in descending order. (Alternatively, the ranking could be shown by using intensities of a single color.) The connections on the SOM are drawn in the order of lowest to highest ranking so in case of intersections the higher ranking connection will overlay the lower ranking one. Because the density ranking does not depend on the size of w_i's receptive field, but only on the relative contribution of each neighbor, line colors express the *local importance* of the connections. The line width and the line color together indicate a combined view of the *global* and *local* properties of the data distribution.

An example of CONN visualization (CONNvis) on the SOM is in Fig. 4 for the Clown data presented in Fig. 1. A detailed explanation for this example will be given in Section III-D. One important aspect we want to note here is that the ranking of the prototypes is not symmetric, i.e., if the rank of w_j for w_i is r, and the rank of w_i for w_j is s, r is not necessarily equal to s. The rank displayed by the color is the higher ranking one of r and s, regardless of the directionality of the connection. Therefore, a prototype may seem to have multiple connections of the same rank. For example, some prototypes at the bottom left of Fig. 4(a) have several red connections.

C. Assessment of Topology Preservation With CONNvis

Superimposing CONN on the SOM grid shows the neighborhood relations of the prototypes both in $\mathcal{M}$ and in $\mathcal{G}$ in the same visualization. Therefore, this new visualization also helps in a detailed assessment of topology preservation. For a perfectly topology preserving mapping, only the immediate SOM neighbors are expected to be connected. However, topology violations may occur, which will manifest in the CONNvis as:

- connected neural units that are not immediate neighbors in $\mathcal{G}$ (forward $(\mathcal{M} \rightarrow \mathcal{G})$ topology violations);
- unconnected neural units that are immediate neighbors in $\mathcal{G}$ (backward $(\mathcal{G} \rightarrow \mathcal{M})$ topology violations).

An example of an indication of forward topology violation is the green connection in Fig. 3: the prototype i has a neighbor in data space (the prototype at the end point of the green line) that is not mapped to an immediate lattice neighbor of i. A backward topology violation is shown by the lack of connection between i and its lattice neighbor to the right. As it is seen from this illustration, both forward and backward topology violations are identified through CONNvis. The visualization of backward topology violations reveals the discontinuities or submanifolds in the data that are obvious cluster boundaries. CONNvis also quantifies the extent of the forward violations. The strength (line width) of a forward topology violating connection characterizes the degree of the violation, which we will call *severity*. The more data vectors contribute to a given connection, the more severe is the violation. For a topology violating connection, low strength (thin lines) usually indicates outliers or noise while greater strengths are due to data complexity or badly formed SOM. The folding length of the violating connection, that is the maximum norm distance between the connected neural units in the SOM lattice, describes whether the topology violation is local (short ranged) or global (long ranged).

In most cases, perfect topology preservation is not necessary for cluster extraction. Weak global violations, or violations that remain within clusters, do not affect the delineation of boundaries. Proper investigation of such conditions for a trained SOM is therefore important. The connectivity matrix and its visualization, introduced above, is a useful tool for such analysis.

D. An Example of CONNvis for a 2-D Data Set

Fig. 4 demonstrates CONNvis for the 2-D Clown data discussed in Figs. 1 and 2. CONN is draped over the 19×17 hexagonal SOM lattice that was used in [26]. Thin dashed lines indicate the areas of the SOM where the different parts of the Clown are mapped. Since there is no dimensionality conflict between $\mathcal{M}$ and $\mathcal{G}$, few (if any) forward topology violations are expected, which is confirmed by the visualization. There are no thick lines connecting distant units. The thin red line of $length = 16$, connecting O2 and the prototype at the upper right corner of the SOM, and the thin blue vertical connection $(length = 10)$ from O1 to the body of the Clown at the right edge are examples of global violations. In this case, the weakness of these connections suggests that the prototypes O1 and O2 are outliers. Prototypes with empty receptive fields often do not have any connections, however, sometimes they may have connections because they can be second BMUs for some data vectors. An example is the prototype circled in the nose of the Clown close to the upper left corner in Fig. 4. Although it has an empty receptive field, it is the second BMU for two data vectors mapped to two adjacent prototypes. Some immediate lattice neighbors are not connected because the corresponding prototypes are not adjacent in $\mathcal{M}$. The resulting separations between neural units expose cluster

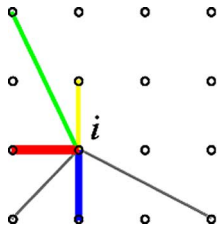


Fig. 3. Example of CONN visualization (CONNvis) on a SOM grid for the connections of a prototype (center node i). A line is drawn between two prototypes if they are adjacent in the data manifold $\mathcal{M}$ according to the induced Delaunay triangulation. The line width is proportional to the strength of the connection, $\mathrm{CONN}(i,j)$, which is the number of data vectors in $\mathrm{RF}_{ij} \cup \mathrm{RF}_{ji}$ (4). It shows the global importance of the connection since it states the number of data samples in $\mathrm{RF}_{ij} \cup \mathrm{RF}_{ji}$ relative to the total number of all data samples. The line colors encode a ranking of the immediate neighbors of this prototype in $\mathcal{M}$: the line to the neighbor with the strongest connection to i is colored red, and blue, green, yellow, and gray shades indicate the connections to the rest of the neighbors in decreasing order of rank. This ranking signifies the local importance of the connections as it displays the relative similarity of adjacent prototypes in data space.

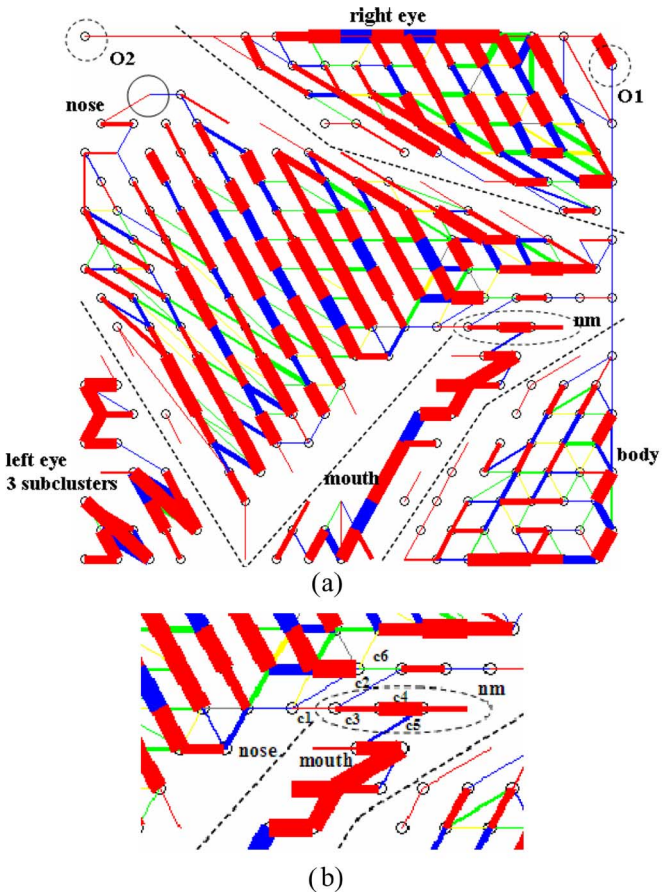


Fig. 4. (a) CONNvis of the 2-D Clown data (from Fig. 1) on the 19×17 hexagonal SOM lattice given in [26]. Prototypes with nonempty RF are shown by small circles. Line widths are proportional to connectivity strengths. The meaning of line widths and line colors is explained in Fig. 3. Dashed lines show major parts of the Clown. Some prototypes that are neighbors on the SOM grid are not connected, which indicates discontinuities in $\mathcal{M}$ (backward topology violations). Some clusters (mouth and body, left eye and nose, right eye and nose) are clearly separated. Others (O1 and O2, O1 and body, nose and nm, nm and mouth) are weakly connected (thin lines). The connections of O2 to the prototype at the top right corner, and O1 to the body are examples of global but weak topology violations. (b) The connections of the subcluster nm to the nose (c1, c2) and to the mouth (c5) are weak: $\mathrm{c1} = 1, \mathrm{c2} = 1$, and $\mathrm{c5} = 2$. In contrast, the connections within nm (c3, c4) are strong: $\mathrm{c3} = 3, = 8$. c2 and c5 exemplify weak local topology violations, which suggest that nm is a subcluster.

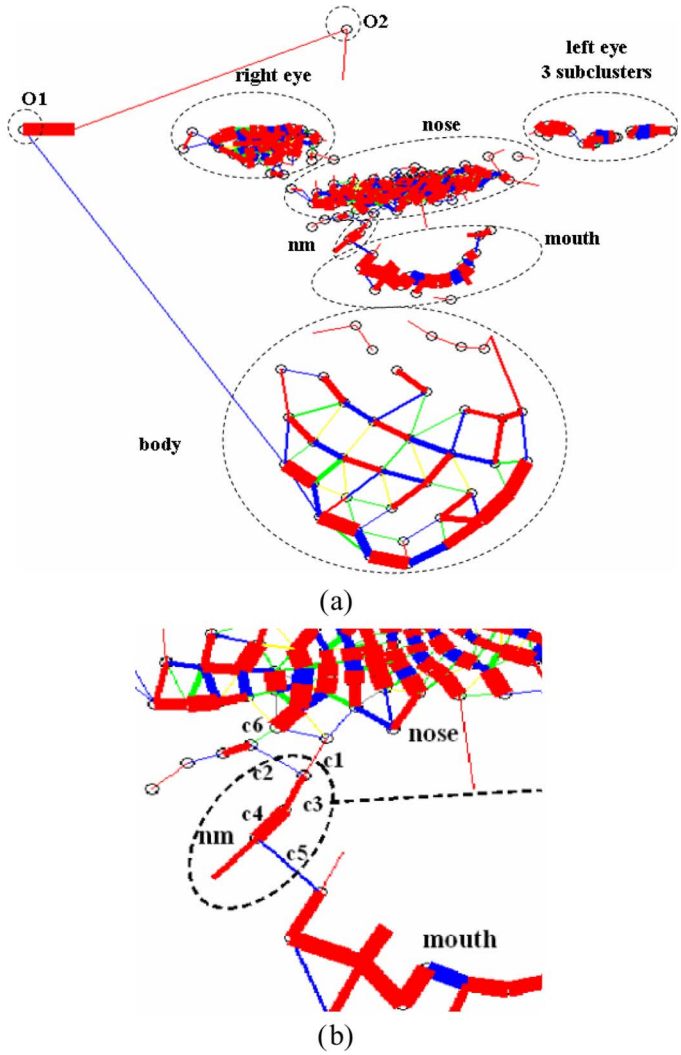


Fig. 5. (a) Visualization of CONN in the data space using the same scheme of line widths and colors as in Fig. 3. This verifies the separations seen in the SOM. (b) The cluster nm in data space, and its relations to the nose and the mouth clusters.

or submanifold structure in $\mathcal{M}$. For example, the separations between left eye and nose, right eye and nose, and mouth and body are obvious. The two global topology violations at the upper and right edges of the CONNvis make the prototypes between the end points seem connected and might obscure the discontinuities. For example, the connection that links O1 to the body of the Clown in Fig. 4 makes it look like O1; the right eye, the nose, and the body are all connected, even though the right eye and the nose are clearly separated from each other, and from the body when this connection is removed. If we first display only the nonviolating connections, then we can get an accurate view of the discontinuities. For this case, the discontinuities between the right eye, the body, and the outliers are clearly outlined by this view.

In Fig. 4(b), we focus on some instructive details. Cluster nm in the dashed oval is a subcluster connecting the nose and the mouth. It has weak and local topology violating connections between the mouth and the nose. The connections of nm to the nose (c1, c2) and to the mouth (c5) are weak compared to the connections within nm (c3, c4). Being both weak and violating, these connections suggest that nm is indeed a subcluster between the

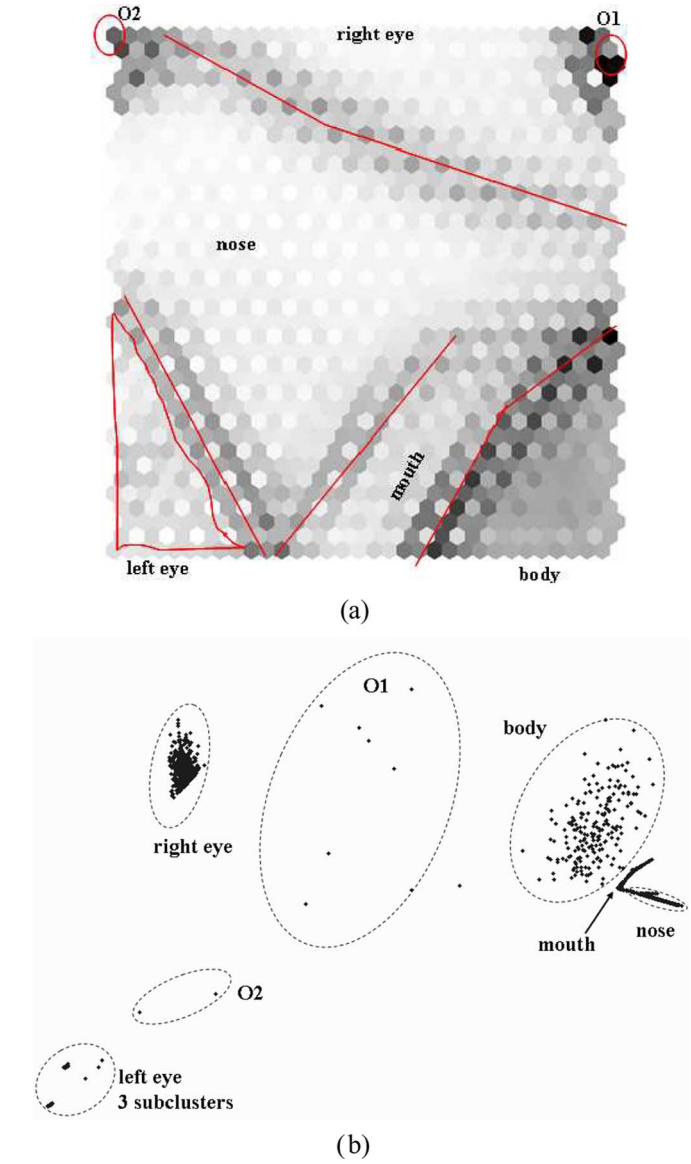


Fig. 6. (a) U-matrix visualization of the 2-D Clown data (from Fig. 1) on the 19×17 hexagonal SOM lattice given in [26]. The lighter the gray intensity of the cell is, the more similar the neighbor prototypes are. The coarse boundaries between the well-separated clusters, indicated by the lines, can be seen through this visualization. However, finer details, such as the three subclusters in the left eye are obscured. (b) ISOMAP of the Clown data. While most clusters can be identified, two major clusters (the nose and the mouth), which are distinct in the U-matrix and in the CONNvis, are not separated.

nose and the mouth. Fig. 5 uses the same scheme to visualize CONN in the data space (which we can do for the special case of 2-D data) to show and validate the structures detected through CONNvis on the SOM grid in Fig. 4.

Fig. 6(a) shows the U-matrix visualization for the SOM of the Clown data. The boundaries between the well-separated natural clusters (such as the right eye and the nose and the body and the mouth) are clearly visible through the U-matrix. However, finer details, such as the three clusters in the left eye, are not emerging. We also compare CONNvis with ISOMAP in Fig. 6(b). ISOMAP is a commonly used manifold learning method [27], mapping a data set onto a 2-D space, while preserving the relationship of the pairwise distances. While most of the natural clusters can be seen through the ISOMAP of the Clown data, two major partitions, the nose and the mouth, are not separated. Because ISOMAP aims to find one underlying submanifold, it may provide a better topographic mapping than the SOM for data sets with no discontinuities. However, for the same reason, ISOMAP may be less informative for cluster extraction than CONNvis, especially for high-dimensional complicated data.

E. CONNvis for Complicated, Large Data Sets

For maps where the number of data vectors is much larger than the number of prototypes, the connectivity strengths span a large range of values. Using a different line width for each connectivity strength becomes infeasible due to limitation by screen resolution and the discrimination capability of the human eye. To help this, line widths can be based on a binning of the $\mathrm{CONN}(i,j)$ values as follows:

$$width(i,j) = \begin{cases} 1 & t_2 > \mathrm{CONN}(i,j) \geq t_1 \\ 2 & t_3 > \mathrm{CONN}(i,j) \geq t_2 \\ \vdots & \vdots \\ n-1 & t_n > \mathrm{CONN}(i,j) \geq t_{n-1} \\ n & \mathrm{CONN}(i,j) \geq t_n \end{cases} \quad (5)$$

where n is a small number. A good choice as threshold t_k is the mean strength μ_{n-k+1} of the $(n-k+1)$th ranking connections: $t_1 \in \{\mu_n, \mu_{n+1}, \ldots, \mu_N\}$, $t_k = \mu_{n-k+1}$, where N is the number of prototypes. This choice provides an automated selection of thresholds based on internal data characteristics as described in the following paragraph. It also employs the limited number of bins efficiently, because each bin reflects the global importance of one rank. Its resolution not only distinguishes strong connections but also reveals weak connections between (separated) clusters.

The above choice of the binning thresholds is motivated by the statistics of connectivity strengths shown in Figs. 7 and 8. These examples are for a 6-D synthetic data set and for an 8-D real remote sensing data set, respectively. Both of these data sets will be described in detail in Section IV. Fig. 7(a) and (b) gives the distribution of connections over ranks for these data sets, respectively. Perhaps surprisingly, the number of neighbors in data space (the number of connection ranks of a given prototype) can be higher than 12 for the 6-D data and more than 20 for the real 8-D data. However, Fig. 8, which shows the average connectivity strength within each rank for the 6-D data and 8-D data sets, tells us that even though the maximum number of connections for a prototype is much larger than 8 (16 for 6-D data and 29 for 8-D data), the average connectivity strength drops sharply after the fourth strongest connection (rank 4) and becomes negligibly small after the eighth strongest connection (rank 8). Fig. 8 also indicates that these averages decay exponentially. This observation suggests these averages as thresholds so that the binning of the line width can reflect the nonlinear distribution of the connectivity strengths. The thresholds chosen this way thus produce relatively wide bins for high ranking connections and narrow bins for weak connections, which can provide a good resolution. It also automatically excludes connections with strengths smaller than t_1. This is advantageous since the

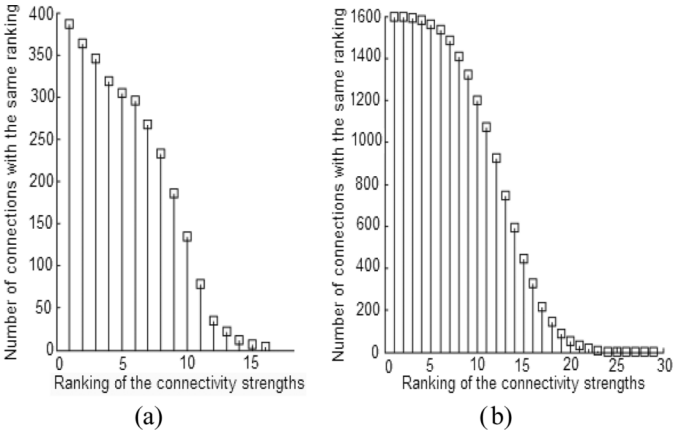


Fig. 7. Number of connections with the same ranking, across all SOM prototypes, for two different data sets. (a) For a 6-D synthetic, low-noise data set (128×128 pixel image in Fig. 9 mapped to a 20×20 SOM). Half of these prototypes have at least eight connections, and some have as many as 16. (b) For an 8-D real, noisy remote sensing image discussed in Section IV (512×512 pixel image mapped to a 40×40 SOM). Ninety percent of the prototypes have at least eight connections, and some have more than 25.

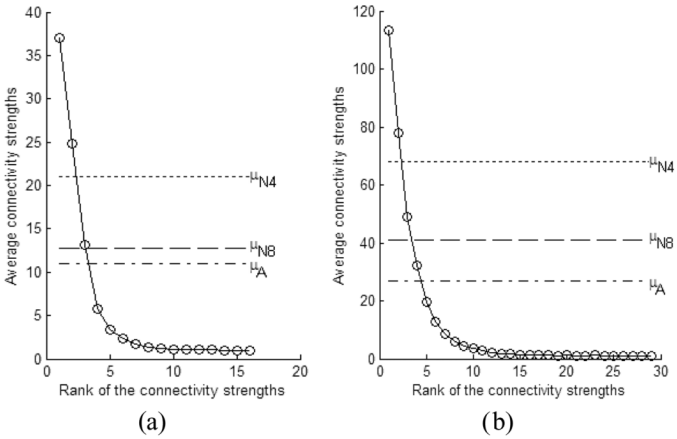


Fig. 8. Average connectivity strengths for the (a) 6-D data set and (b) 8-D data set. μ_A is the mean of all connections, μ_{N4} is the mean of connections with $length_{L2-\text{norm}} = 1$ (connections between the prototypes that are in 4–neighborhood in a rectangular lattice), μ_{N8} is the mean of connections with $length_{\max\ \text{norm}} = 1$ (connections between the prototypes that are in 8–neighborhood). The average strengths drops sharply after rank 4 and becomes negligibly small after rank 8, even though the number of connections is much higher (see Fig. 7).

connections with such low strengths are likely to be caused by noise or outliers.

One might be inclined to use equally sized bins between $t_1 = \mu - 2\sigma$ and $t_n = \mu + 2\sigma$ where μ and σ are the mean and standard deviation of all connections, respectively. However, using all connections in calculating μ and σ may produce very small μ and large σ, which is not useful for an informative binning in (5) due to noisy connections or outliers. More reasonable thresholds for extracting cluster structure and suppressing noise may be achieved by using the immediate SOM lattice neighbors, as some noisy and violating connections may be excluded from the statistics in that case. Fig. 8 shows μ_A, μ_{N8}, and μ_{N4} for the 6-D data and 8-D data sets where μ_A is the mean of all connections while μ_{N8} and μ_{N4} are the means of connections of prototypes that are also immediate lattice neighbors (eight neighbors or four neighbors for rectangular lattice), respectively. For the 6-D data set, there are few topology violating connections, therefore μ_{N8} is very similar to μ_A. For the complicated 8-D data set, μ_{N8} is much larger than μ_A. However, the binning may still not be useful to discriminate among strong connections because thresholds set by μ_{N8} and σ_{N8} will only bin the low-strength connections while lumping most rank 1 and rank 2 connections into one bin, as can be seen from Fig. 8. μ_{N4} and σ_{N4} produce a high μ and relatively low σ but equal-size bins diminish its usefulness. The thresholds t_k $(1 \leq k \leq n)$ and n should support the specific data and application, which may also call for other approaches to binning, such as in [28].

IV. Clustering Through CONNvis

CONNvis guides accurate capture of cluster boundaries by showing how strongly (weakly) various parts of the data manifold are connected. It provides a tool to filter out weak connections, which are mostly caused by noise or negligible residual errors in the learning, and therefore, are unimportant for the description of the data structure. Since connections across cluster boundaries are typically weak and few, filtering out weak connections using the automatic thresholding, described in Section III-E, can result in almost clean-cut boundaries, outlining "coarse clusters." In the following, we will give a recipe of the exact procedure of cluster extraction including steps to separate the coarse clusters interactively.

How do "entanglements" (topology violations) affect our cluster extraction procedure? Fortunately, CONNvis shows the exact locations as well as the severity (the strength) and the folding length of the violating connections. In a reasonably well-trained SOM most violating connections are weak, and at short folding lengths, not extending across clusters. Severe violations (long, thick lines), when present, are signs of incorrect mapping. With only a small number of such connections, one can verify and recover twisted clusters manually, ignoring these connections (temporarily visually removing them) while evaluating the rest. With a large number of strong global violations, a new SOM training may be needed.

We define a "global violation" as a connection with a folding length exceeding the radius of the "tightest" SOM neighborhood into which all prototypes that are neighbors in data space should be packed when the mapping is (as) topology preserving (as possible). The tightest SOM neighborhood depends on the data, and is defined and computed the following way: let m be the maximum number of prototypes adjacent to any prototype in the data manifold. For a rectangular lattice, the number of neighbors within $length_{L1-\text{norm}} = l$ is $8l$, hence the number of neighbors within $length_{L1-\text{norm}} \leq n$ is $8\sum_{l=1}^{l=n} l$. Thus, m prototypes can fit into the tightest neighborhood with the neighbor prototypes within a distance of $l_{\min}$ where

$$m \leq \min_{l_{\min}} \sum_{l=1}^{l=l_{\min}} 8l. \tag{6}$$

Any connection with length greater than $l_{\min}$ will then be called a "global violation." For example, for the maximum number of connections given in Fig. 6(a), which is 16, connections with $length > 2$ will be global violations, and

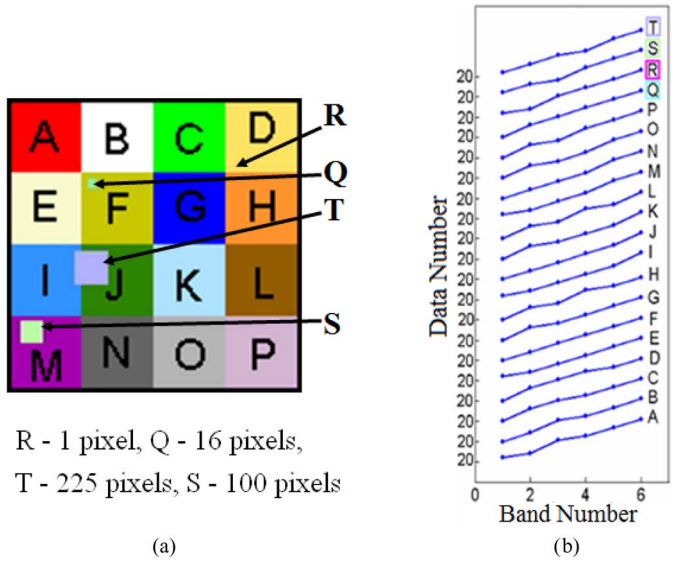


Fig. 9. The 6-D (6-band) synthetic image data set consisting of 20 classes. Each pixel is a 6-D stacked feature vector. (a) Spatial distribution of classes in the 128 × 128 pixels image. Four classes are relatively rare (R, Q, T, and S). (b) Mean of the feature vectors for each class, vertically offset for clarity. The data number shows the mean feature value of classes at the corresponding image band numbers.

similarly, for the one in Fig. 6(b), which is 29, connections with $length > 3$ will be global violations.

After analysis of the global violating connections, the manual cluster extraction is based on the strength (width) and the rank (color) of the connections as well as on the number of connections between the prototypes bridging coarse clusters. We remove weak connections (those with the lowest strength) that link any two coarse clusters X and Y at their boundary, as follows.

Step 1) Remove all weak connections to cluster X if the number of weak connections to X is less than the number of weak connections to the other cluster Y.

Step 2) Remove the weakest connection if the connections of the prototype to the two clusters have different widths.

Step 3) Remove the lowest ranking connection if the number of weak connections to both clusters is the same and all connections at the boundary of these clusters are weak.

Step 4) Repeat Steps 1)–3) until this prototype has been disconnected from one of the clusters.

Step 5) Repeat Steps 1)–4) for all prototypes at this boundary.

Below, we give two examples of cluster extraction from CONNvis, one for a synthetic noisy image data set consisting of 6-D pixel vectors, and one for a real, noisy remote sensing image data set with 8-D pixel vectors (6-band and 8-band images, respectively). In our CONNvis of these images, we use a 4-level binning with thresholds determined by averages of same ranking connectivity strengths, as described in Section III-E under (5). This binning has provided sufficient resolution for cluster capture for the cases we present here. We call a connection with $width = 1$ "weak" (unimportant), and a connection with $width > 1$ "strong." This distinction, derived from the statistics of the data, works well for our applications in this paper.

TABLE I
REMOVAL OF CONNECTIONS AT THE BOUNDARIES OF THE COARSE CLUSTERS IN THE CONNVIS OF THE 20-CLASS DATA SET FOR CLUSTER EXTRACTION

	Reason for removal		
	Low number of connections	Weak connections	Low ranking connections
Connections at the boundaries of coarse clusters	G-H	P-Q	A-B
	F-G	P-S	B-F
	F-J		C-D
	E-I		C-G
	M-I		N-O
	P-O		

A. An Explanatory Example for Cluster Extraction

To illustrate cluster extraction from CONNvis, we use a synthetic 6-band, 128 × 128 spectral image. A spectral image is composed of images acquired simultaneously at a given set of wavelengths and registered together. At each pixel, the data vector composed of the measured values at the n wavelengths (image bands) is the spectrum of the material in that pixel. The spectra are the n-dimensional input vectors to the clustering. Our synthetic image consists of 20 known classes, four of which are rare. The data vector at each image pixel was generated from the mean vector of the class that the given pixel belongs to, by adding 10% 6-D Gaussian noise to it. Fig. 9(a) shows the spatial layout of the classes in the image, color coded, and annotated with labels A-S. The color coding of these classes is shown in the online version. Fig. 9(b) displays the mean feature vectors (signatures) of the classes. To include all connections in the CONNvis of the SOM of this data set, as in Fig. 10(a), we set $t_1 = 0$. The unconnected and weakly connected prototypes form nearly empty corridors, which outline coarse cluster boundaries. The known cluster labels are shown in Fig. 10(b)–(d) to help discuss this cluster extraction procedure. Some clusters such as classes R and T are already outlined in the initial view [Fig. 10(a)] by unconnected neighbor prototypes. One can start from here, pruning connections based on our interpretation of CONNvis. First, we observe that most topology violations are weak and the majority of them are between the prototypes in the same (coarse) cluster. There is no strong global violation (no long thick line) in this case. We then remove weak global topology violations ($width = 1, length > 2$). This results in clear separation of some classes from others [Fig. 10(b)], such as classes K and L. The choice of $length > 2$ is given by (6), with the maximum number of connections for any prototype being 16 for this data set [Fig. 7(a)].

From this visualization, we can start manual extraction as prescribed by Steps 1)–5). Fig. 10(b) and (c) illustrates the extraction of clusters. For example, in Fig. 10(b), clusters E and I are bridged by a prototype (shown as a black dot), which has two weak lines (one red and one blue) to I and a weak blue one to E. In Fig. 10(c), the connection to E has been removed (as per Step 1) because E had only one connection to this prototype whereas

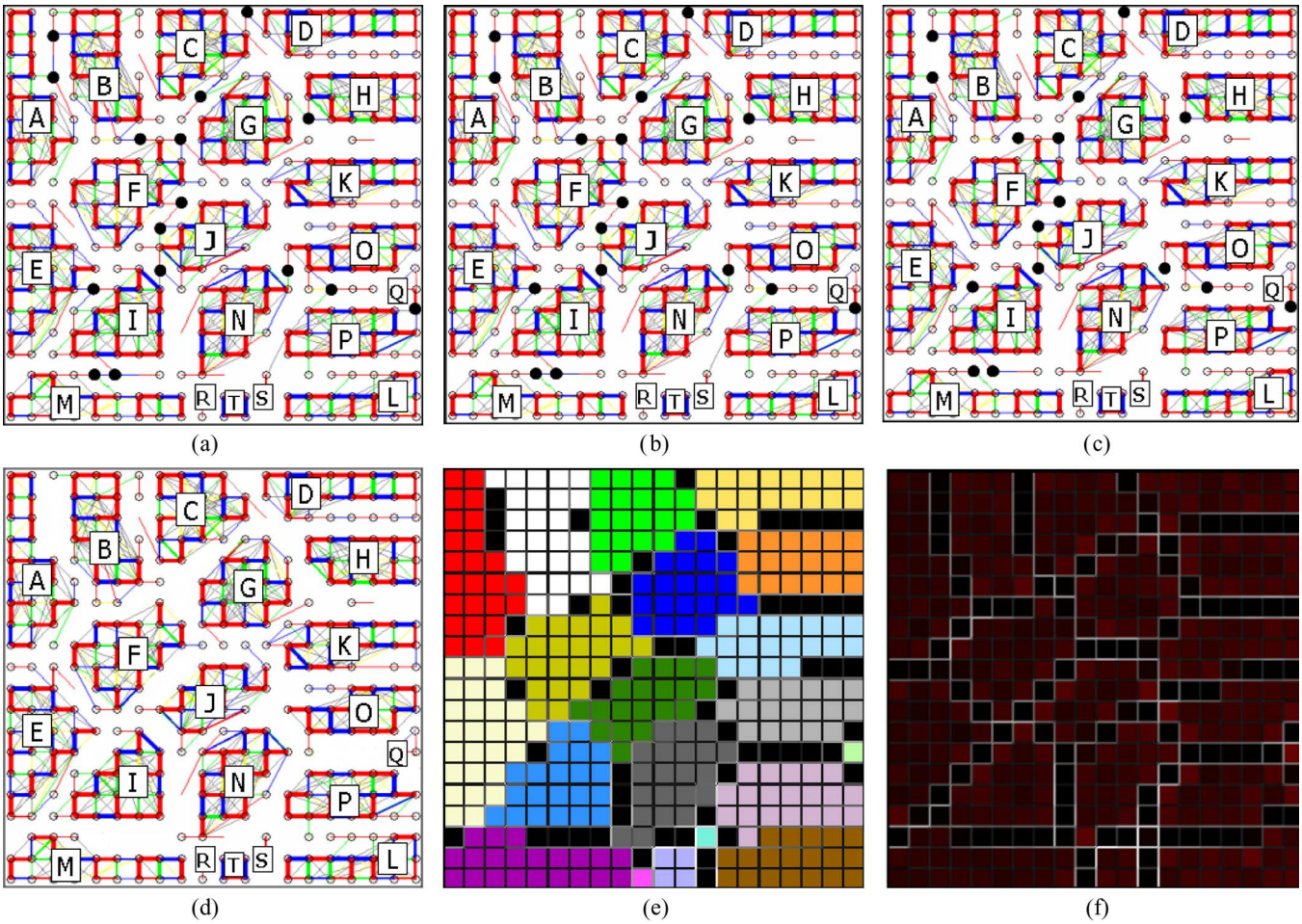
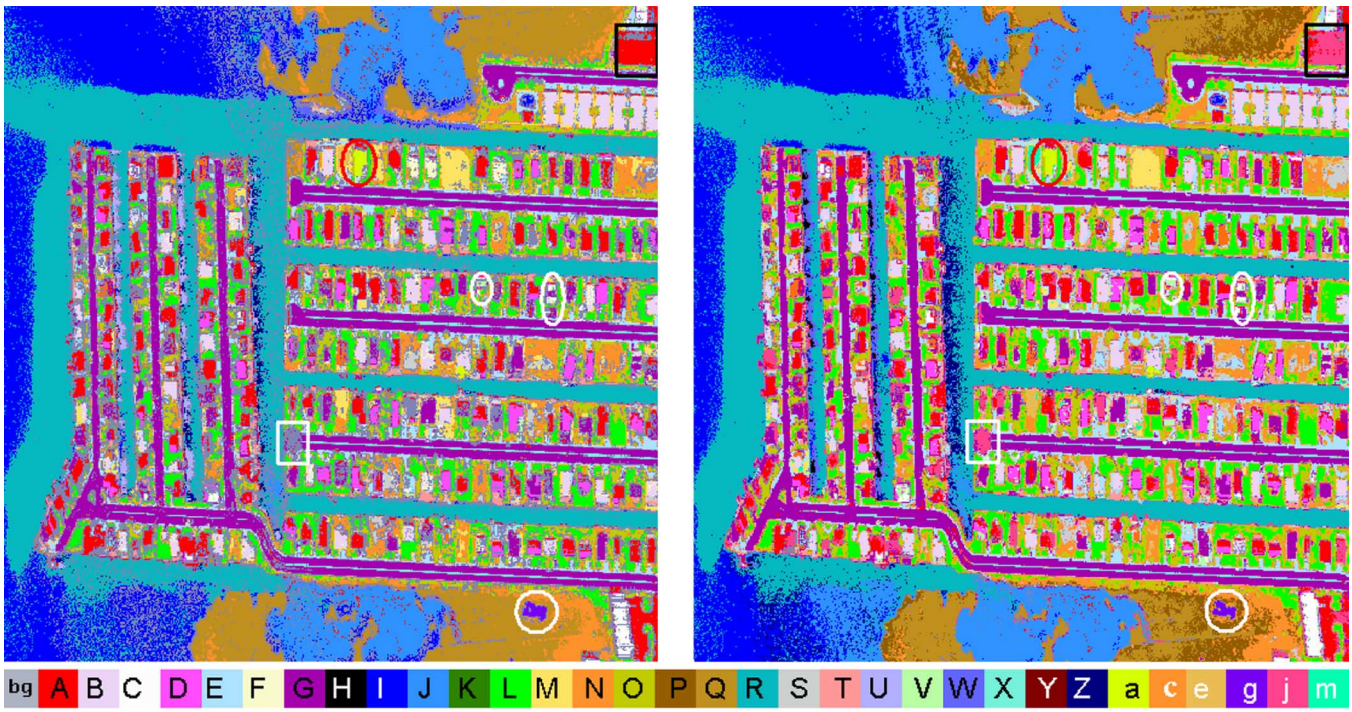


Fig. 10. (a) CONNvis on the SOM lattice for the 6-D, 20-class data. The prototypes are at the junctions of the connections (open circles). Thin connections mean weak similarity. CONNvis reveals coarse clusters through the nearly empty corridors formed by unconnected or weakly connected prototypes. In this case, topology violations remain mostly within these coarse clusters. (b) Weak global violating connections [$length > 2$ for this case as per Fig. 7(a) and (6)] have been removed. The prototypes at the boundaries of coarse clusters are shown by black dots. The coarse clusters are annotated with the known class labels. (c) The weak connections that link two coarse clusters have been removed based on the criteria given in Section IV-A in Steps 1)–5). All classes are correctly identified this way, including the rare ones (R, T, S, and Q). Some of the border prototypes with small receptive fields may need a second look to decide if they really belong to the respective cluster or should be regarded as outliers. (d) Same as in (c), but the border prototypes are removed. (e) The known labels, color coded as in Fig. 9 (shown in the online version), are overlain on the SOM for verification of the extracted clusters. The color of a grid cell shows the cluster membership of its prototype. (f) Modified U-matrix (mU-matrix) over the same SOM. The intensity level of each grid cell is proportional to the size of the receptive field of the corresponding prototype. The intensity of the "fences" between each pair of grid cells, in all eight directions (including the diagonals) is proportional to the Euclidean distance of the respective pair of prototypes, in data space. White fence is large dissimilarity, and dark means strong similarity. Most of the grid cells between the double fences are empty, or have very few data points mapped to them. The white fences perfectly delineate the 20 known classes. For this simple data set, the mU-matrix and CONNvis provide equally good clues for the determination of the cluster boundaries.

I had two. Removal of the two other connections between I and this prototype depends on further choices made by the user. One choice can be the inclusion of this prototype in the cluster to which it has the strongest connection (I), as shown in Fig. 10(c), because it is most similar to that cluster with respect to the data. Another choice can be the exclusion of this prototype, as well as all those at the cluster boundaries [as in Fig. 10(d)], which have very small receptive fields because they are often representatives of noise or outliers. M and I are separated through a similar procedure as E and I. An example for separation based on weakness of the connections (as in Step 2) is the separation of clusters P and Q. The prototype at the boundary of P and Q has one connection to Q and two connections to P, but the connections to P are weak whereas the connection to Q is a strong one, hence the weak connections are removed. Clusters C and G share a prototype, which is connected with one weak connection to each. However, the connection to G is lower ranking (green, lower strength) than the connection to C (blue); thus the green connection is removed (Step 3). Table I lists the pairs of clusters, which have a common boundary and the method for removing connections between those clusters. By this semimanual procedure, the clusters are extracted easily. For comparison, we overlay the known labels on the SOM as shown in Fig. 10(e) (color coded as in Fig. 9). Here, each grid cell represents a prototype, located in its center. The cell is colored according to the cluster membership of its prototype. The extracted clusters in Fig. 10(d) show a striking match to the true clusters.

We show a modified U-matrix (mU-matrix) representation in Fig. 10(f) to illustrate the differences in knowledge representation between CONN and U-matrix type (distance based) visualizations. First, we need to point out that the mU-matrix (our modification of the U-matrix [10]) is more detailed than



(a) (b)

Fig. 11. Comparison of cluster maps of the 8-band 512 × 512 pixel remote sensing image of Ocean City, MD. There are unclustered pixels in both cluster maps indicated by the background color "bg." (a) Earlier cluster map extracted by using a mU-matrix (see [13] for details). Red and white ovals point out the locations of rare clusters [C, V, a and g in Fig. 12(b)]. Clusters extracted from CONNvis in Fig. 12. The agreement between the two cluster maps is very good. In (a), there are more pixels unclustered than in (b), which results in more appearances of the background color "bg" in (a), and more coverage by some colors such as turquoise and green in (b). We also easily capture the formerly identified rare clusters (shown in the ovals). Some clusters in (a) are split into subclusters in (b). An example is the cluster A (red, concrete) which is split into A (red) and j (pink). A region that is clustered as j is the large building with concrete roof outlined by a black rectangle at the top right of the image. Subcluster j also covers some regions that are not clustered in (a), for example, the end of a road outlined by the white rectangle in (a) and (b). See Fig. 12 for their labels and locations in the SOM.

the original U-matrix. Instead of displaying the average of the distances to the SOM neighbor prototypes over each grid cell, we display the individual distances to each neighbor in all directions including the diagonal neighbors. This allows crisper delineation of cluster boundaries than with the U-matrix, and facilitates the detection of small clusters such as R and T at the bottom center in Fig. 10(f) [hot pink and grayish blue, respectively, in Fig. 10(e)]. The cluster R is represented by a single prototype that has large distances to all of its SOM neighbors. In the customary U-matrix, the average of the distances to its neighbors would assign a high-intensity color (nearly white) to the entire cell containing R, separating the cluster to its right (T) from the cluster to its left (M), while R itself would disappear under this wide fence. The cluster T, which has four prototypes, would be smeared because the averaging would produce a medium high fence on each of those four prototypes. For this simple data set, with low noise and slightly overlapping clusters, the mU-matrix and CONNvis work equally well for cluster capture. For complicated data, however, CONNvis offers more support. We demonstrate this next.

B. A Real-Data Application

A real remote sensing spectral image of Ocean City, MD, comprising 512×512 pixels, represents fairly complicated data. Each pixel has an 8-D feature vector, called spectrum, associated with it. The feature vector is composed of the measured radiance values at a given set of wavelengths. The image was acquired on April 30, 1997, with a Daedalus AAds-1260 multispectral scanner, which records data in 12 spectral bands, ten in the 0.38–1.1-μm range, and two in the 11–14-μm thermal infrared region. The flight altitude of approximately 600 m and a FOV of 2.5 mrad yield an average of approximately 1.5 m per pixel ground resolution [29]. The first two and the last two spectral bands were excluded from our processing because of extreme noise.

Ocean City is a long linear settlement on the seashore with rows of closely spaced buildings separated by straight parallel roads and water canals. The spatial layout of different surface types in the city is shown in Fig. 11(a) through an earlier cluster map [13] where different colors label spectrally different materials. Ocean (blue, I) surrounds the city from the left ending in small bays (medium blue, J, at the top center and bottom center of the scene), which contain suspended sediments and algae. These small bays are surrounded with coastal marshlands (brown, P; ocher, Q). Shallow water canals (turquoise, R) separate the double rows of houses, trending in roughly North–South (N–S) direction in the left of the scene and East–West (E–W) direction in the right of the scene. The canals provide a waterway to boats. Many houses here have private docks (flesh-colored pink, T) and as a consequence, dirty water at such locations (black, H). The streets have paved roads (magenta, G)

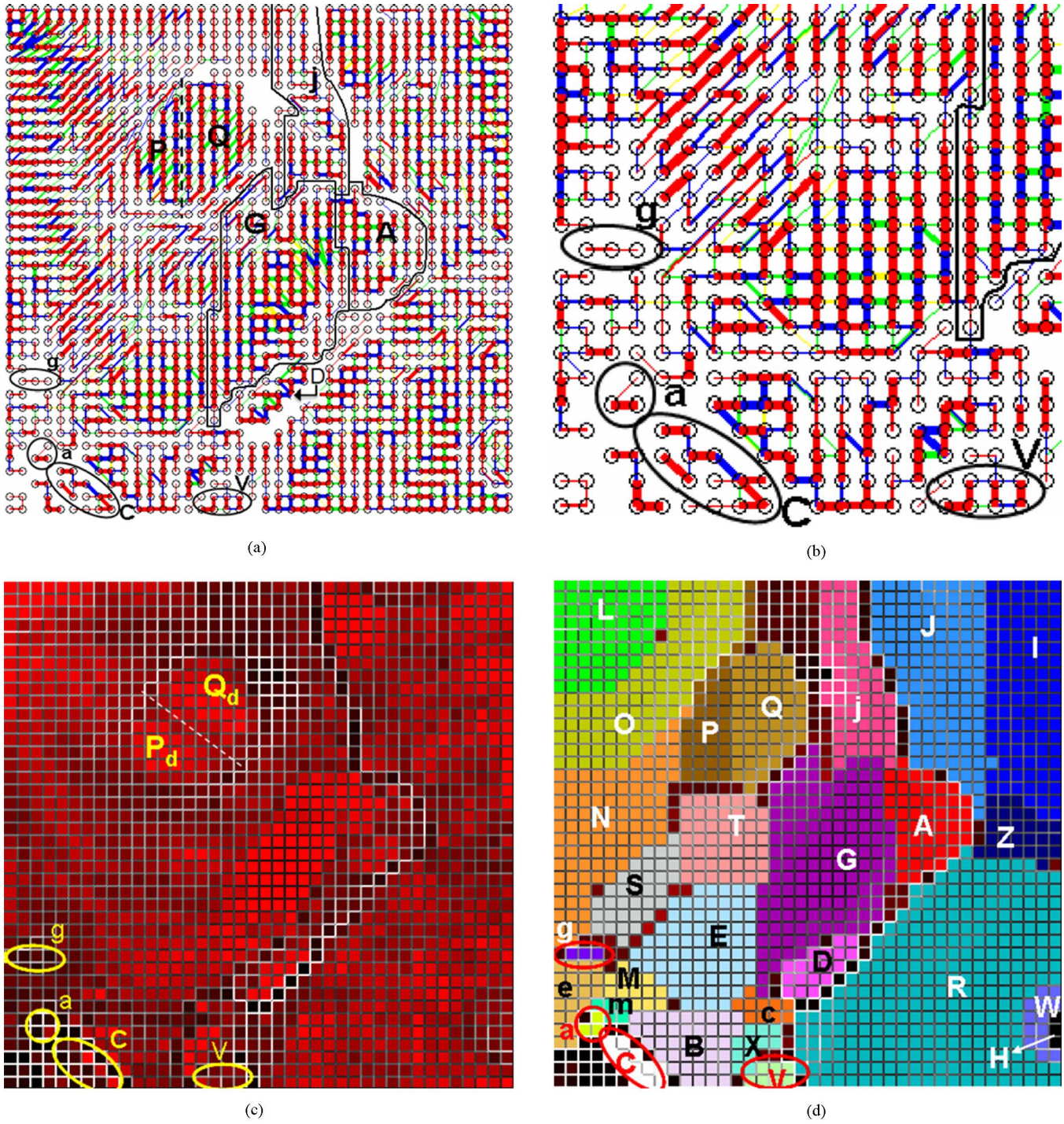


Fig. 12. Cluster extraction for the Ocean City, MD, data based on CONNvis. (a) CONNvis for the 40 × 40 rectangular SOM lattice. The prototypes, shown by circles, are at the junctions of the connections. The weak global violations were excluded. As an example for cluster extraction, clusters G, j, and A are outlined by solid black lines. The ovals point out small clusters (C, V, a, and g) extracted in previous work [13]. The separation of those clusters is clearly indicated by the lack of connections to other clusters. Separation between clusters P and Q is shown by the dashed vertical line where the connectivity strengths between P and Q are weak. (b) An enlarged view of the bottom left quarter of the CONNvis in Fig. 12(a) This view provides a clearer representation, for easier interpretation. (c) mU-matrix. The prototypes are at the center of the grid cells. The density distribution is also shown by the red intensity of the cells. Boundaries of some clusters, small clusters (C, V, g, and a shown in ovals) in particular, are obscure because of the high fence values between the prototypes within the same clusters. (d) The extracted clusters from CONNvis annotated with the respective labels.

with reflective paint in the middle (light blue, E). The colors of small rectangles, outlining houses, indicate various roof materials (A, B, C, D, E). Typical vegetation types around buildings are healthy lawn, trees, and bushes (pure green, L), yellowish lawn (split-pea green, O), and dry grass (orange, N). There are also some rarely occurring material types that only exist at the locations shown by the ovals in Fig. 11(a). The spatial extent of the largest one (white, C) is 0.4% of the image while one type of roof material (pale green, V) in the narrow white oval has only 239 pixels and the material (dark purple in the white circle, g)

in the middle of the marshland at the bottom of the scene has 251 pixels.

The cluster map in Fig. 11(a) was produced in an earlier work from mU-matrix representation of a 40×40 SOM [13] and was verified against expert knowledge. We take this cluster map as a baseline and show that we achieve the same quality of clustering or better, using CONNvis. We take the same 40×40 learned SOM as was used for the capturing of the clusters from a mU-matrix, in Fig. 11(a), and apply the procedure described in Section IV. The statistics of the connections, given in Figs. 7(b) and 8(b), indicate that the prototypes have up to 29 connections and large average strengths even for low ranking connections. As described at the beginning of Section IV, a large number of strong connections at high ranks, in general, indicates folding whereas the existence of low-strength connections at small ranks most likely indicates noise.

Fig. 11(b) presents the cluster map extracted from CONNvis. The general agreement between the two cluster maps in Fig. 11 indicates a good clustering based on CONNvis. In what follows, we discuss the processing and point out similarities with, and improvements to, the mU-matrix-based cluster map in Fig. 11(a).

Fig. 12(a) is the CONNvis of the SOM with a 4-level binning where thresholds are the average strengths of connections of the same rank [Fig. 8(b)]. This results in $t_4 = \mu_1 = 113, t_3 = \mu_2 = 78, t_2 = \mu_3 = 50$, and $t_1 = \mu_4 = 35$. The connections with $length > 3$ and $width = 1$ are removed since they are weak global violations by the argument presented earlier in Section IV about the relationship between the maximum number of connections per prototype and the SOM neighborhood radius within which they can map without being considered globally violating (6). In the CONNvis [Fig. 12(a)], some coarse clusters are very obvious. One example is the cluster near the lower center (D) with a wide empty corridor at one side. Some of the small clusters (g, a, C, V) are clearly separated. Other coarse clusters may be harder to recognize in this busy figure. To help the reader, we outlined a few (but not all) coarse clusters (A, G, j) with solid black lines. These also have nearly empty corridors around them (where the black lines show), which means they are just as well separated as the ones with the wide corridors around them, but these corridors have only the width of one cell, and therefore, are more difficult to see. Fig. 12(b) shows an enlargement of the bottom left quarter of Fig. 12(a), for an easier interpretation.

For reference, Fig. 12(c) shows a static snapshot of the mU-matrix view of the same SOM. An interactive process was used to find clusters by adjusting the intensity (gray) levels of the fences between grid cells, which provides maximum flexibility in viewing the distances between prototypes. However, the inherent limitations of what is visualized can conceal some—real, existing—details because the distances between prototypes, by themselves, do not necessarily reveal all structural variations. An example is the small cluster "a" in the lower left corner of the SOM. When we use the connectivity strength as the similarity measure as in Fig. 12(a), the prototypes reveal the small clusters that are harder to find in other visualizations.

The dissimilarities, indicated by high fence values in the mU-matrix, are shown by the corridors outlined by no or weak connections in the CONNvis [Fig. 12(a)]. The boundaries between the small clusters, labeled C, V, a, and g, in the lower left corner of the SOM become obvious and are extracted easily and fast with the CONNvis [Fig. 12(b)]. Because of either clear separation or weak connections, it is also much easier to capture other clusters from CONNvis with more certainty. All extracted clusters are shown in the SOM in Fig. 12(d).

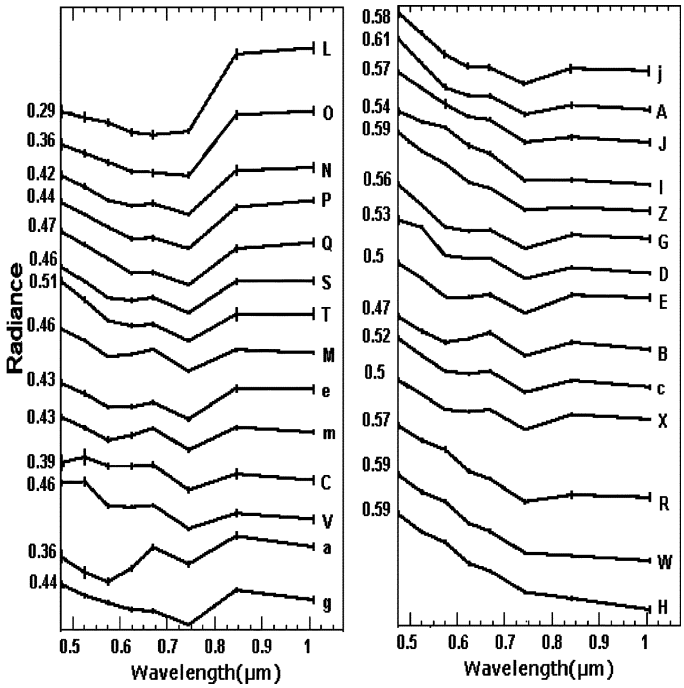


Fig. 13. Mean signatures (feature vectors) of the extracted clusters shown in Fig. 12(d). The signatures are offset for clarity, with standard deviations shown by the vertical bars. All represent different materials, verified from ground truth [29]. The subtle differences between some of the signatures indicate the clustering challenge which CONNvis effectively handled.

Fig. 13 shows the averages and standard deviations of the spectral radiance signatures (feature vectors) of the extracted clusters. Many signatures are distinct, but some are very similar with slight differences, which pose a clustering challenge. The small clusters C, V, a, and g have unique signatures, yet it was difficult to find them in the mU-matrix. Using CONNvis clearly helped capture the clusters including rare ones, in this data set.

One difference between the two cluster maps in Fig. 11 is that cluster A (red, concrete) in Fig. 11(a) is split into two subclusters A (red) and j (pink) in Fig. 11(b) because of the weak connections between them in the CONNvis. An example of the subcluster j is the large building with concrete roof at the top right of the image, in a black rectangle. Another region clustered as j is the end of a road, shown in a white rectangle, which remained unclustered (colored "bg") in Fig. 11(a). The signatures of A and j have appreciable differences (Fig. 13).

Another difference is the detection of clusters P (brown) and Q (ocher). The CONNvis delineates the border between the clusters P and Q through the weak connections across them. The mU-matrix in Fig. 12(c) clearly indicates the boundary around the combined cluster P $\cup$ Q by high fence values. However, the separation between P and Q is hard to distinguish even by tuning the fence heights to scrutinize the local similarity relations of the prototypes. This leads to the extraction of P

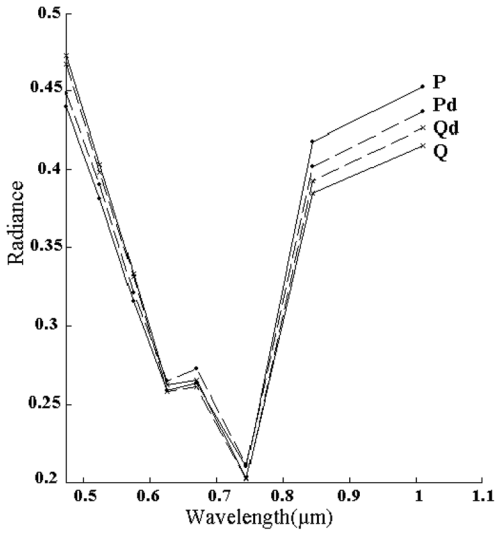


Fig. 14. Comparison of the signatures of P, Q to P_d and Q_d. P and Q are extracted from CONNvis in Fig. 12(a) while P_d and Q_d are extracted based on the density distribution seen in Fig. 12(b). The signatures of P and Q are more distinct than those of P_d and Q_d.

and Q together as one cluster from the mU-matrix [Q, ocher regions in Fig. 11(a)].

As an additional point, one might be tempted to think that the separation of clusters P and Q is along the direction of the dashed line shown in Fig. 12(c) because the density in that direction is much lower than the density of the surrounding prototypes. P_d and Q_d label the subclusters of P $\cup$ Q extracted based on this density evaluation. In contrast, according to the CONNvis [Fig. 12(a)], P $\cup$ Q should be separated in the vertical direction indicated in Fig. 12(a) due to weak and few connections between prototypes. We denote the resulting subclusters by P and Q. Fig. 14 compares the mean signatures of P, Q, P_d, and Q_d. The signatures of P and Q are more different from each other than the signatures of P_d and Q_d. This demonstrates that density distribution may be misleading for cluster identification due to the fact that it only shows the total receptive field size of the prototypes but does not show how the data is distributed among the neighbor prototypes. Yet many practitioners rely on receptive field size alone for cluster identification. The CONNvis shows the prototypes that are neighbors in data space and the data distribution among them, which in turn produces a better topology representation for cluster extraction.

We made an attempt to compare the quality of the two cluster maps in Fig. 11 quantitatively, by using two commonly accepted cluster validity indices. One is the Davies–Bouldin index (DBI), which is based on centroid distance metrics [30]. The other index is the generalized Dunn index (GDI) with centroid linkage as between cluster distance metric and average distance to centroid as within cluster distance metric. The best clustering is indicated by the minimum (maximum) index of DBI (GDI). GDI favors the CONN-based clustering (GDI = 0.63) over the mU-matrix clustering (GDI = 0.44) whereas DBI favors the mU-matrix clustering over CONNvis clustering (DBI = 1.03 versus 1.30). One contributing factor to this contradiction may be that the two maps contain different numbers of unclustered prototypes. We do not know of any validity index that has been shown to yield meaningful comparison in such situation.

V. Discussions and Conclusion

We define a new connectivity measure for the similarity of SOM prototypes. It integrates data distribution into the customary Delaunay triangulation, which, when displayed on the SOM grid, enables 2-D visualization of the manifold structure regardless of the data dimensionality. We are not aware that other existing SOM visualizations have this capability. This representation also enables more detailed detection of manifold structures than the ones that work solely with prototype distances or those that are limited to low dimensions. An additional contribution is that the binning scheme used in CONNvis is automatically derived from internal data characteristics. This makes the CONN matrix (divorced from visualization) suitable for automation of cluster extraction. Our CONNvis is also unique among SOM representations in that it shows both forward and backward topology violations on the SOM grid. This allows the assessment of the quality of SOM learning, data complexity, and dimensionality match between the data manifold and the SOM, and thus helps decide whether correct data mining is possible or a new SOM learning or modification of the grid structure is necessary.

An unresolved issue with this representation and its use in cluster capture is that the binning scheme, defined globally in Section III-E, may be ineffective for some large data sets. This is because the global scheme gathers all connectivity strengths of prototypes in high-density regions of the SOM into the largest bin. That results in hiding the underlying (sub)cluster structures in those regions. One way to overcome this problem may be a region-based binning by using local statistics of the connectivity strengths within subregions of the SOM. For example, one could calculate the means μ_i separately for user-defined subregions of different connectivity densities.

An interesting open problem is how to compare clusterings produced by different methods that can leave some prototypes unclustered. For example, in the semimanual clustering we described in this paper, some prototypes remain unclustered because of uncertainty on the part of the analyst based on (limited) visualization. Automated methods can also produce unclustered prototypes with various threshold (or parameter) settings. The number of unclustered prototypes can vary based on the decision of the analyst or the threshold settings in an automated procedure. In such situation, it is unclear how useful existing cluster validity indices can be. This may necessitate the development of new measures to provide meaningful assessment of cluster validity under such circumstances.

Finally, we point out that the connectivity matrix CONN is applicable to prototypes obtained by any quantization process since the knowledge represented by CONN is independent of visualization. It can be integrated into similarity measures in any prototype-based clustering algorithms in addition to the more customary distance-based similarity measures.

Acknowledgment

The authors would like to thank Prof. B. Csathó from the Department of Geology, University of Buffalo, for the Ocean City

image and ground truth, and to Dr. Alhoniemi and Dr. Vesanto from Department of Information Technology, Turku University, for sharing their Clown data as well as the SOM weight vectors from their processing in [26]. They would also like to thank the anonymous reviewers for their valuable comments. P. Tracadas from Rice University contributed greatly with software development for the HyperEye environment,[1] in which the presented higher dimensional simulations were run.

REFERENCES

[1] T. Kohonen, *Self-Organizing Maps*, 2nd ed. Berlin, Germany: Springer-Verlag, 1997.
[2] E. Merényi, "Precision mining of high-dimensional patterns with self-organizing maps: Interpretation of hyperspectral images," in *Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence. Studies in Fuzziness and Soft Computing*. Berlin, Germany: Physica-Verlag, 2000, vol. 54.
[3] T. Villmann and E. Merényi, "Extensions and modifications of the Kohonen SOM and applications in remote sensing image analysis," in *Self-Organizing Maps: Recent Advances and Applications*, U. Seiffert and L. C. Jain, Eds. New York: Springer-Verlag, 2001, pp. 121–145.
[4] G. Pölzlbauer, A. Rauber, and M. Dittenbach, "Advanced visualization techniques for self-organizing maps with graph-based methods," in *Proc. 2nd Int. Symp. Neural Netw.*, Z. Y. Jun Wang and X. Liao, Eds., Chongqing, China, Jun. 1, 2005, pp. 75–80.
[5] J. Venna and S. Kaski, "Neighborhood preservation in nonlinear projection methods: An experimental study," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2001, vol. 2130, pp. 485–492.
[6] T. Martinetz and K. Schulten, "Topology representing networks," *Neural Netw.*, vol. 7, no. 3, pp. 507–522, 1993.
[7] B. Fritzke, "A growing neural gas network learns topologies," in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1995, pp. 625–632.
[8] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Netw.*, vol. 15, no. 8-9, pp. 1041–1058, 2002.
[9] M. Aupetit, "Learning topology with the generative gaussian graph and the EM algorithm," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 83–90.
[10] A. Ultsch, "Self-organizing neural networks for visualization and classification," in *Information and Classification-Concepts, Methods and Applications*, O. B. Lausen and R. Klar, Eds. Berlin, Germany: Springer-Verlag, 1993, pp. 307–313.
[11] M. Kraaijveld, J. Mao, and A. Jain, "A nonlinear projection method based on Kohonen's topology preserving maps," *IEEE Trans. Neural Netw.*, vol. 6, no. 3, pp. 548–559, May 1995.
[12] A. Ultsch, "Maps for the visualization of high-dimensional data spaces," in *Proc. 4th Workshop Self-Organizing Maps*, 2003, vol. 3, pp. 225–230.
[13] E. Merényi, A. Jain, and T. Villmann, "Forbidden data," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 786–197, May 2007.
[14] M. Cottrell and E. de Bodt, "A Kohonen map representation to avoid misleading interpretations," in *Proc. 4th Eur. Symp. Artif. Neural Netw.*, 1996, pp. 103–110.
[15] E. Hakkinen and P. Koikkalainen, "The neural data analysis environment," in *Proc. 1st Workshop Self-Organizing Maps*, Espoo, Finland, Jun. 4–6, 1997, pp. 69–74.
[16] J. Himberg, "A SOM based cluster visualization and its application for false colouring," in *Proc. IEEE/INNS/ENNS Int. Joint Conf. Neural Netw.*, Como, Italy, 2000, vol. 3, pp. 587–592.
[17] S. Kaski, T. Kohonen, and J. Venna, "Tips for SOM processing and colourcoding of maps," in *Visual Explorations in Finance Using Self-Organizing Maps*, T. K. G. Deboeck and T. Kohonen, Eds. London, U.K.: Springer Finance, 1998.
[18] S. Kaski, J. Venna, and T. Kohonen, "Coloring that reveals cluster structures in multivariate data," *Austral. J. Intell. Inf. Process. Syst.*, vol. 6, pp. 82–88, 2000.
[19] J. Vesanto, "SOM-based data visualization methods," *Intell. Data Anal.*, vol. 3, no. 2, pp. 111–126, 1999.
[20] E. Pampalk, A. Rauber, and D. Merkl, "Using smoothed data histograms for cluster visualization in self-organizing maps," in *Proc. Int. Conf. Artif. Neural Netw.*, 2002, pp. 871–876.
[21] D. Merkl and A. Rauber, "Alternative ways for cluster visualization in self-organizing maps," in *Proc. 1st Workshop Self-Organizing Maps*, Espoo, Finland, Jun. 4–6, 1997, pp. 106–111.
[22] M.-C. Su and H.-T. Chang, "A new model of self-organizing neural networks and its applications," *IEEE Trans. Neural Netw.*, vol. 12, no. 1, pp. 153–158, Jan. 2001.
[23] H. Ressom, D. Wang, and P. Natarajan, "Adaptive double self-organizing maps for clustering gene expression profiles," *Neural Netw.*, vol. 16, pp. 633–640, 2003.
[24] J. Blackmore and R. Miikkulainen, "Visualizing high-dimensional structure with the incremental grid growing neural network," in *Proc. 12th Int. Conf. Mach. Learn.*, San Francisco, CA, 1995, pp. 55–63.
[25] H. Yin, "ViSOM- a novel method for multivariate data projection and structure visualization," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 237–243, Jan. 2002.
[26] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.
[27] J. B. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[28] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 194–2002.
[29] B. Csatho, W. Krabill, J. Lucas, and T. Schenk, "A multisensor data set of an urban and coastal scene," *Int. Arch. Photogramm. Remote Sens.*, pp. 26–31, 1998.
[30] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 28, no. 3, pp. 301–315, Jun. 1998.

[1]http://www.ece.rice.edu/HYPEREYE

Kadim Taşdemir received the B.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2001, the M.S. degree in computer science from Istanbul Technical University, Istanbul, Turkey, in 2003, and the Ph.D. degree in electrical and computer engineering from Rice University, Houston, TX, in 2008.

In September 2008, he became an Assistant Professor of Computer Engineering at Yasar University, Izmir, Turkey. His research interests include detailed knowledge discovery from high-dimensional large data, especially multi- and hyperspectral imagery, artificial neural networks, self-organized learning, data mining, and pattern recognition.

Erzsébet Merényi (M'98–SM'05) received the M.Sc. degree in mathematics and the Ph.D. degree in computer science from Szeged (Attila József) University, Szeged, Hungary, in 1975 and 1980, respectively.

Currently, she is a Research Professor in the Electrical and Computer Engineering Department, Rice University, Houston, TX. She previously worked as a Staff Scientist at the Lunar and Planetary Laboratory, University of Arizona, Tucson. Her interests include artificial neural networks, self-organized learning, manifold learning, segmentation and classification of high-dimensional patterns, data fusion, data mining, knowledge discovery, and application to information extraction from multi- and hyperspectral data for identification of surface composition, including geologic, ecosystem and urban mapping from planetary remote sensing imagery, and analysis of multivariate medical data. She has been analyzing data from various space missions and terrestrial remote sensing projects for over 20 years, including the development of custom algorithms for unique data such as those obtained by the (then) Russian Vega spacecraft from their close flyby of Comet Halley in 1986. More recently, she has been involved in various projects for geologic mapping of Martian and terrestrial regions, through analyses of imagery from the Imager for Mars Pathfinder, the Mars Exploration Rovers, and from NASA's airborne hyperspectral sensor AVIRIS.