Dominant Sets and Pairwise Clustering

Marcello Pelillo

Department of Computer Science

Ca' Foscari University, Venice

pelillo@dsi.unive.it

http://www.dsi.unive.it/~pelillo

[joint work with Massimiliano Pavan]

Talk's Outline

- Dominant sets and their characterization
- Evolutionary game dynamics for clustering
- Experiments on intensity/color/texture image segmentation
- Extension of the framework to hierarchical clustering
- Experiments on the (hierarchical) organization of an image database

Talk's Outline

- **Dominant sets and their characterization**
- Evolutionary game dynamics for clustering
- Experiments on intensity/color/texture image segmentation
- Extension of the framework to hierarchical clustering
- Experiments on the (hierarchical) organization of an image database

Notations

We represent the data to be clustered as an undirected edge-weighted graph with no self-loops $G = (V, E, w)$, where $V = \{1, \ldots, n\}$ is the vertex set, $E \subseteq V \times V$ is the edge set, and $w : E \rightarrow \mathbb{R}_+^*$ is the (positive) weight function.

We represent the graph G with the corresponding weighted adjacency (or similarity) matrix, which is the $n \times n$ symmetric matrix $A = (a_{ij})$ defined as:

$$a_{ij} = \begin{cases} w(i,j)\,, & \text{if } (i,j) \in E \\ 0\,, & \text{otherwise.} \end{cases}$$

Basic Definitions

Let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$. The (average) weighted degree of i w.r.t. S is defined as:

$$\text{awdeg}_S(i) = \frac{1}{|S|} \sum_{j \in S} a_{ij} \ .$$

Moreover, if $j \notin S$ we define:

$$\phi_S(i,j) = a_{ij} - \text{awdeg}_S(i) \ .$$

Intuitively, $\phi_S(i,j)$ measures the similarity between nodes j and i, with respect to the average similarity between node i and its neighbors in S.

Note that $\phi_S(i,j)$ can be either positive or negative.

Assigning Node Weights / 1

Let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$. The weight of i w.r.t. S is

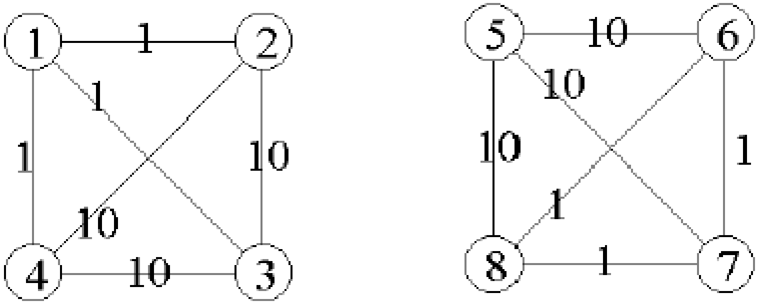
$$\mathrm{w}_S(i) = \begin{cases} 1, & \text{if } |S| = 1 \\ \sum_{j \in S \setminus \{i\}} \phi_{S \setminus \{i\}}(j, i)\, \mathrm{w}_{S \setminus \{i\}}(j), & \text{otherwise.} \end{cases}$$

Moreover, the total weight of S is defined to be:

$$\mathrm{W}(S) = \sum_{i \in S} \mathrm{w}_S(i) .$$

Assigning Node Weights / 2

Intuitively, $w_S(i)$ gives us a measure of the overall similarity between vertex i and the vertices of $S \setminus \{i\}$ with respect to the overall similarity among the vertices in $S \setminus \{i\}$.

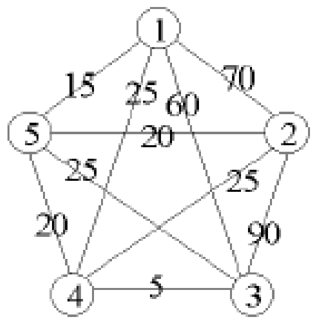


$w_{\{1,2,3,4\}}(1) < 0$ and $w_{\{5,6,7,8\}}(5) > 0$.

Dominant Sets

A non-empty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$, is said to be dominant if:

1. $w_S(i) > 0$, for all $i \in S$ (internal homogeneity)
2. $w_{S \cup \{i\}}(i) < 0$, for all $i \notin S$ (external inhomogeneity)



The set $\{1, 2, 3\}$ is dominant.

Dominant sets $\equiv$ clusters

For 0/1 matrices: dominant sets $\equiv$ (strictly) maximal cliques

From Dominant Sets to Local Optima (and Back) / 1

Given an edge-weighted graph $G = (V, E, w)$ and its weighted adjacency matrix A, consider the following Standard Quadratic Program (StQP):

$$
\begin{array}{ll}
\text{maximize} & f(\mathbf{x}) = \mathbf{x}' A \mathbf{x} \\
\text{subject to} & \mathbf{x} \in \Delta
\end{array}
$$

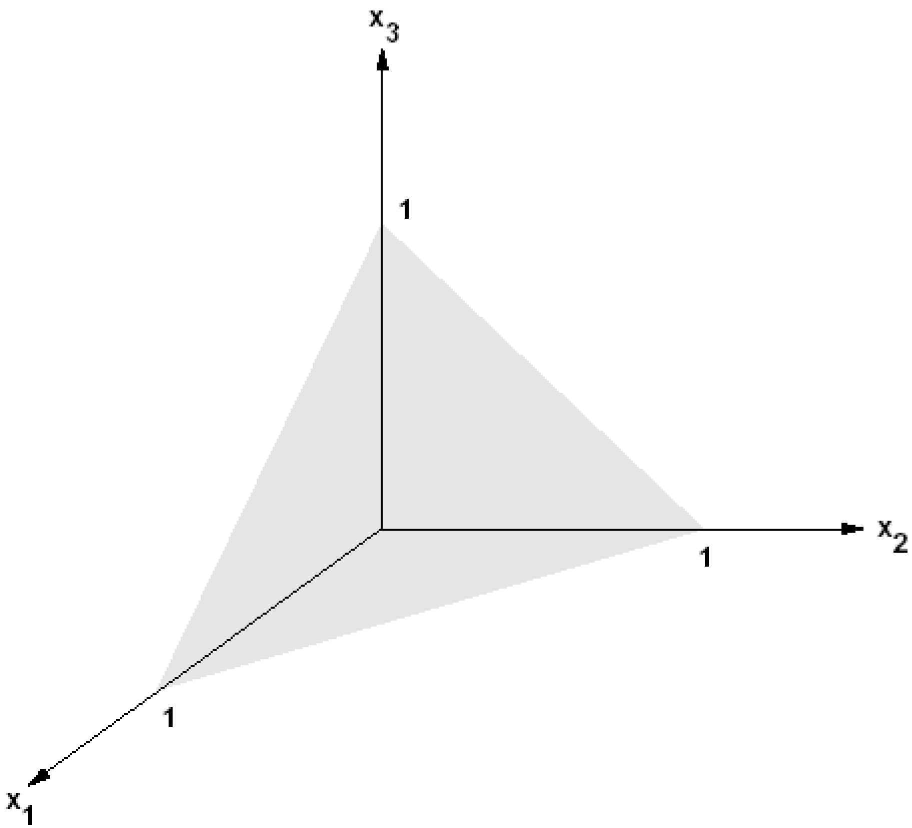
where

$$\Delta = \left\{ \mathbf{x} \in \mathbf{R}^n : \mathbf{e}'\mathbf{x} = 1 \text{ and } x_i \geq 0 \ \forall i \in V \right\}$$

is the standard simplex of $\mathbf{R}^n$ and $\mathbf{e} = (1, 1, \cdots, 1)'$.

Note. Other approaches to clustering lead to similar quadratic optimization problems (e.g., Sarkar and Boyer, 1998).

The Standard Simplex



From Dominant Sets to Local Optima (and Back) / 2

Theorem *If S is a dominant subset of vertices, then its weighted characteristics vector $\mathbf{x}^S$, defined as*

$$x_i^S = \begin{cases} \frac{\mathrm{w}_S(i)}{\mathrm{W}(S)}, & \text{if } i \in S \\ 0\,, & \text{otherwise} \end{cases}$$

is a strict local maximizer of f in Δ.

Conversely, if $\mathbf{x}^*$ is a strict local maximizer of f in Δ then its support

$$\sigma = \sigma(\mathbf{x}^*) \doteq \{i \in V \ : \ x_i^* \neq 0\}$$

is a dominant set, provided that $\mathrm{w}_{\sigma \cup \{i\}}(i) \neq 0$ for all $i \notin \sigma$.

Talk's Outline

- Dominant sets and their characterization
- **Evolutionary game dynamics for clustering**
- Experiments on intensity/color/texture image segmentation
- Extension of the framework to hierarchical clustering
- Experiments on the (hierarchical) organization of an image database

Replicator Equations

Developed in evolutionary game theory to model the evolution of behavior in animal conflicts (Hofbauer & Sigmund, 1998).

Let $W = (w_{ij})$ be a non-negative real-valued $n \times n$ matrix.

Continuous-time version:

$$\frac{d}{dt}x_i(t) = x_i(t)\left[(W\mathbf{x}(t))_i - \mathbf{x}(t)'W\mathbf{x}(t)\right]$$

Discrete-time version:

$$x_i(t+1) = x_i(t)\frac{(W\mathbf{x}(t))_i}{\mathbf{x}(t)'W\mathbf{x}(t)}$$

Δ is invariant under both dynamics, and they have the same stationary points.

The Fundamental Theorem of Natural Selection

If $W = W'$, then the function

$$F(\mathbf{x}) = \mathbf{x}'W\mathbf{x}$$

is strictly increasing along any non-constant trajectory of both continuous-time and discrete-time replicator dynamics.

In other words, $\forall t \geq 0$:

$$\frac{d}{dt}F(\mathbf{x}(t)) > 0$$

for the continuous-time dynamics, and

$$F(\mathbf{x}(t+1)) > F(\mathbf{x}(t))$$

for the discrete-time dynamics, unless $\mathbf{x}(t)$ is a stationary point.

Grouping by Replicator Equations

Let A denote the weighted adjacency matrix of the similarity graph.

Let

$$W = A \quad (= W' \geq 0) \ .$$

The replicator systems, starting from an arbitrary initial state, will eventually converge to a maximizer of the function $f(\mathbf{x}) = \mathbf{x}' A \mathbf{x}$, over the simplex.

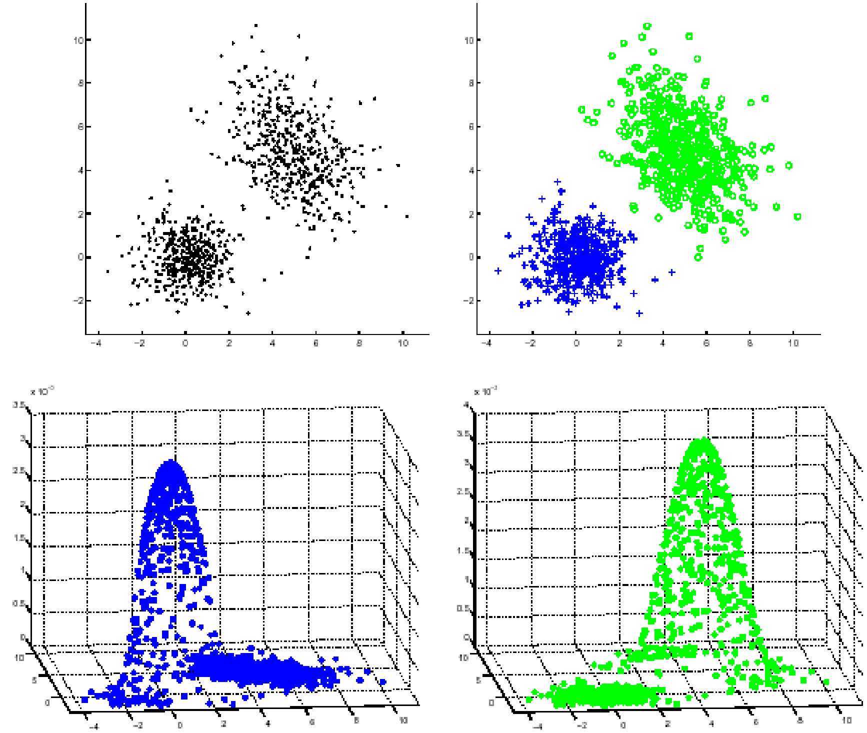
This will correspond to a dominant set in the graph, and hence to a cluster of vertices.

MATLAB Code for Replicator Dynamics

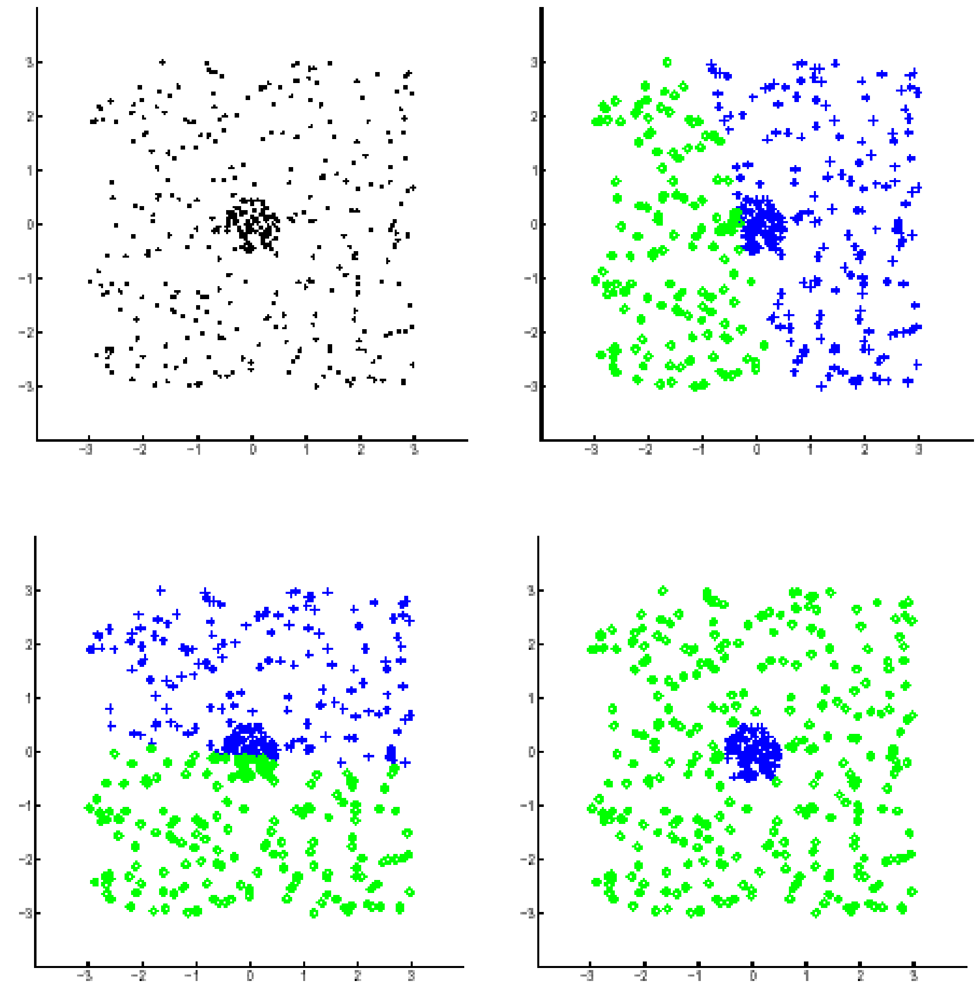
```
while true
    x = x.*(A*x);
    x = x./sum(x);
end
```

Characteristic Vectors

Note. The components of the weighted characteristic vectors give us a measure of the participation of the corresponding vertices in the cluster, while the value of the objective function provides a measure of the cohesiveness of the cluster (*cfr.* Sarkar and Boyer, 1998).



Separating Structure from Clutter



Talk's Outline

- Dominant sets and their characterization
- Evolutionary game dynamics for clustering
- **Experiments on intensity/color/texture image segmentation**
- Extension of the framework to hierarchical clustering
- Experiments on the (hierarchical) organization of an image database

Image Segmentation

An image is represented as an edge-weighted undirected graph, where vertices correspond to individual pixels and the edge-weights reflect the "similarity" between pairs of vertices.

Our clustering algorithm basically consists of iteratively finding a dominant set in the graph using replicator dynamics and then removing it from the graph, until all vertices have been clustered.

In our experiments, we used the discrete-time replicator equations. The process was started from the simplex barycenter and stopped after a few iterations.

On average, the algorithm took only a few seconds to converge, on a machine equipped with a 750 MHz Intel Pentium III.

Experimental Setup

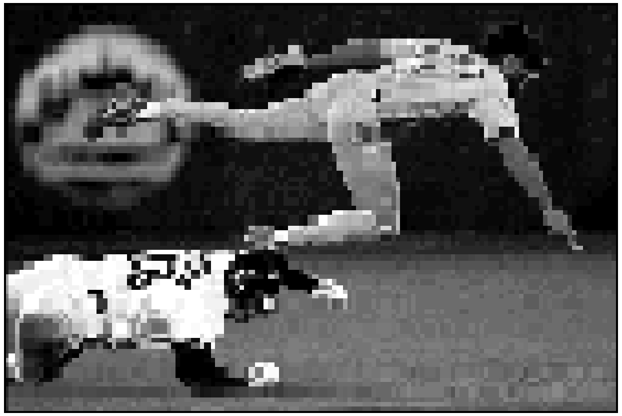
The similarity between pixels i and j was measured by:

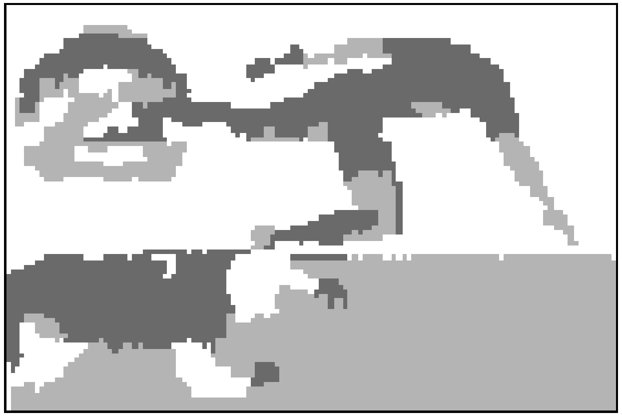
$$w(i,j) = \exp\left(\frac{-\|\mathbf{F}(i) - \mathbf{F}(j)\|_2^2}{\sigma^2}\right)$$

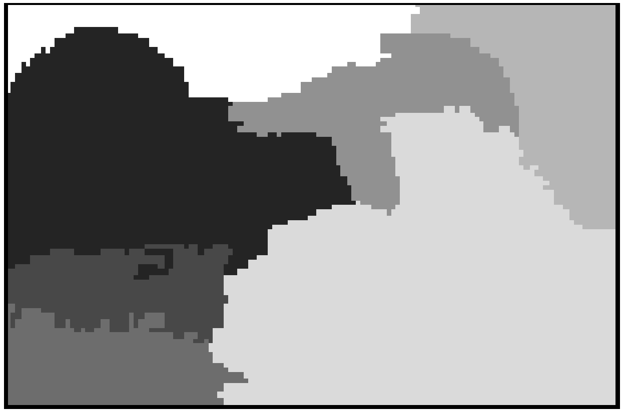
where σ is a positive real number which affects the decreasing rate of w, and:

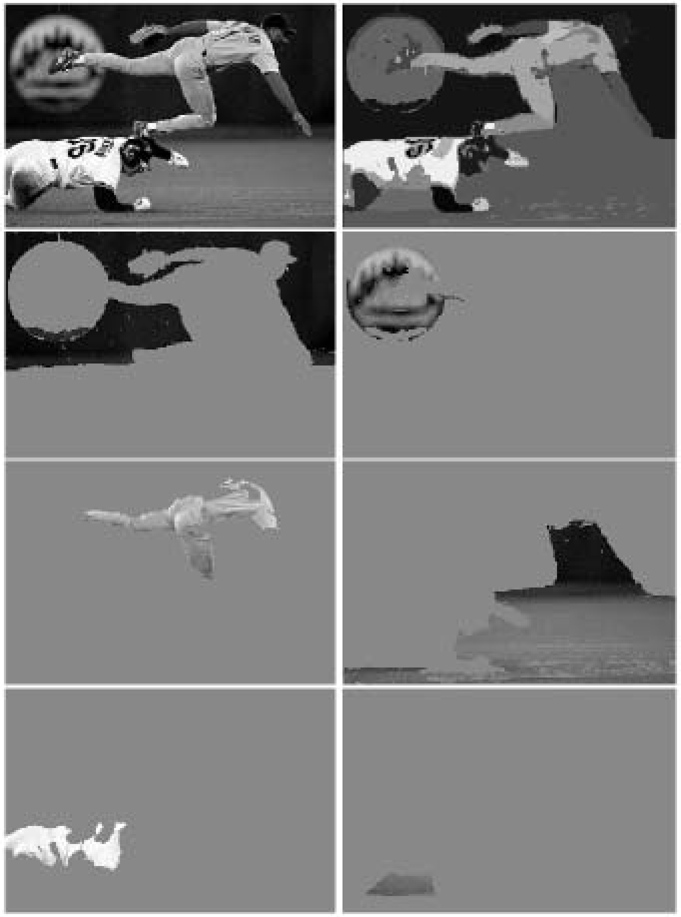
- $\mathbf{F}(i) \equiv$ (normalized) intensity of pixel i, for intensity segmentation
- $\mathbf{F}(i) = [v, vs\sin(h), vs\cos(h)](i)$, where h, s, v are the HSV values of pixel i, for color segmentation
- $\mathbf{F}(i) = [|I * f_1|, \ldots, |I * f_k|](i)$ is a vector based on texture information at pixel i, the f_i being DOOG filters at various scales and orientations, for texture segmentation

Intensity Segmentation Results / 1

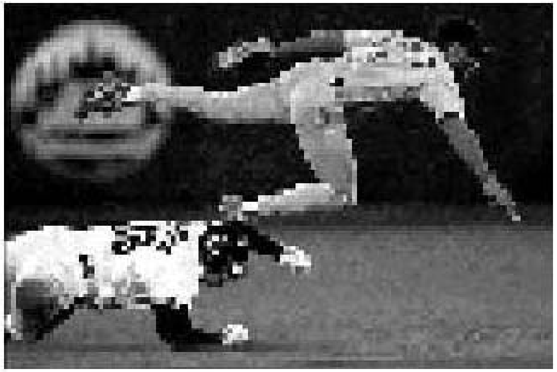


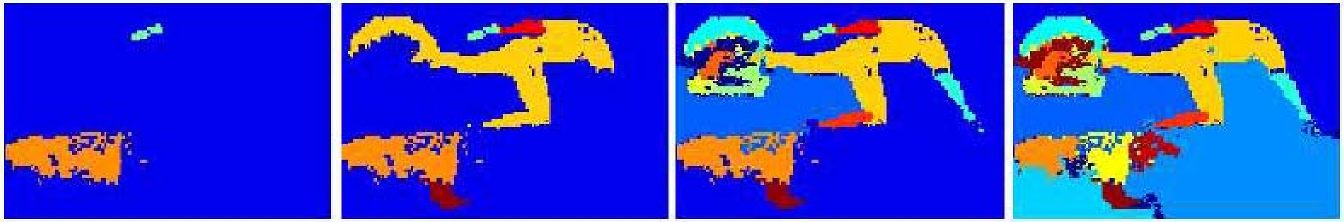






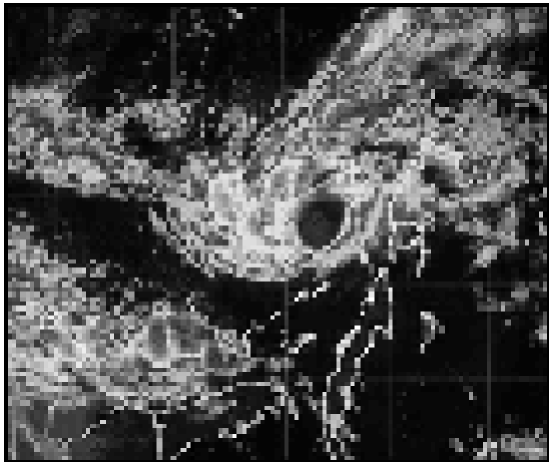
Felzenszwalb and Huttenlocher (2003).

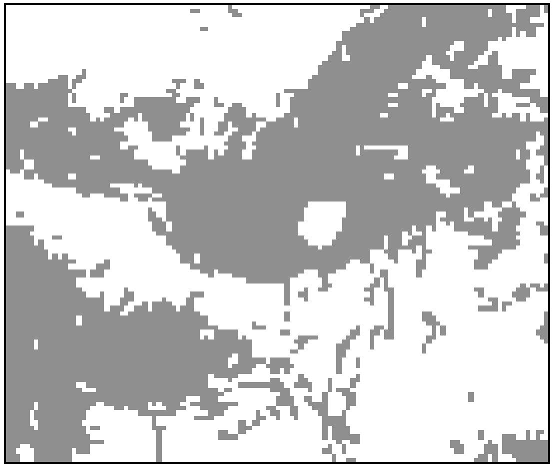


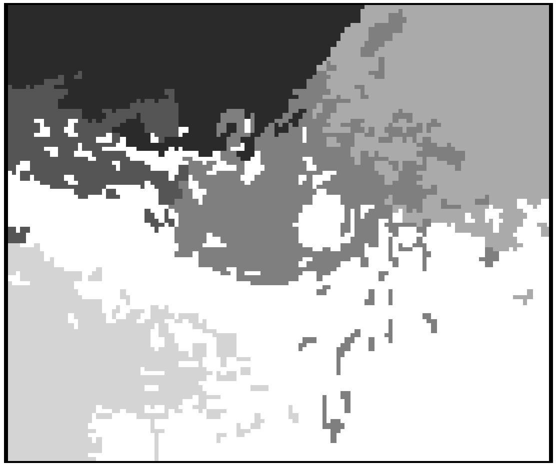


Gdalyahu, Weinshall, and Werman (2001).

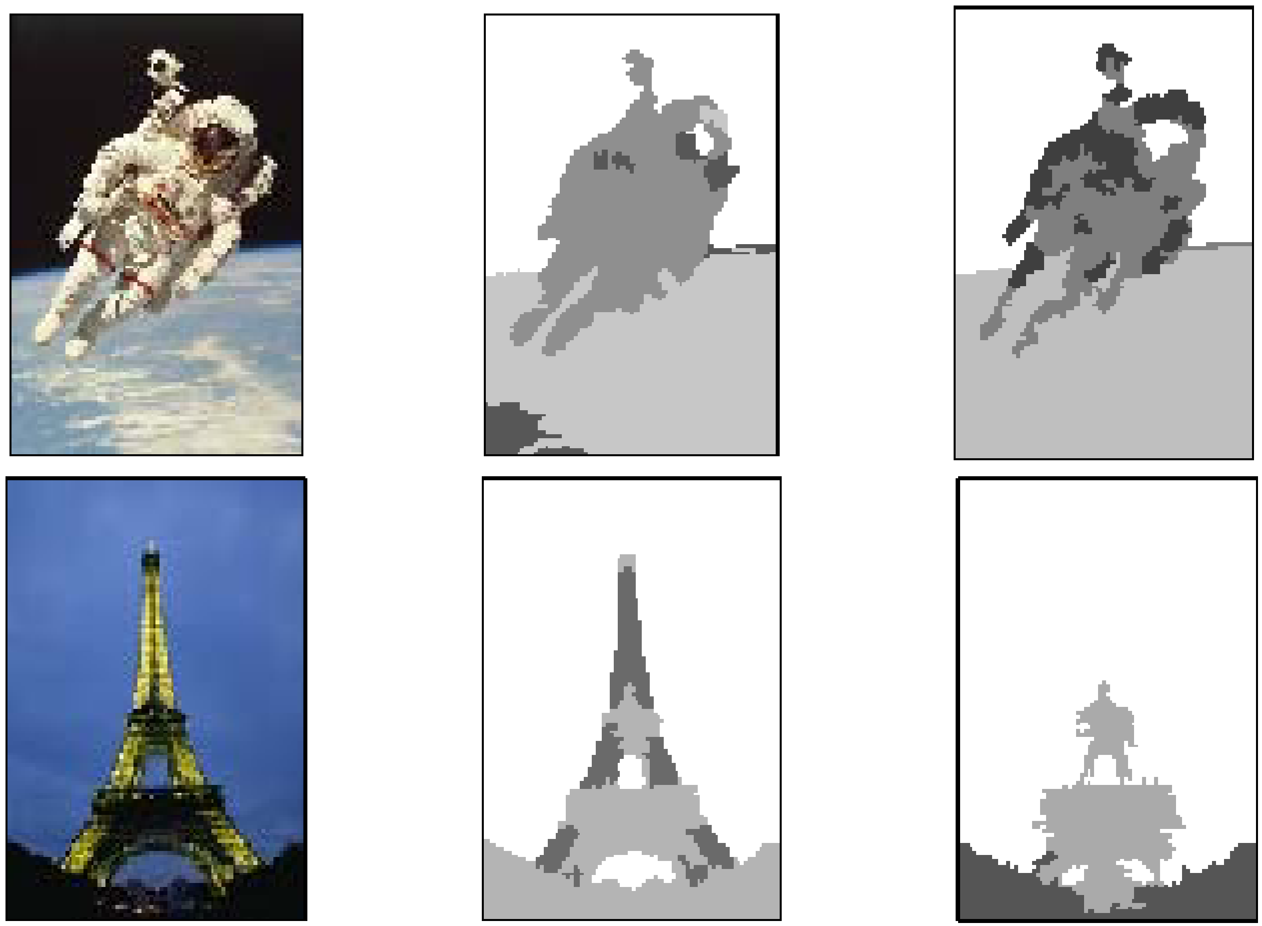
Intensity Segmentation Results / 2



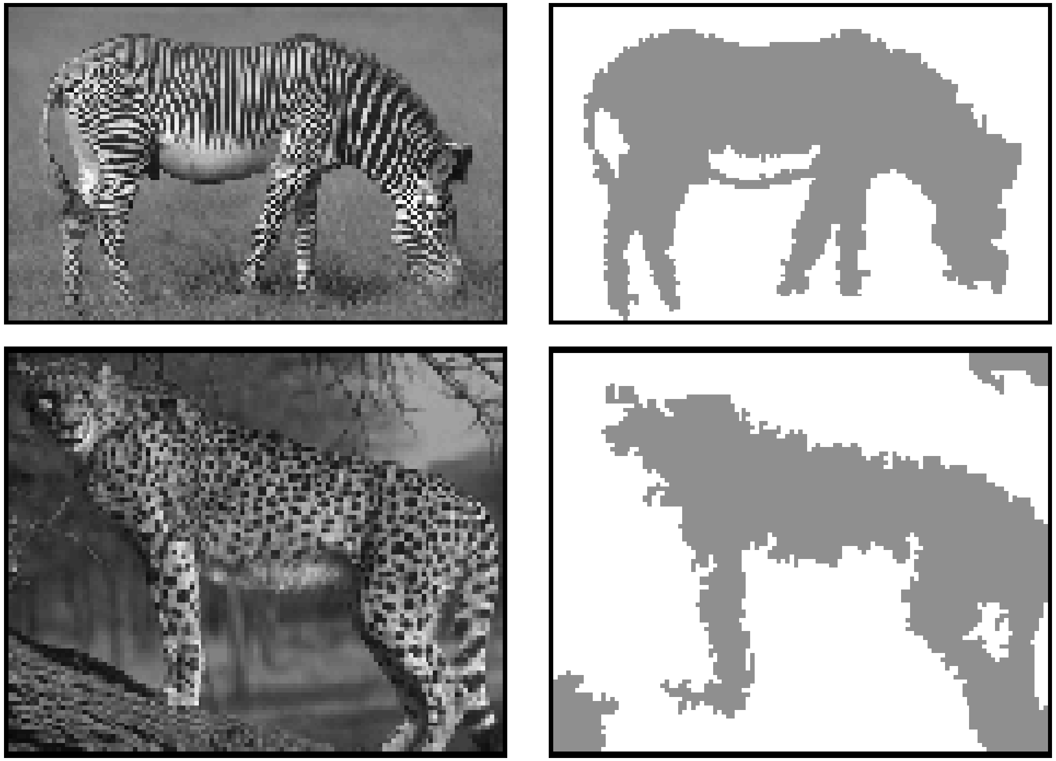




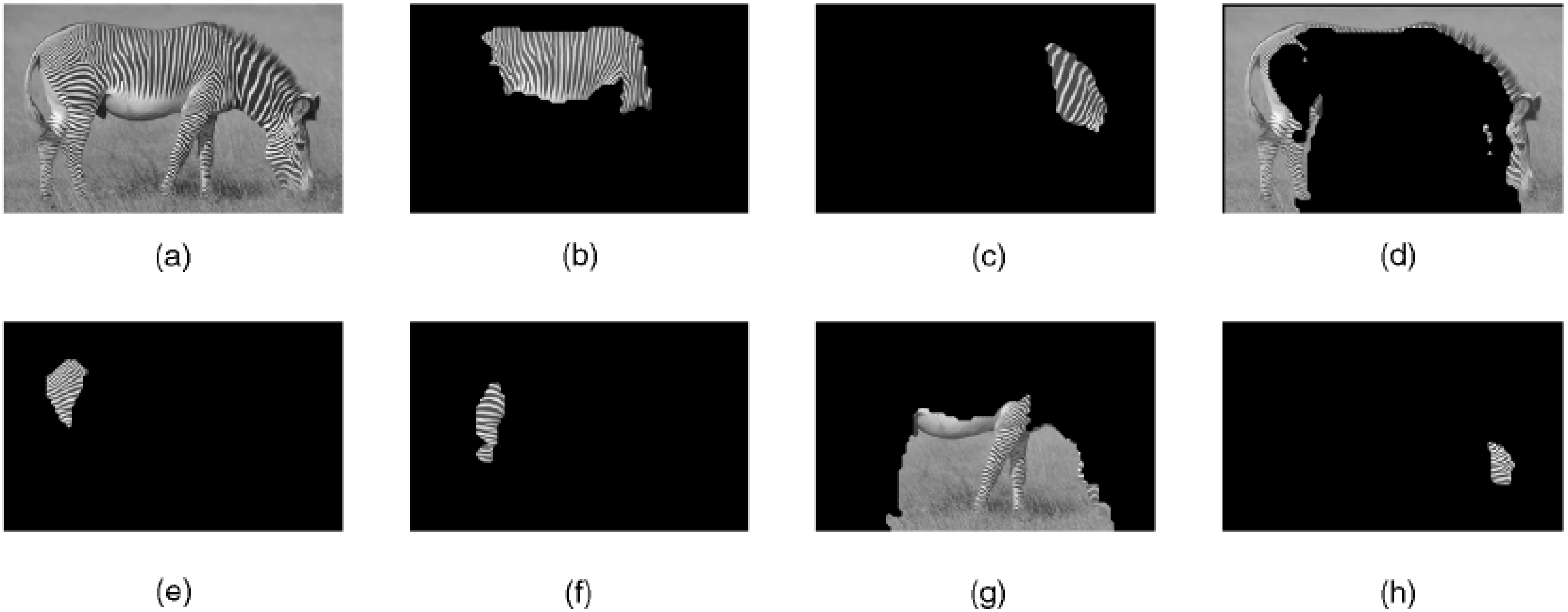
Color Segmentation Results



Texture Segmentation Results



Ncut Results



Dealing with Large Data Sets

We address the problem of grouping *out-of-sample* (i.e., unseen) examples after the clustering process has taken place.

This may serve to:

1. substantially reduce the computational burden associated to the processing of very large data sets, by extrapolating the complete grouping solution from a small number of samples,

2. deal with dynamic situations whereby data sets need to be updated continually.

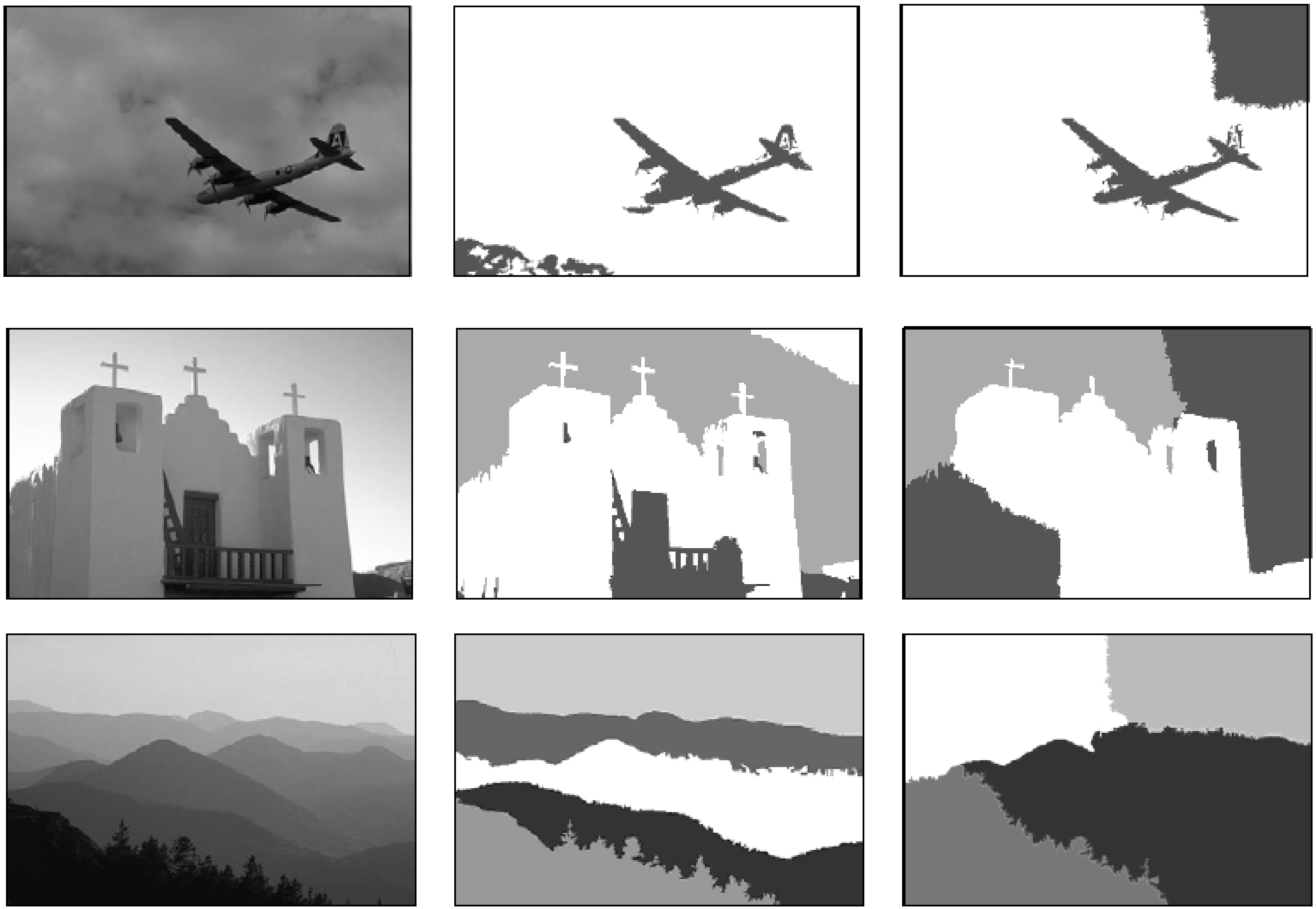
Grouping Out-of-Sample Data

Recall that the sign of $w_{S \cup \{i\}}(i)$ provides an indication as to whether i is tightly or loosely coupled with the vertices in S.

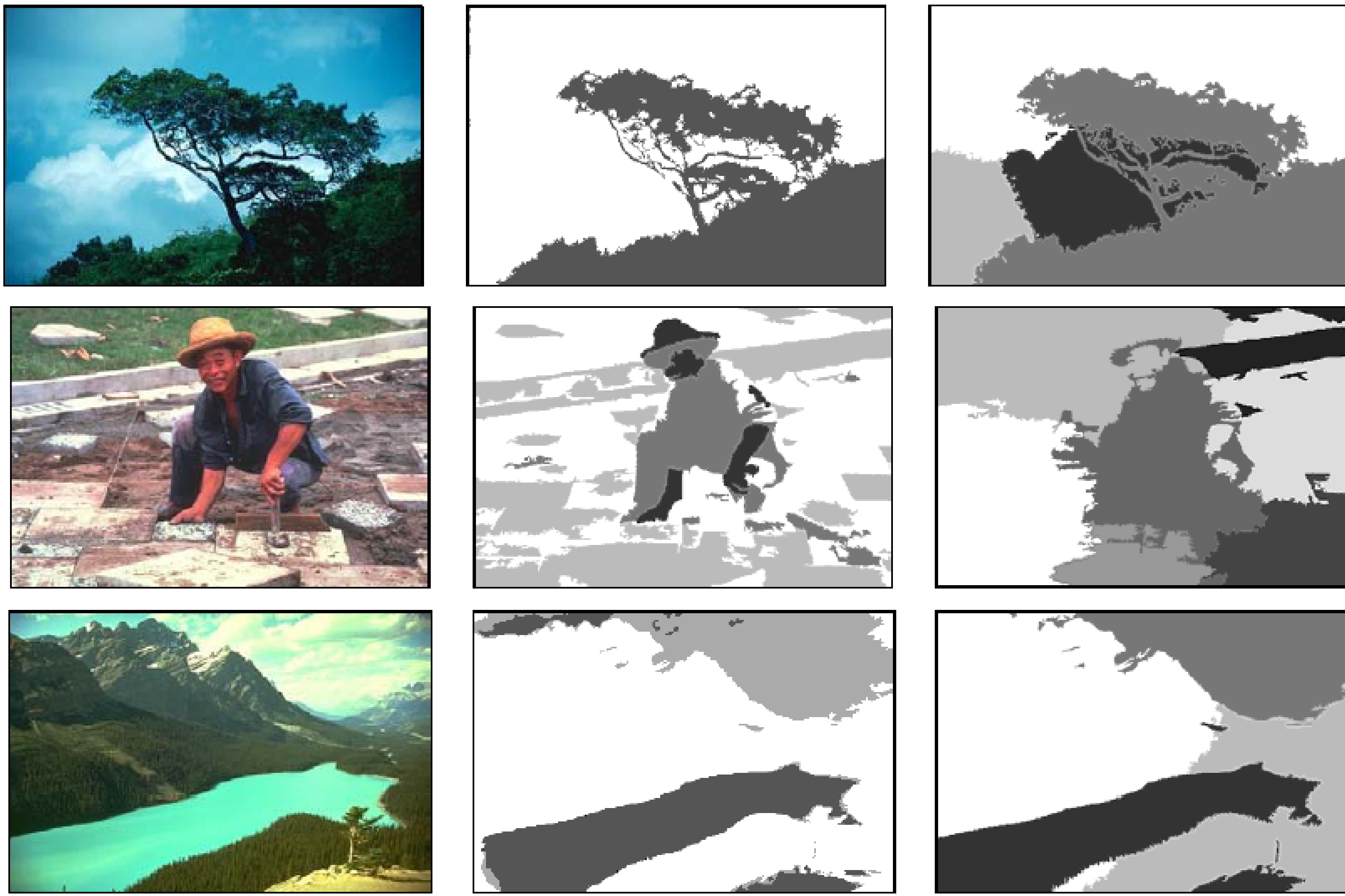
Accordingly, we use the following rule for predicting cluster membership of unseen data i:

if $w_{S \cup \{i\}}(i) > 0$, then assign vertex i to cluster S .

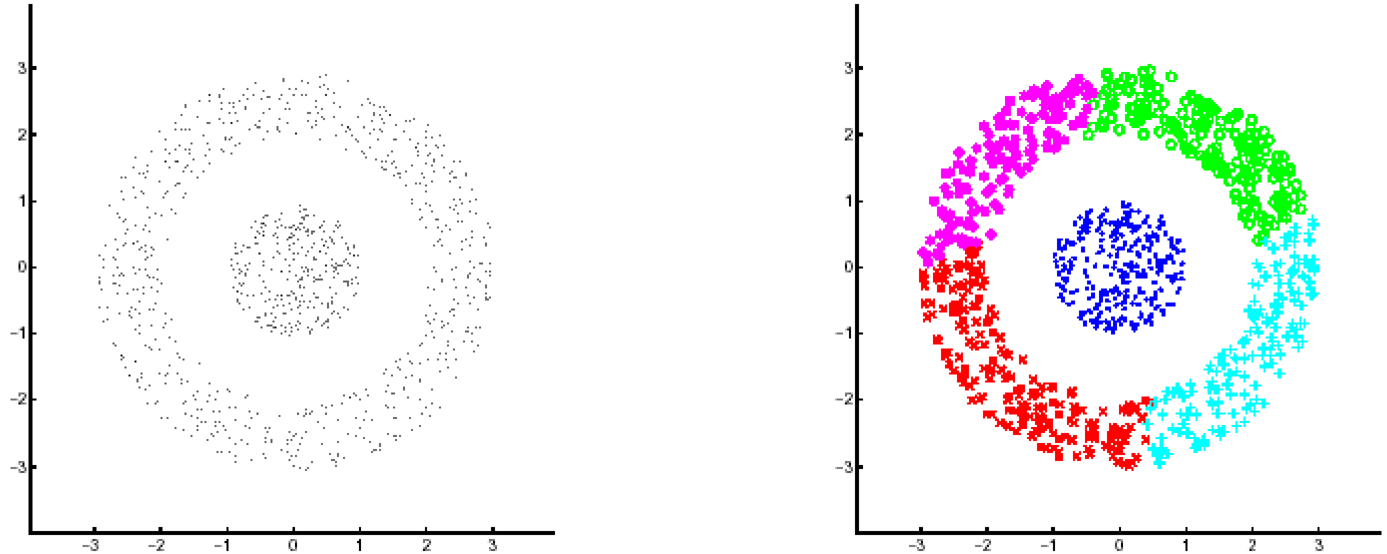
Results on Berkeley Database Images (321 x 481)



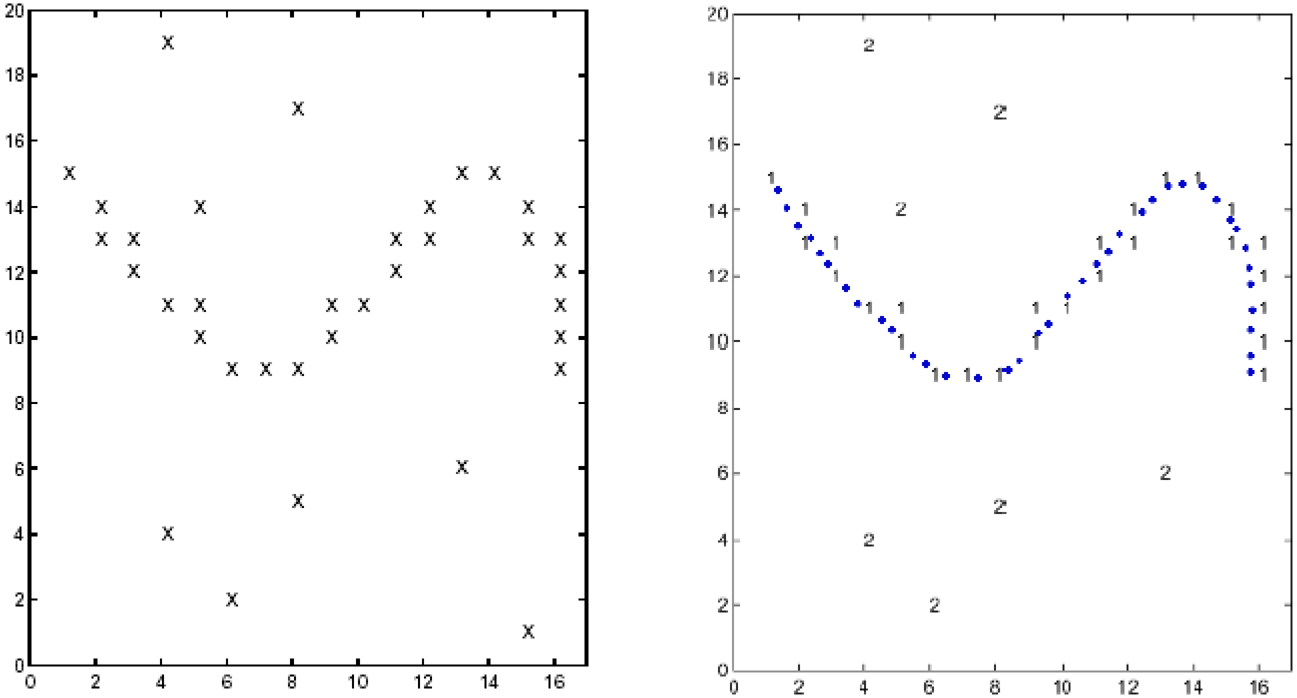
Results on Berkeley Database Images (321 x 481)



Capturing Elongated Structures / 1



Capturing Elongated Structures / 2



“Closing” the Similarity Graph

Basic idea: Trasform the original similarity graph G into a “closed” version thereof (G_{closed}), whereby edge-weights take into account chained (path-based) structures.

Unweighted (0/1) case:

$$G_{closed} = \text{Transitive Closure of } G$$

Note: G_{closed} can be obtained from:

$$A + A^2 + \ldots + A^n$$

Weighted Closure of *G*

Observation: When G is weighted, the ij-entry of A^k represents the sum of the total weights on the paths of length k between vertices i and j.

Hence, our choice is:

$$A_{closed} = A + A^2 + \ldots + A^n$$

Example: Without Closure ($\sigma = 2$)

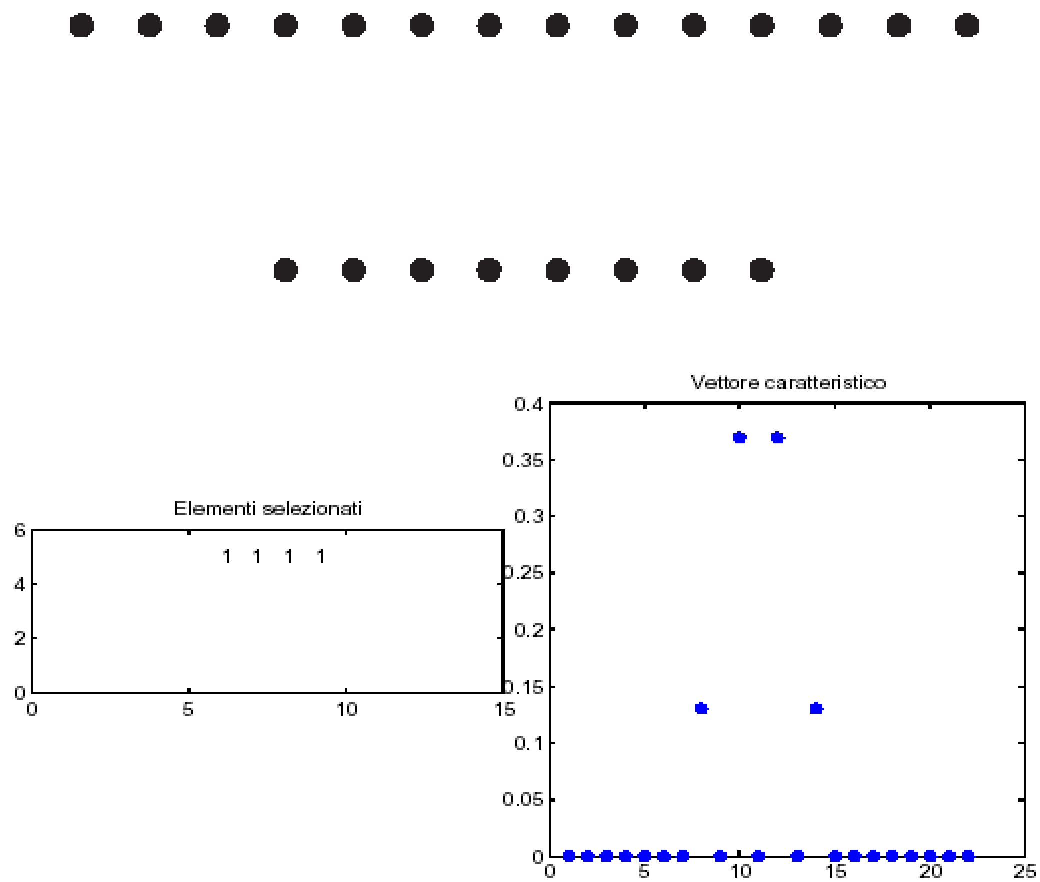
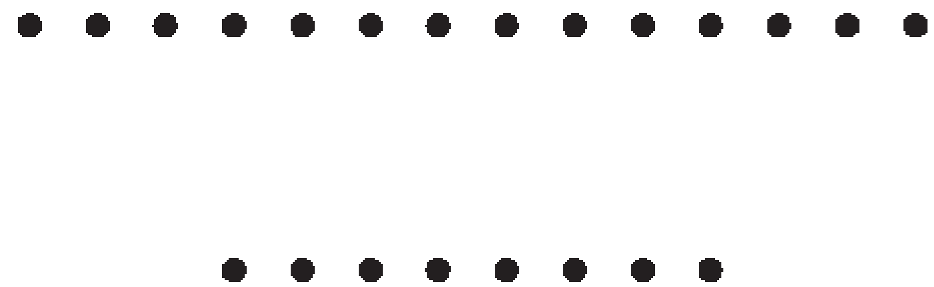


Figura 4.11: Cluster senza chiusura: $\sigma = 2$

Example: Without Closure (σ = 4)



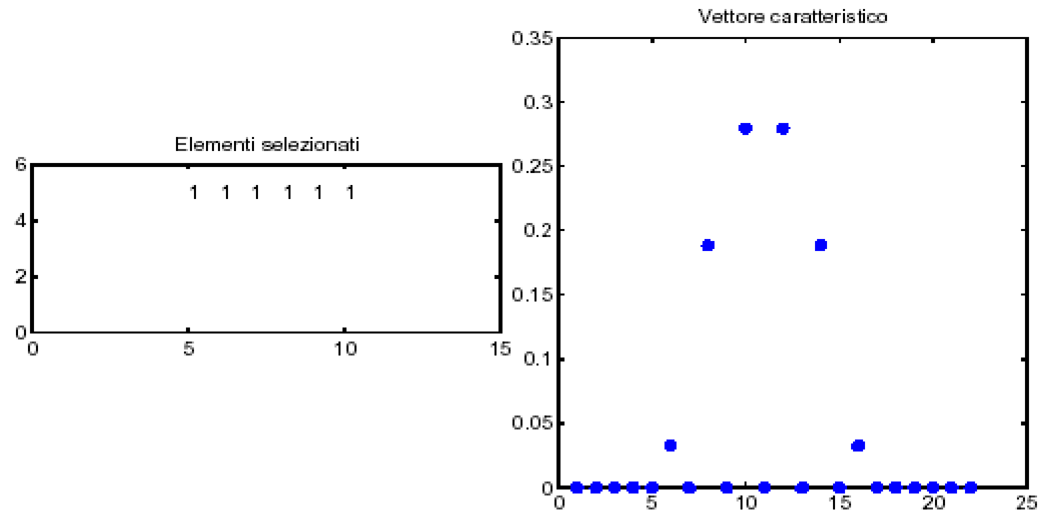
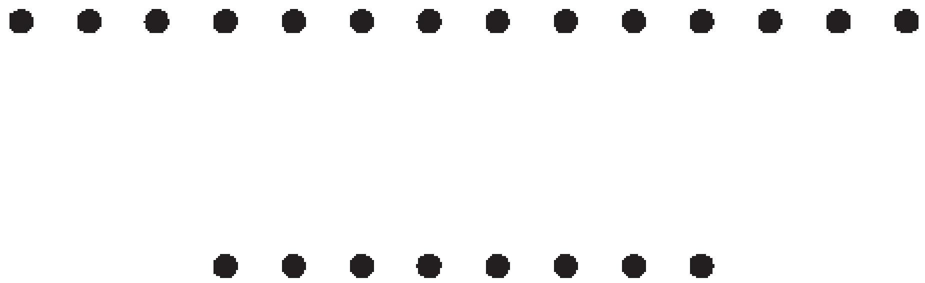


Figura 4.12: Cluster senza chiusura: $\sigma = 4$

Example: Without Closure (σ = 8)



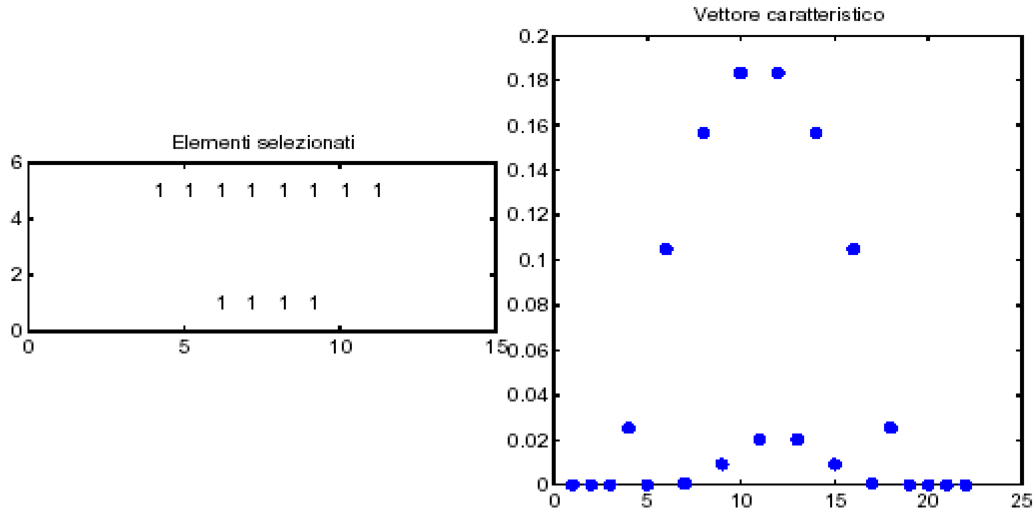


Figura 4.13: Cluster senza chiusura: $\sigma = 8$

Example: With Closure (σ = 0.5)

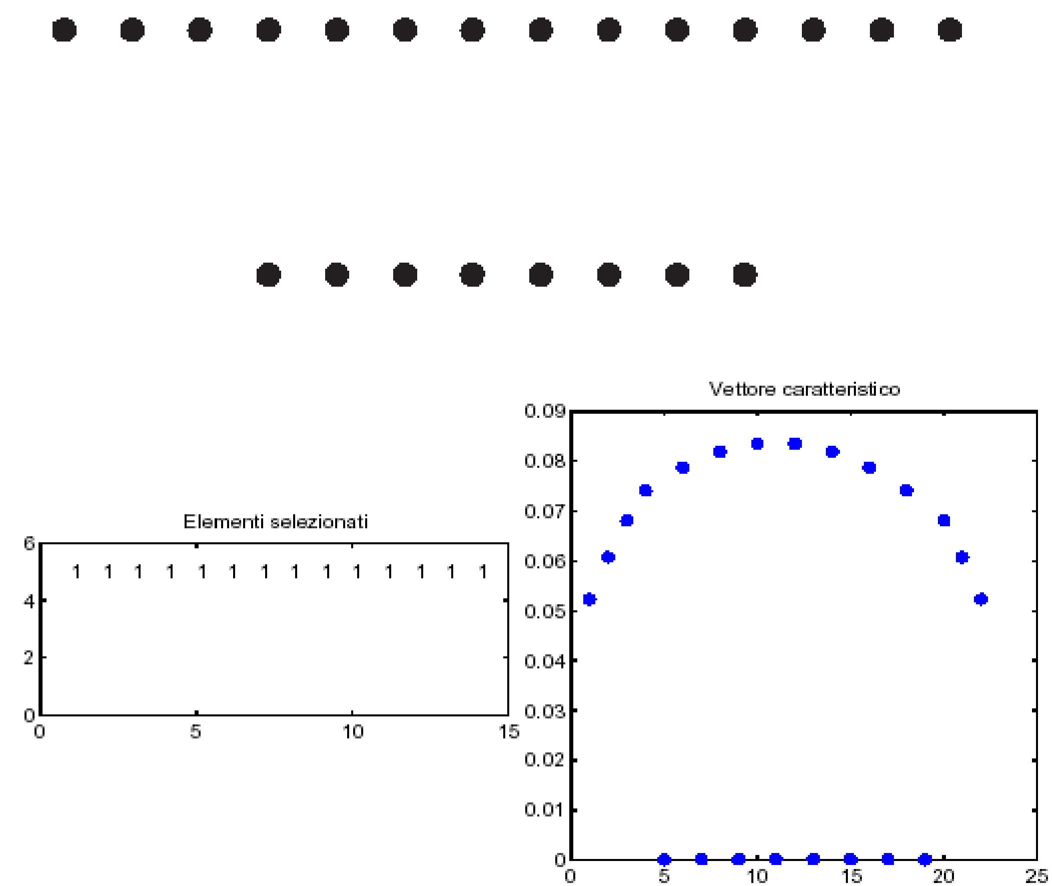
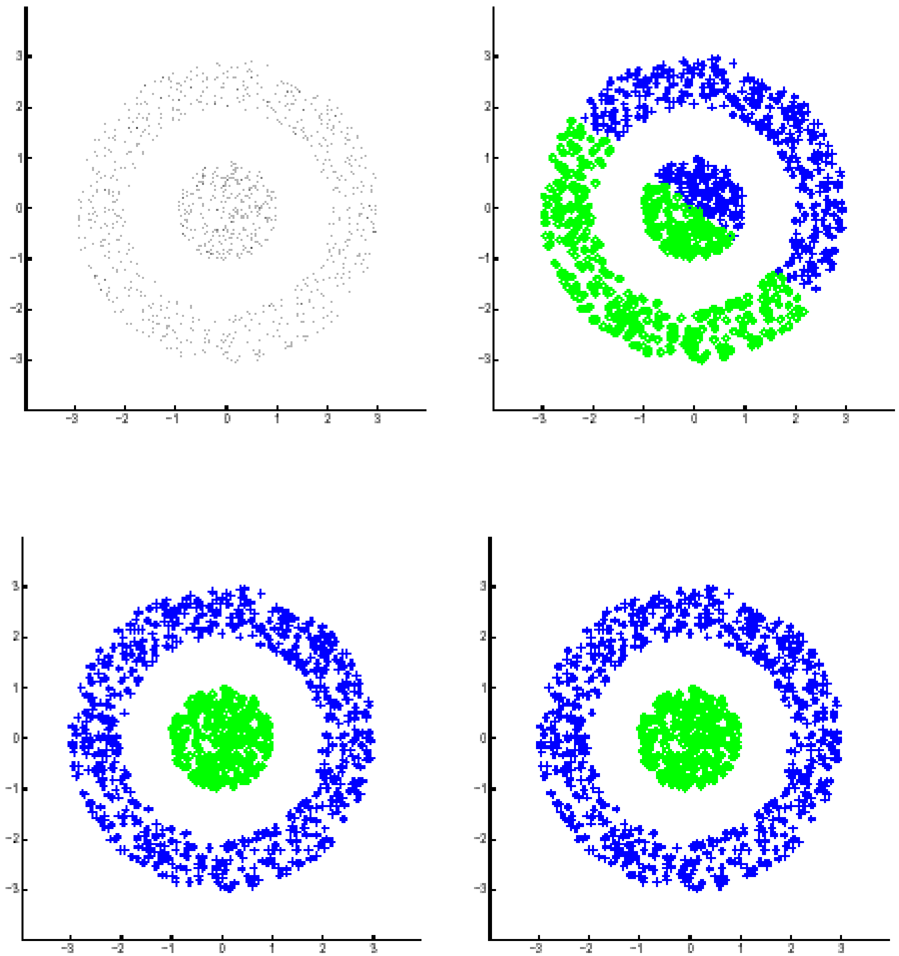


Figura 4.14: Cluster mediante chiusura: $\sigma = 0,5$



Grouping Experiments

The elements to be grouped are edgels.

We used Herault/Horaud (1993) similarities, which combine the following four terms:

1. Co-circularity
2. Smoothness
3. Proximity
4. Contrast

Comparison with Mean-Field Annealing (MFA).

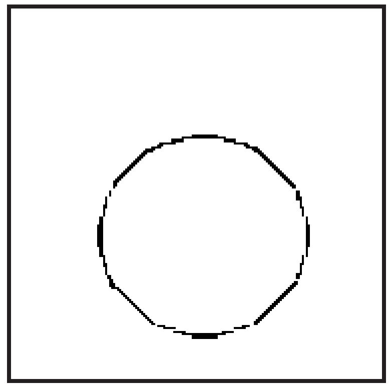


Immagine originale
204 edge

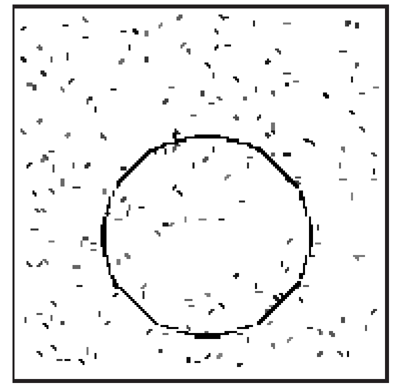
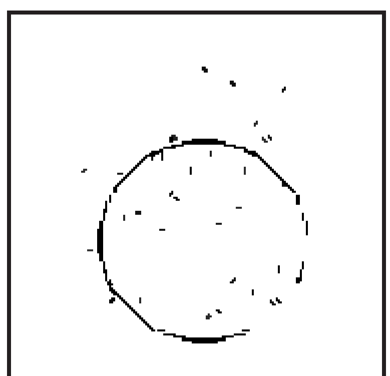
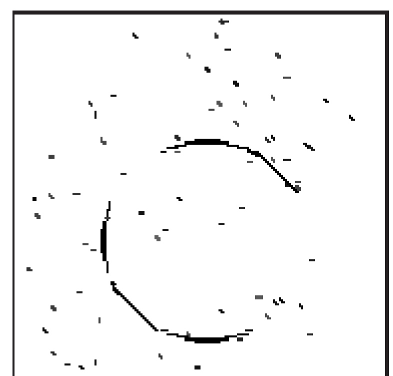


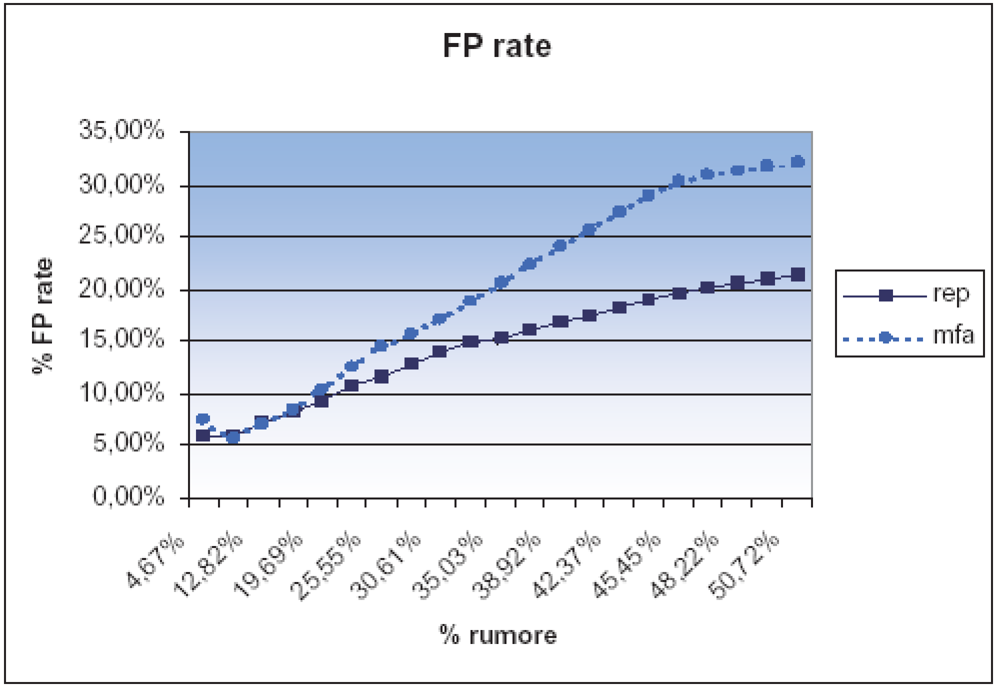
Immagine con rumore al 50%



Insiemi dominanti
FP rate: 16,67%



Mean Field Annealing
FP rate: 34, 31%



FP rate
35,00%
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
% FP rate
4,67%
12,82%
19,69%
25,55%
30,61%
35,03%
38,92%
42,37%
45,45%
48,22%
50,72%
% rumore
rep
mfa

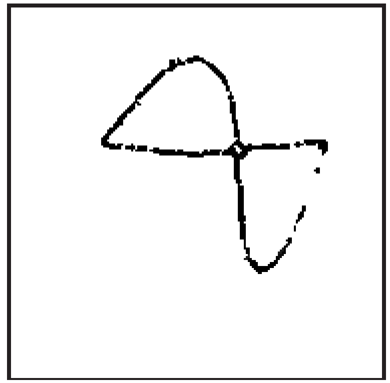


Immagine originale
278 edge

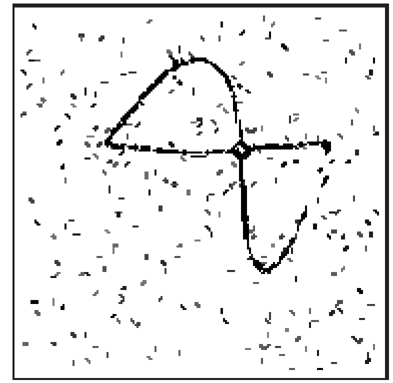
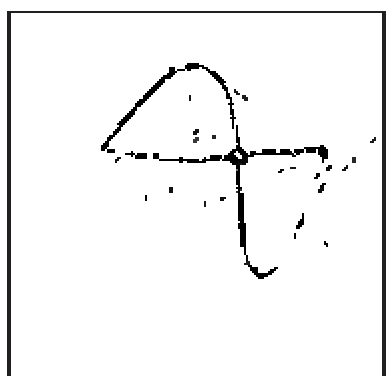
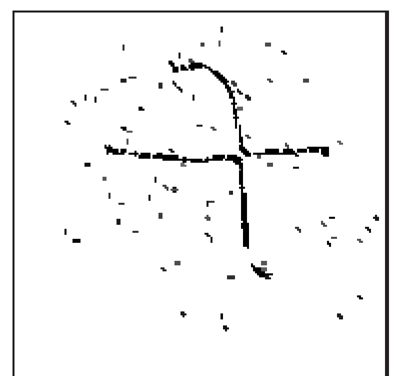


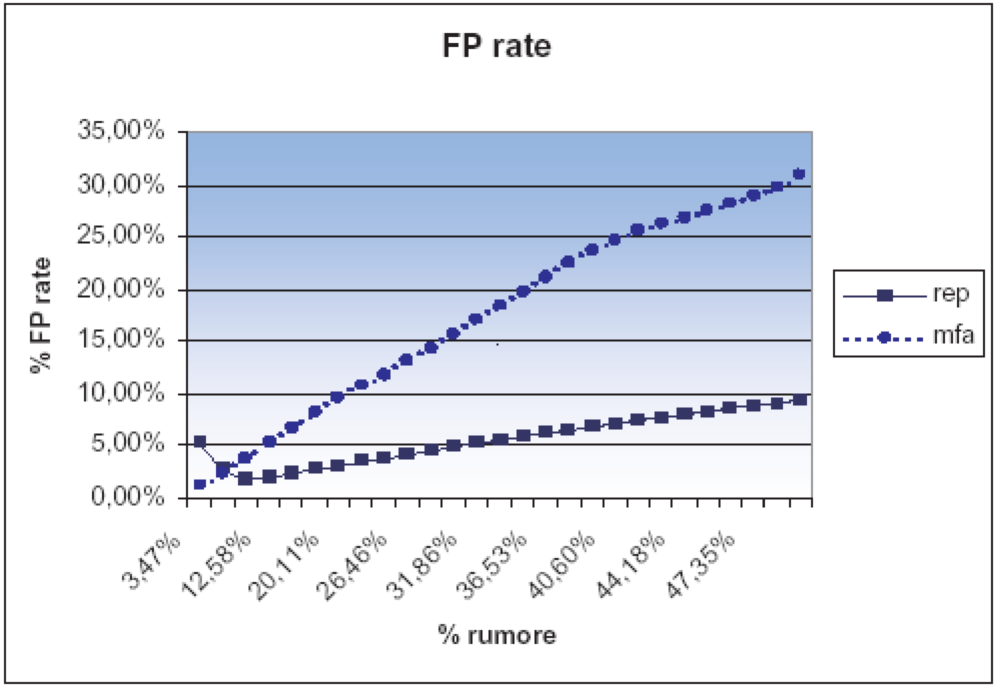
Immagine con rumore al 50%



Insiemi dominanti
FP rate: 8,99%



Mean Field Annealing
FP rate: 29,5%


FP rate
35,00%
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
% FP rate
3,47%
12,58%
20,11%
26,46%
31,86%
36,53%
40,60%
44,18%
47,35%
% rumore
rep
mfa

Talk's Outline

- Dominant sets and their characterization
- Evolutionary game dynamics for clustering
- Experiments on intensity/color/texture image segmentation
- **Extension of the framework to hierarchical clustering**
- Experiments on the (hierarchical) organization of an image database

Building a Hierarchy: A Family of Quadratic Programs

Consider the following family of StQP's:

$$\begin{array}{ll} \text{maximize} & f_\alpha(\mathbf{x}) = \mathbf{x}'(A - \alpha I)\mathbf{x} \\ \text{subject to} & \mathbf{x} \in \Delta \end{array}$$

where $\alpha \geq 0$ is a parameter and I is the identity matrix.

The objective function f_α consists of:

- a data term ($\mathbf{x}'A\mathbf{x}$) which favors solutions with high internal coherency
- a regularization term ($-\alpha\mathbf{x}'\mathbf{x}$) which acts as an entropic factor: it is concave and, on the simplex Δ, it is maximized at the barycenter and it attains its minimum value at the vertices of Δ

An Observation

The solutions of the StQP remain the same if the matrix $A - \alpha I$ is replaced with $A - \alpha I + \kappa \mathbf{e}\mathbf{e}'$, where κ is an arbitrary constant, since

$$\mathbf{x}'(A - \alpha I + \kappa \mathbf{e}\mathbf{e}')\mathbf{x} = \mathbf{x}'(A - \alpha I)\mathbf{x} + \kappa$$

for all $\mathbf{x} \in \Delta$.

In particular, if $\kappa = \alpha$ the resulting matrix is nonnegative and has a null diagonal.

Hence all (strict) solutions of the StQP correspond to dominant sets for the scaled similarity matrix $A + \alpha(\mathbf{e}\mathbf{e}' - I)$ having the off-diagonal entries equal to $a_{ij} + \alpha$.

Bounds for the Regularization Parameter / 1

When α is large enough the regularization term $(-\alpha \mathbf{x}'\mathbf{x})$ dominates, and the only solution of the StQP is in the interior of Δ: this corresponds to a unique large cluster which comprises all the data points.

Proposition *If*

$$\alpha > \lambda_{\max}(A)$$

then f_α is a strictly concave function in $\mathbb{R}^n$, and the only solution $\mathbf{x}$ of the StQP belongs to the interior of Δ, i.e., $\sigma(\mathbf{x}) = V$.

Bounds for the Regularization Parameter / 2

Given a subset of vertices $S \subseteq V$, the face of Δ corresponding to S is defined as:

$$\Delta_S = \{\mathbf{x} \in \Delta \ : \ \sigma(\mathbf{x}) \subseteq S\}$$

and its relative interior is:

$$\text{int}(\Delta_S) = \{\mathbf{x} \in \Delta \ : \ \sigma(\mathbf{x}) = S\} \ .$$

Theorem *Let $S \subset V$ be a proper subset of vertices ($S \neq V$), and let A_S denote the submatrix of A formed by the rows and columns indexed by the elements of S. If*

$$\alpha > \lambda_{\max}(A_S)$$

then there is no point $\mathbf{x} \in \text{int}(\Delta_S)$ that is a local maximizer of f_α in Δ.

Bounds for the Regularization Parameter / 3

Suppose for simplicity that $a_{ij} \leq 1$ for all $i, j \in V$, i.e.

$$0 \leq A \leq \mathbf{e}\mathbf{e}^T - I \ .$$

For any $S \subseteq V$ we get:

$$\lambda_{\max}(A_S) \leq \lambda_{\max}(\mathbf{e}\mathbf{e}^T - I) = |S| - 1$$

Hence, if we want to avoid clusters of size $|S| \leq m < |V|$ we could let

$$\alpha > m - 1$$

In so doing, no face Δ_S with $|S| \leq m$ will contain solutions of the StQP, in other words:

at this scale *all* clusters will have more than m data points

The Landscape of f_α

Key observation: For any fixed α, the energy landscape of f_α is populated by two kinds of solutions:

- solutions which correspond to dominant sets for the original matrix A
- solutions which do not correspond to any dominant set for the original matrix A, although they are dominant for the scaled matrix $A + \alpha(\mathbf{ee}' - I)$

The latter represent large subsets of points that are not sufficiently coherent to be dominant with respect to A, and hence they should be split.

Sketch of the Hierarchical Clustering Algorithm

Basic idea: start with a sufficiently large α and adaptively decrease it during the clustering process:

1) let α be a large positive value (e.g., $\alpha > |V| - 1$)

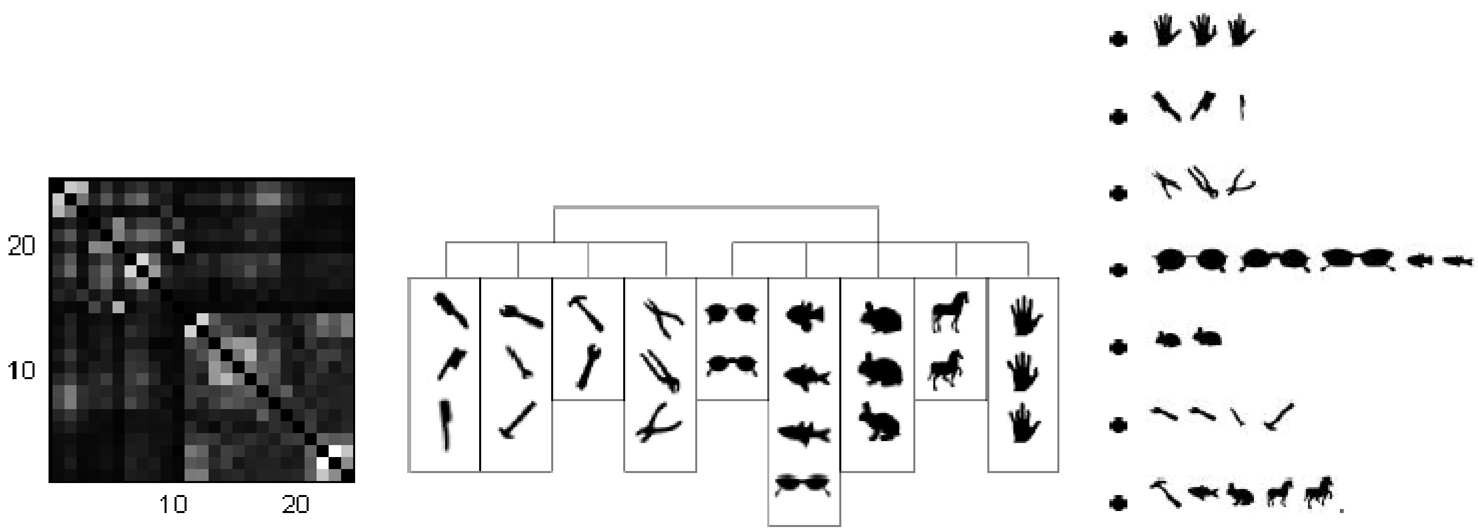
2) find a partition of the data into α-*clusters*

3) for all the α-*clusters* that are not 0-*clusters* recursively repeat step 2) with decreased α

Pseudo-code of the Algorithm

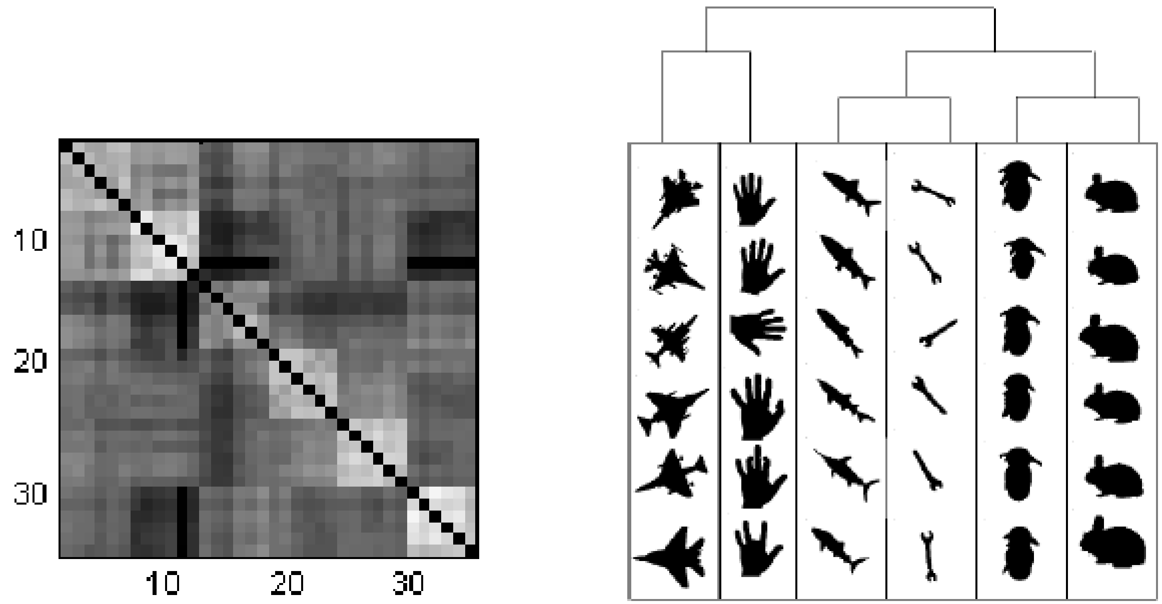
```
Algorithm HIER_CLUSTERING( V, A )
begin
  if V is dominant then return V
  let α > |V| − 1
  repeat
    decrease α
    if α < 0 then α ← 0
    V_1, ..., V_k ← SPLIT( V, A, α )
  until k > 1
  return ∪_{i=1}^{k} { HIER_CLUSTERING( V_i, A_{V_i} ) }
end
```

Luo and Hancock's Similarities (CVPR'01)



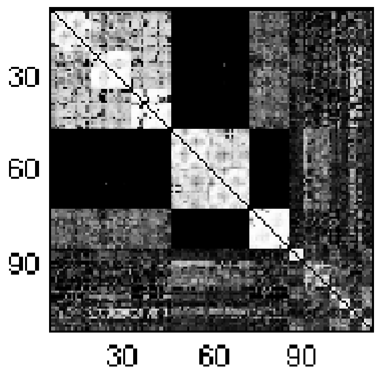
Left: Similarity matrix used in the experiment. Middle: Hierarchy produced by our algorithm. Right: (Flat) partition produced by Luo and Hancock.

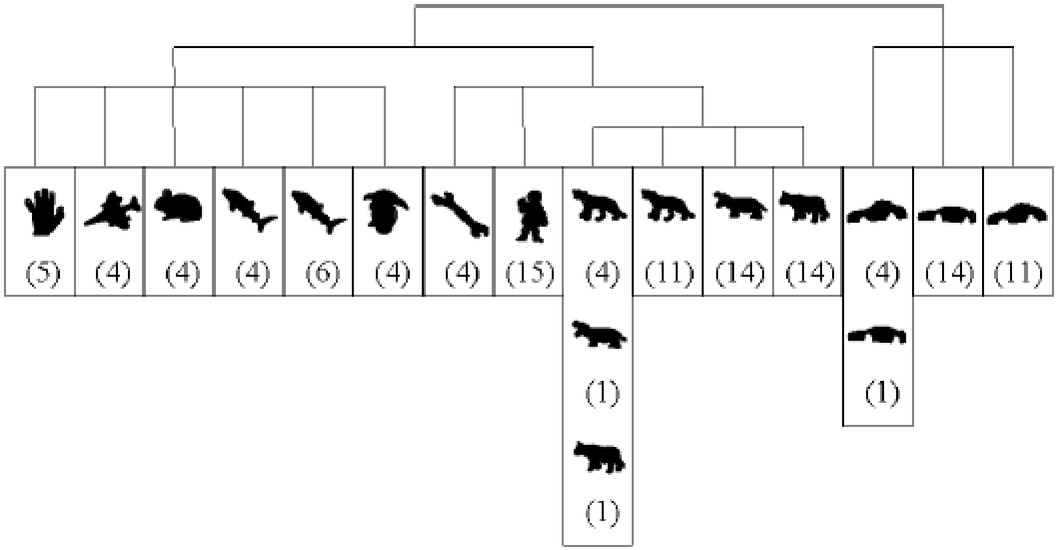
Klein and Kimia's Similarities (SODA'01)



Left: Similarity matrix used in the experiment. Right: Hierarchy produced by our algorithm.

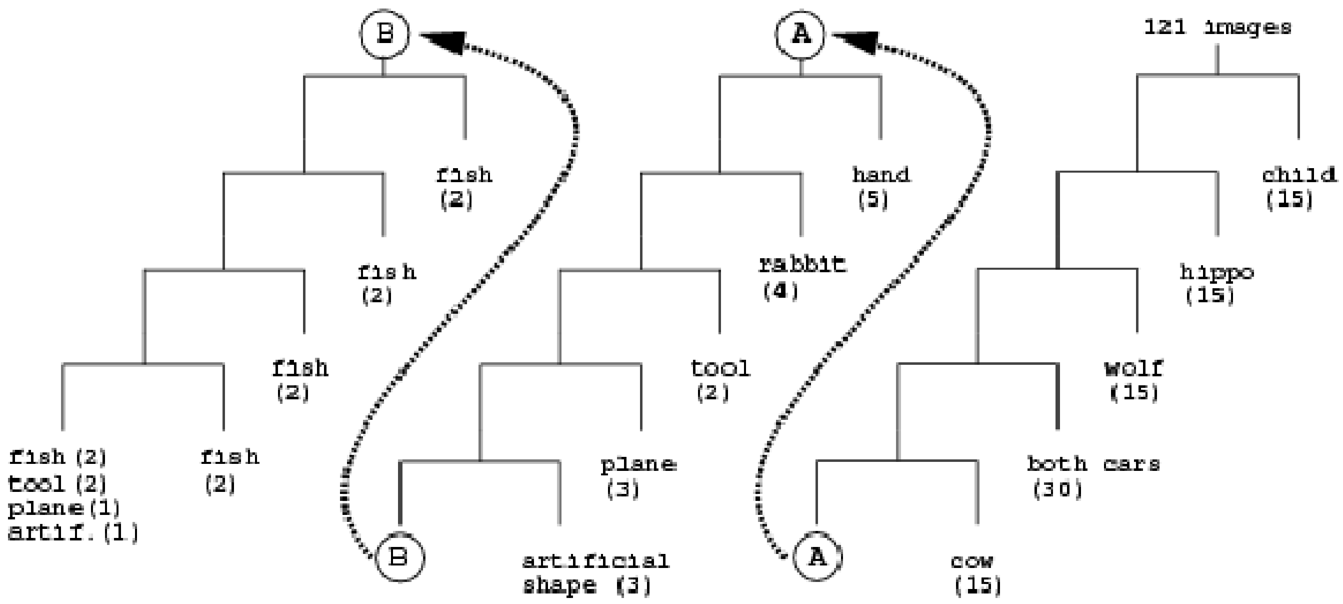
Gdalyahu and Weinshall's Similarities (PAMI 01)



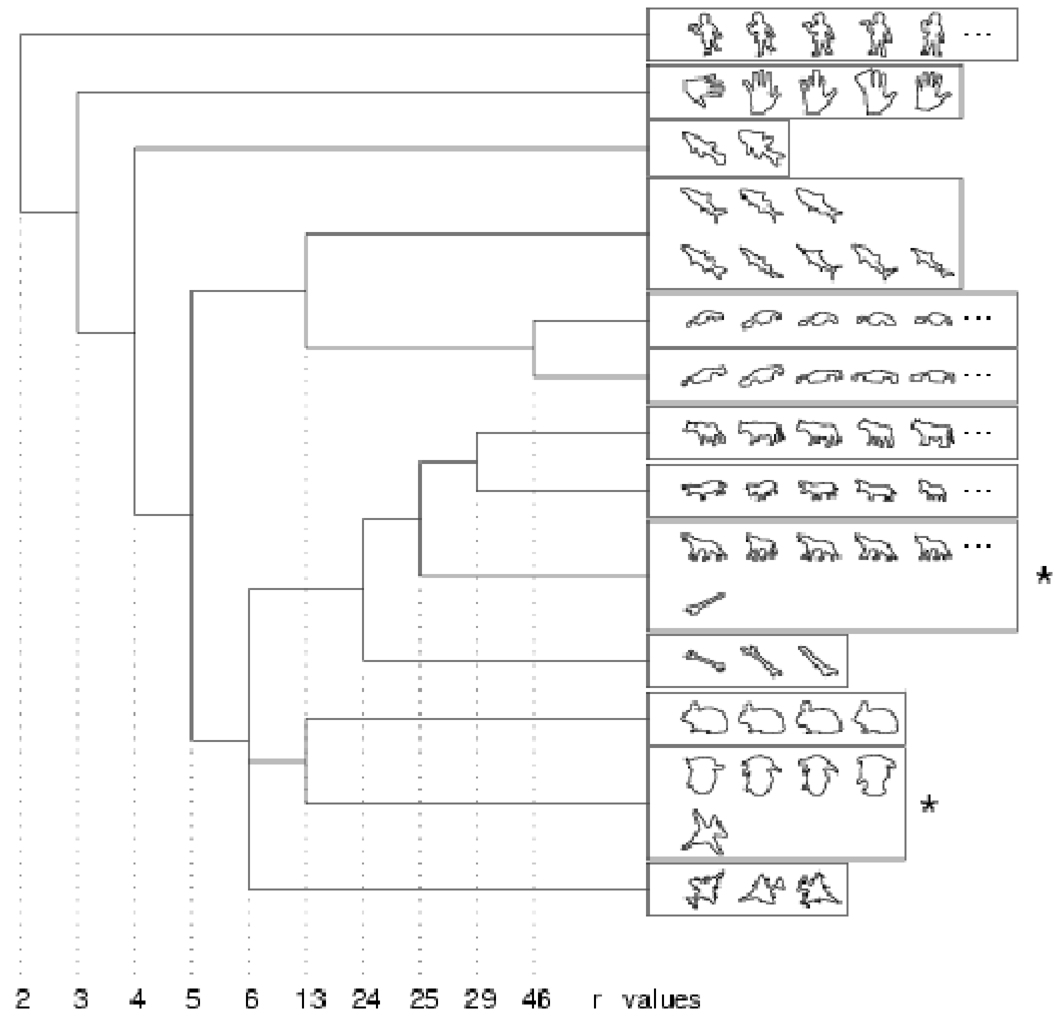


Left: Similarity matrix used in the experiment (courtesy of Y. Gdalyahu). Right: Hierarchy produced by our algorithm.

Factorization Results (Perona and Freeman, 98)



Typical-cut Results (From Gdalyahu, 1999)



Conclusions

- Introduced the notion of a dominant set of vertices in an edge-weighted graph, and defined a new notion of a cluster.
- Established a connection between the (combinatorial) problem of finding dominant sets and (continuous) quadratic programming.
- Used straightforward parallel dynamics from evolutionary game theory that can be coded in a few lines of MATLAB.
- Demonstrated potential of the approach on image segmentation.
- Extended the framework to hierarchical clustering
- Demonstrated its potential on the problem of organizing a shape database.

References

M. Pavan and M. Pelillo.
A new graph-theoretic approach to clustering and segmentation.
Proc. *CVPR 2003*.

M. Pavan and M. Pelillo.
Dominant sets and hierarchical clustering.
Proc. *ICCV 2003*.

M. Pavan and Pelillo.
Efficient out-of-sample extension of dominant-set clusters.
Proc. *NIPS*04*

A. Sperotto, M. Pavan, and M. Pelillo.
Perceptual grouping and dominant sets.
In preparation (2005).

A. Torsello and M. Pelillo
Grouping with asymmetric affinities: A game-theoretic perspective.
In preparation (2005).