Uniformity and Homogeneity Based Hierachical Clustering *

Peter Bajcsy and Narendra Ahuja

Beckman Institute
University of Illinois at Urbana-Champaign
405 N. Mathews Ave., Urbana, IL 61801
E-mail: peter@stereo.ai.uiuc.edu and ahuja@vision.ai.uiuc.edu

Abstract

This paper presents a clustering algorithm for dot patterns in n-dimensional space. The n-dimensional space often represents a multivariate (n_f-dimensional) function in a n_s-dimensional space ($n_s + n_f = n$). The proposed algorithm decomposes the clustering problem into the two lower dimensional problems. Clustering in n_f-dimensional space is performed to detect the sets of dots in n-dimensional space having similar n_f-variate function values (location based clustering using a homogeneity model). Clustering in n_s-dimensional space is performed to detect the sets of dots in n-dimensional space having similar interneighbor distances (density based clustering with a uniformity model). Clusters in the n-dimensional space are obtained by combining the results in the two subspaces.

1. Introduction

Clustering explores inherent tendency of a dot pattern to form sets of dots (clusters) in multidimensional space. The multidimensional space represents parameters of some phenomenon, for example, image texture may contain overlapping multiple textures having inherent densities (one subspace) with different colors or shapes of texels within each texture (another subspace). This paper presnts a new clustering method that separates the n_s-dimensional spatial (e.g., location and density) and n_f-dimensional intrinsic properties represented by the dot distribution ($n_s + n_f = n$). In this sense it differs from many of the existing methods (single link, complete link, minimum spanning tree, Zahn's clustering, nearest neighbors, Voronoi neighbors, K-means and mode seeking [6, 3, 2, 11, 10, 1, 8, 9, 5, 4]). The clustering problem is decomposed into two lower-dimensional problems. The dot pattern in n-dimensional space is projected onto the two subspaces. The specific choice of subspaces is determined by the application at hand. Clustering is performed in each subspace and the results then combined.

Thus, clustering is viewed as an extension of the problem of segmenting a noisy multivariate multidimensional function. A location uniformity model for clustering is used in n_s-dimensional subspace (modeling uniform sampling) to detect clusters with similar interior distances between dots (density based clustering), and a homogeneity model for clustering is used in n_f-dimensional subspace (modeling constant multivariate function values) to detect clusters with similar locations of dots (location based clustering). Similarity is defined as the Euclidean distance, e.g., between two interior distances or two locations. The two models are used in the corresponding two subspaces and the links and dot locations are clustered using a new method. Overall clustering is carried out by clustering the two dot patterns independently in n_s and n_f dimensional subspaces and then combining the results. Hierarchical organization of clusters is obtained by (1) varying the degrees of uniformity ε and homogeneity δ to create several clusterings and (2) capturing the relationship among the detected clusters as a function of uniformity ε and homogeneity δ. The proposed clustering method can be related to the graph theoretic algorithms.

2. Uniformity and homogeneity based clustering

First, a mathematical framework is established in section 2.1. n-dimensional (nD) points are projected onto the two lower dimensional subspaces giving rise to the n_s-dimensional (n_sD) sample points and n_f-dimensional (n_fD) attribute points. Clustering of sample points is proposed with the uniformity model in section 2.2 (uniformity of sample point locations or homogeneity of interior link distances). Clustering of attribute points is proposed with the homogeneity model in section 2.3 (homogeneity of point locations). A procedure for hierarchical clustering is outlined in section 2.4. The result is exclusive (nonoverlapping

* This research was supported in part by the Advanced Research Projects Agency under grant N00014-93-1-1167 administered by the Office of Naval Research

clusters), intrinsic (no a priori knowledge), agglomerative (grouping points) and a graph based hierarchical classification of a dot pattern.

2.1. Mathematical formulation

An nD dot pattern is defined as a set of points p_i with coordinates $(x_1, x_2, \cdots, x_{n_s}, f_1, f_2, \cdots, f_{n_f})$, which represent a discrete sample point $x_i = (x_1, x_2, \cdots, x_{n_s})$ and a discrete attribute point $f(x_i) = (f_1, f_2, \cdots, f_{n_f})$ in the two n_s and n_f dimensional subspaces. f is defined as a mapping $f : \Re^{n_s} \longrightarrow \Re^{n_f}$ at sample points x_i.

Dissimilarity measure d of two dots p_1 and p_2 is defined by the Euclidean distance of the minimum path between p_1 and p_2 (denoted as link l_{p_1,p_2}), i.e., $d(l_{p_1,p_2}) = \| p_1 - p_2 \|$.

Given sample points x_i, a link is assigned to every possible pair of sample points. All links over given sample points x_i create a complete graph $H = \{l_{x_{i1},x_{i2}} = l_k\}$. Let us suppose that all links from a complete graph H are partitioned into nonoverlapping clusters of links CL_m, where m is the index of a cluster. The uniformity ε of one cluster of links CL_m (denoted as CL_m^ε) is valid if for all links l_k in the cluster CL_m^ε the following is true:

(1) A connected graph $G(CL_m^\varepsilon) \subset H$ is created, i.e., if every sample point is a vertex in the graph then a path exists between any two vertices in the connected graph.

(2) Distances $d(l_k \in CL_m^\varepsilon)$ associated with links l_k vary by no more than ε, i.e., $\mid d(l_1 \in CL_m^\varepsilon)) - d(l_2 \in CL_m^\varepsilon)) \mid \leq \varepsilon$. We can write that all links $l_k \in CL_m^\varepsilon$ must have link distances within an ε wide distance interval $d(l_k) \in [d_{midp}(CL_m^\varepsilon) - \frac{\varepsilon}{2}, d_{midp}(CL_m^\varepsilon) + \frac{\varepsilon}{2}]$, where $d_{midp}(CL_m^\varepsilon)$ is the average value of the maximum and minimum link distances from the connected graph $G(CL_m^\varepsilon)$; $d_{midp}(CL_m^\varepsilon) = \frac{1}{2}(\max\{d(l_k)\} + \min\{d(l_k)\})$ (see Figure 1).

Having the final partition of all links l_k into clusters of links CL_m^ε with ε-uniformity, we can obtain the final partition of sample points x_i into clusters of sample points CS_j^ε with ε-uniformity based on the priority of minimum average link distances within clusters of links (the minimum spanning tree of clusters of links CL_m^ε).

Let us suppose that all attribute points $f(x_i)$ are partitioned into clusters of attribute points CF_j, where j is the index of a cluster. The homogeneity δ of one cluster CF_j (denoted as CF_j^δ) is defined as the maximum distance between any pair of attribute points from the cluster, i.e., $\| f(x_1 \in CF_j^\delta) - f(x_2 \in CF_j^\delta) \| \leq \delta$. We can also write that any attribute point $f(x_i) \in CF_j^\delta$ must have a location within an $n_f D$ sphere $f(x_i) \in Sph(Center = f_{midp}; radius = \frac{\delta}{2})$, where f_{midp} has the coordinates of the middle attribute point from the two attribute points $f(x_q)$ and $f(x_t)$ being the most distant; $\| f(x_q) - f(x_t) \| = \max\{\| f(x_{i1}) - f(x_{i2}) \|\}$ and $f_{midp} = \frac{1}{2}(f(x_q) + f(x_t))$. The

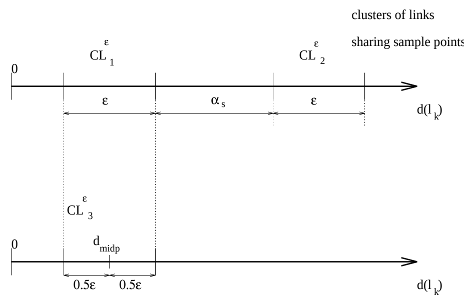

Figure 1. Uniformity and separation of clusters of links.

Uniformity and separation of clusters of links are illustrated on the axis of link distances $d(l_k)$. Clusters of links with ε-uniformity contain links with link distances occupying ε wide interval on the axis of link distances ($CL_1^\varepsilon, CL_2^\varepsilon, CL_3^\varepsilon$). Separation of any pair of clusters of links, which share at least one common sample point x_i by their links (CL_1^ε, CL_2^ε), is defined as $\alpha_s = \min\{\mid d(l_k \in CL_1^\varepsilon) - d(l_k \in CL_2^\varepsilon) \mid\}$. There is no separation defined between clusters of links, which do not share at least one common sample point ($CL_1^\varepsilon, CL_3^\varepsilon$).

δ-homogeneity of a cluster is illustrated in Figure 2.

Definition 1 *ε-uniformity and δ-homogeneity based dot pattern clustering.*

Given the uniformity parameter ε, the homogeneity parameter δ and nD dots $p_i = (x_i, f(x_i))$, (ε, δ) based dot pattern clustering partitions nD dots p_i into a set of clusters $C_t^{\varepsilon,\delta}$ such that the clusters $C_t^{\varepsilon,\delta}$ (t is the index of a cluster) satisfy the following properties:

1. *ε-uniformity of sample points x_i.*
2. *δ-homogeneity of attribute points $f(x_i)$.*
3. *Cluster intersection; $C_{t1}^{\varepsilon,\delta} \cap C_{t2}^{\varepsilon,\delta} = 0$ for all $t1 \neq t2$.*
4. *Cluster union; $\cup C_t^{\varepsilon,\delta} = \cup p_i$.*

2.2. Clustering of sample points x_i

The clustering method for unknown clusters of links having a large separation of link distances with respect to their interior uniformity is proposed first in 2.2.1. Statistical descriptors of clusters are introduced to cope with unknown clusters of links having a small separation of link distances with respect to their interior uniformity in 2.2.2. Descriptors and a sequential decrease of the number of links reduce complexity of the proposed clustering method. The clustering algorithm is provided at the end in 2.2.3.

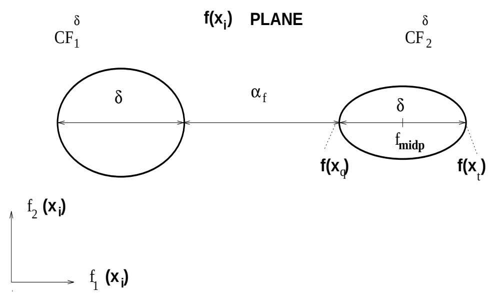

Figure 2. Homogeneity and separation of clusters of 2D attribute points.

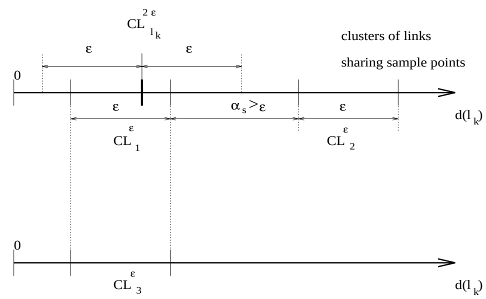

Figure 3. Link distances for CL_m^ε and $CL_{l_k}^{2\varepsilon}$.
An ε-uniform unknown cluster of links CL_m^ε with a large separation of link distances with respect to its interior uniformity $\alpha_s > \varepsilon$ contains the same links as any created 2ε-uniform cluster $CL_{l_k}^{2\varepsilon}$ from a link $l_k \in CL_m^\varepsilon$ (see $CL_{l_k}^{2\varepsilon} = CL_1^\varepsilon$).

All attribute points from one δ-homogeneous cluster are within a sphere having center at f_{midp} and radius $\frac{\delta}{2}$ (see CF_2^δ). Separation α_f of a pair of clusters is defined as the minimum distance between two attribute points each from one cluster; $\alpha_f = \min\{\| (f(x_{i1}) \in CF_1^\delta) - (f(x_{i2}) \in CF_2^\delta) \|\}$.

2.2.1 The clustering method for modelled clusters

Let us suppose that M nonoverlapping ε-uniform clusters of links $\{CL_m^\varepsilon;\ m = 1, ..M\}$ are created from a complete graph $H = \{l_k\}$ over sample points x_i. Let us assume that for all pairs of clusters of links $CL_{m1}^\varepsilon, CL_{m2}^\varepsilon$ sharing at least one sample point x_i by their links, the separation α_s of link distances is more than their interior uniformity ε $\alpha_s > \varepsilon$; (see Figure 3). This scenario represents a modelled dot pattern.

An unknown cluster of links CL_m^ε can be created from any link $l_{k1} \in CL_m^\varepsilon$ by grouping together all links l_k satisfying the inequality $\mid d(l_{k1}) - d(l_k) \mid \leq \varepsilon$. The ε-uniform cluster CL_m^ε is identical with the 2ε-uniform cluster $CL_{l_{k1}}^{2\varepsilon}$ created from the link l_{k1} such that $d_{midp} = d(l_{k1})$; $\mid d_{midp} - d(l_k) \mid \leq \varepsilon$ and d_{midp} was defined in section 2.1.

Starting from individual links l_k,

> clusters of links CL_m^ε can be created by grouping those links l_{k1} and l_{k2} together, which (1) are connected (l_{k1} and l_{k2} share one common point x_i) and (2) lead to identical 2ε-uniform clusters $CL_{l_{k1}}^{2\varepsilon} = CL_{l_{k2}}^{2\varepsilon} = CL_m^\varepsilon$.

Let us order clusters of links $CL_m^\varepsilon = \{l_k\}_{k=1}^{M_m}$ based on their average link distances (first moments $d_{1stm}(l_k \in CL_m^\varepsilon) = \frac{1}{M_m}\sum_{k=1}^{M_m} d(l_k \in CL_m^\varepsilon)$) from the shortest average link distances to the longest average link distances; $d_{1stm}(l_k \in CL_1^\varepsilon) \leq d_{1stm}(l_k \in CL_2^\varepsilon) \leq ... \leq d_{1stm}(l_k \in CL_M^\varepsilon)$. Then clusters of sample points CS_j^ε are uniquely derived from the cluster of links CL_m^ε by using minimum spanning tree of CL_m^ε with the link distances equal to $d_{1stm}(l_k \in CL_m^\varepsilon)$. Thus links $l_k \in CL_1^\varepsilon$ from the ordered set of CL_m^ε will assign labels to sample points x_i first, then links $l_k \in CL_2^\varepsilon$ for the remaining unlabelled sample points, etc.

Proposed clustering of sample points x_i
(1) Create 2ε-uniform clusters of connected links $CL_{l_k}^{2\varepsilon}$ for each link l_k. (2) Compare all pairs of 2ε-uniform clusters $CL_{l_1}^{2\varepsilon}$ and $CL_{l_2}^{2\varepsilon}$, which started from links l_1, l_2 sharing one sample point x_i. (3) Assign links into clusters of links CL_m^ε based on comparisons. (4) Map uniquely already created clusters of links CL_m^ε into clusters of sample points CS_j^ε.

These four steps of the proposed clustering method for a modelled dot pattern ($\alpha_s > \varepsilon$) are applied to any dot pattern having unknown clusters of links with large or small separation with respect to their interior uniformity; $\alpha_s > \varepsilon$ or $\alpha_s \leq \varepsilon$. Performance improvement of the method is achieved by using cluster descriptors for the comparison of 2ε-uniform clusters in the step 2. Exhaustive comparison of two clusters is replaced by comparing two values (descriptors), which leads to a reduction of the computational complexity. Reduction of the computational complexity is improved even more by sequentially decreasing the number of the processed links.

2.2.2 Complexity reduction and performance improvement

A. Descriptors of clusters $CL_{l_k}^{2\varepsilon}$:
A simplified comparison of two 2ε-uniform clusters $CL_{l_k}^{2\varepsilon}$ can be performed using descriptors $D(CL_{l_k}^{2\varepsilon})$ derived from link distances $d(l_i \in CL_{l_k}^{2\varepsilon})$. The descriptor was selected as the mean estimator (first moment) of the correct link distances μ_m within a cluster CL_m^ε; $D(CL^{2\varepsilon}_{l_k \in CL_m^\varepsilon}) = d_{1stm}(l_i \in CL_{l_k}^{2\varepsilon}) = \frac{1}{M_{l_k}}\sum_{i=1}^{M_{l_k}} d(l_i \in CL_{l_k}^{2\varepsilon}) \propto \mu_m$.

Analysis of clustering accuracy using descriptors for

$\alpha_s > \varepsilon$ and $\alpha_s \leq \varepsilon$ led to simplified comparisons of pairs of clusters $CL^{2\varepsilon}_{l_1}$ and $CL^{2\varepsilon}_{l_2}$ in the form of inequalities $\mid D(CL^{2\varepsilon}_{l_1}) - D(CL^{2\varepsilon}_{l_2}) \mid \leq \varepsilon$ rather than equalities $D(CL^{2\varepsilon}_{l_1}) = D(CL^{2\varepsilon}_{l_2})$ if two links l_1, l_2 are going to be assigned to the same cluster (cluster detection approach). Inequalities improve the noise robustness of the method (decrease probability of misclassification).

B. Number of links:

The number of processed links is decreased by merging links in the order of the link distances $d(l_k)$ (from the shortest links to the longest links) into CL^{ε}_m and deriving clusters of sample points CS^{ε}_j immediately. No other links, which contain already merged sample points $x_i \in CS^{\varepsilon}_j$, will be processed afterwards. When the union of all clusters of sample points includes all given sample points ($\cup CS^{\varepsilon}_j = \cup x_i$) then no more links are processed.

2.2.3 Clustering procedure

(1) Calculate link distances $d(l_k)$ for the complete graph H over sample points x_i.
(2) Order $d(l_k)$ from the shortest to the longest.
(3) Create 2ε-uniform clusters of links $CL^{2\varepsilon}_{l_k}$ for each individual link l_k such that $d(l_k) = d_{midp} \leq d(l_1) + \varepsilon$ of the cluster $CL^{2\varepsilon}_{l_k}$.
(4) Calculate descriptors $d_{1stm}(CL^{2\varepsilon}_{l_k})$.
(5) Group together connected pairs of links l_{k1} and l_{k2} into a common cluster of links CL^{ε}_m if $\mid d_{1stm}(l_k \in CL^{2\varepsilon}_{l_{k1}}) - d_{1stm}(l_k \in CL^{2\varepsilon}_{l_{k2}}) \mid \leq \varepsilon$.
(6) Assign those unassigned sample points to clusters CS^{ε}_j, which belong to links creating clusters CL^{ε}_m.
(7) Remove all links from the ordered set, which contain already assigned sample points.
(8) Perform calculations from step (3) for $d(l_1) = d(l_1) + \varepsilon$ until there are unassigned sample points.

2.3. Clustering of attribute points $f(x_i)$

Clustering of attribute points is analogous to the clustering of sample points with replacing links by attribute point locations. The clustering method for unknown clusters having a large separation with respect to their homogeneity $\alpha_f > \delta$ is derived first.

Proposed clustering for attribute points $f(x_i)$

(1) Create 2δ-homogeneous clusters $CF^{2\delta}_{f(x_i)}$ for every attribute point $f(x_i)$. (2) Compare pairs of clusters $CF^{2\delta}_{f(x_i)}$. (3) Assign attribute points $f(x_i)$ to clusters CF^{δ}_j based on comparisons.

Unknown clusters CF^{δ}_j having a small separation with respect to their interior homogeneity $\alpha_f \leq \delta$ are tackled by using descriptors of 2δ-homogeneous clusters in the step 2, which estimate the correct centroid value μ_j of attribute points within a cluster CF^{δ}_j; $D(CF^{2\delta}_{f(x_i) \in CF^{\delta}_j}) = f_{1stm}(f(x_l) \in CF^{2\delta}_{f(x_i)}) = \frac{1}{M_{f(x_i)}} \sum_{l=1}^{M_{f(x_i)}} (f(x_l) \in CF^{2\delta}_{f(x_i)}) \propto \mu_j$. The clustering algorithm is provided next.

Clustering procedure

(1) Create 2δ-homogeneous clusters $CF^{2\delta}_{f(x_i)}$ for each attribute point $f(x_i)$.
(2) Calculate descriptors $f_{1stm}(f \in CF^{2\delta}_{f(x_i)})$.
(3) Group together attribute points $f(x_1)$ and $f(x_2)$ into a common cluster CF^{δ}_j if $\| f_{1stm}(f \in CF^{2\delta}_{f(x_1)}) - f_{1stm}(f \in CF^{2\delta}_{f(x_2)}) \| \leq \delta$.

2.4. Hierarchical clustering

A hierarchy of clusters of dots $C^{\varepsilon,\delta}_t$ is defined as a combination of the hierarchy of clusters of links CL^{ε}_m and the hierarchy of clusters of attribute points CF^{δ}_j for varying uniformity and homogeneity parameters (ε, δ). Clusters of links or attribute points are organized hierarchically by allowing the clusters only to grow for increasing parameter.

The hierarchy of clusters of links CL^{ε}_m and clusters of attribute points CF^{δ}_j is guaranteed by modifying link distances and attribute points within created clusters at each parameter value (ε, δ) to the first moments of link distances $d_{1stm}(l_k \in CL^{\varepsilon}_m)$ and attribute points $f_{1stm}(f(x_i) \in CF^{\delta}_j)$ for performed agglomerative clustering.

3. Performance evaluation

Theoretical and experimental evaluations are focused on (1) clustering accuracy and (2) performance for real applications. Clustering accuracy is tested for (1) synthetic (modelled) dot patterns and (2) standard test dot patterns (80x, IRIS), which were used by several other researches to illustrate properties of clusters (80x is used in [6] and IRIS in [7, 6]). Experimental results are compared with four other clustering methods (single link, complete link, FORGY and CLUSTER). In addition, experimental performance of both proposed methods is tested using dot patterns from [11] to compare clustering results with the two related methods, the Zahn's clustering [11] (ε-uniformity method) and the centroid method [6] (δ-homogeneity method). Experimental results for real applications are conducted for dot patterns obtained from botanical analysis of plants and image texture analysis and synthesis.

From all aforementioned experiments we show only one result of image texture detection to demonstrate an exceptional property of the proposed clustering method with respect to all other known clustering techniques.

A synthetic image with three overlapping textures having different densities of circles (texels) contains subset of circles having different color. Created gray scale image is shown in Figure 4 (top). The image was segmented and

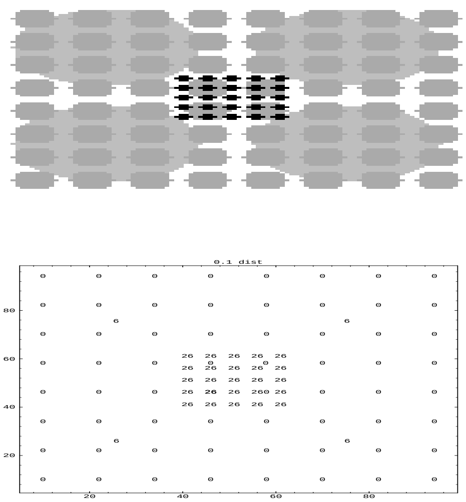

Figure 4. Image with overlapping transparent textures.

Top - original synthetic image. Bottom - ε-uniformity clustering of a dot pattern obtained by taking centroid locations of detected regions in the segmented synthetic image; clusters are denoted by numerical labels (6 for large circles, 0 for middle size circles and 26 for small circles) and are detected at the uniformity value $\varepsilon = 0.1$.

2D dots were obtained as centroids of detected regions. ε-uniformity clustering of a dot pattern provided separation of the three different textures shown in Figure 4 (bottom). The texture separation was successful despite partial occlusion of circles an therefore irregularity of dot locations obtained from segmented regions. Within each detected texture a δ-homogeneity clustering grouped together circles with similar color.

4. Conclusions

We have presented a new hierarchical clustering method that decomposes the n-dimensional clustering problem into two lower dimensional problems. Decomposing allows us to apply two different models to n-dimensional dots, the ε-uniformity model in n_s-dimensional subspace and the δ-homogeneity model in n_f-dimensional subspace ($n_s + n_f = n$). A new ε-uniformity method for density based clustering is proposed for $n_s D$ spatial points. The use of density allows us to detect multiple interleaved noisy clusters that represent projections of different clusters on transparent surfaces into a single image. δ-homogeneity clustering is proposed for $n_f D$ attribute points to detect intrinsic property represented by dots.

References

[1] N. Ahuja. Dot pattern processing using voronoi neighborhoods. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 4(3):336–343, May 1982.

[2] N. Ahuja and B. J. Schachter. *Pattern Models*. John Wiley and sons inc., USA, 1983.

[3] E. by K. S. Fu. *Digital Pattern Recognition*. Springer-Verlag, New York, 1976.

[4] B. S. Everitt. *Cluster analysis*. Edward Arnold, a division of Hodder and Stoughton, London, 1993.

[5] H. Hanaizumi, S. Chino, and S. Fujimura. A binary division algorithm for clustering remotely sensed multispectral images. *IEEE Transaction on Instrumentation and Measurement*, 44(3):759–763, June 1995.

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

[7] P. H. Sneath and R. R. Sokal. *Numerical Taxonomy*. W. H. Freeman and company, San Francisco, 1973.

[8] M. Tuceryan and A. Jain. Texture segmentation using voronoi polygons. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(2):211–216, 1990.

[9] Y. Wong and E. C. Posner. A new clustering algorithm applicable to multiscale and polarimetric sar images. *IEEE Transaction on Geoscience and Remote Sensing*, 31(3):634–644, May 1993.

[10] B. Yu and B. Yuan. A global optimum clustering algorithm. *Engineering applications of artificial inteligence*, 8(2):223–227, April 1995.

[11] C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transaction on Computers*, C-20(1):68–86, January 1971.