Locally Scaled Density Based Clustering

Ergun Biçici and Deniz Yuret

Koç University, Istanbul, Turkey
www.ku.edu.tr

April 13, 2007

ICANNGA 2007 Presentation

Today's Talk

- Density Based Clustering
- Local Scaling
- Locally Scaled Density Based Clustering (LSDBC)
- Experiments
- Conclusion

Density Based Clustering

- ε: Radius of the volume of data points to look for
- $\wp$: Minimum number of points that has to be exceeded
- Let $d(p, q)$ give the distance between two points p and q
- ε neighborhood of a point p:

$$N_\varepsilon(p) = \{q \in Points \mid d(p, q) \leq \varepsilon\}$$

Density Based Clustering

- **Definition 1.** (**Directly density-reachable**)
  A point p is directly density reachable from a point q wrt. ε and $\wp$, if $p \in N_\varepsilon(q)$ and $|N_\varepsilon(q)| \geq \wp$ (core point condition). $\square$

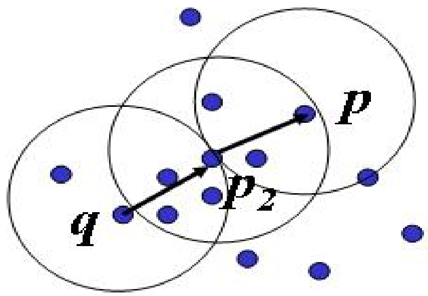
- **Definition 2.** (**Density-reachable**)
  A point p is density reachable from a point q wrt. ε and $\wp$, if there is a chain of points $p_1, p_2, ..., p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density reachable from p_i. $\square$

- **Definition 3.** (**Density-connected**)
  A point p is density connected to a point q wrt. ε and $\wp$, if there is a point r such that both p and q are density reachable from r wrt. ε and $\wp$. $\square$

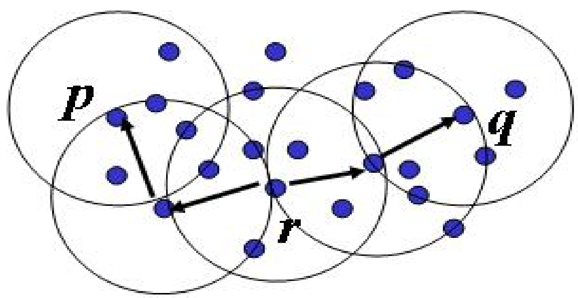
- A *cluster* C wrt. ε and $\wp$ is a non-empty set of points such that $\forall p, q \in C$, p is density connected to q wrt. ε and $\wp$.

Density Based Clustering

- Density-reachable



- Density-connected



DBSCAN's Results [1]

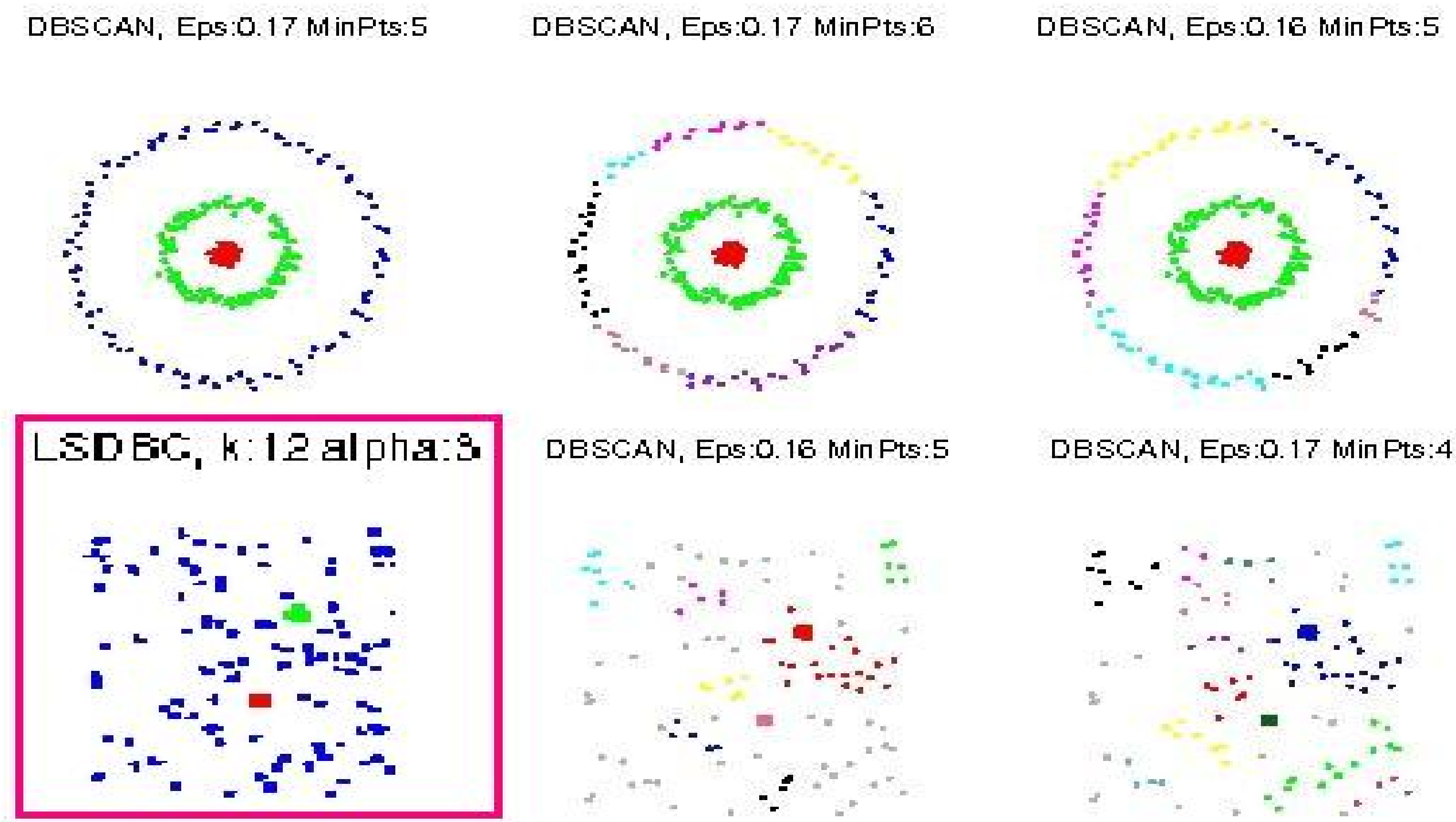


Figure 1: Density based clustering is sensitive to minor changes in ε and $\wp$

Local Scaling

- Scale distances proportional to its distance to its kth nearest neighbor.

- Given two points x_i and x_j, let A_{x_i,x_j} denote the affinity between the two points.

- $A_{x_i,x_j} = exp(-\frac{d^2(x_i,x_j)}{\sigma^2})$, where σ is a threshold distance below which two points are thought to be similar.

- A local scaling parameter: $\sigma_i = d(x_i, x_i^k)$ (k=7 in [2])

- $\hat{A}_{ij} = exp(-\frac{d^2(x_i,x_j)}{\sigma_i \sigma_j})$

Locally Scaled Density Based Clustering

Input: D: Distance matrix, k: input to *kNN*-dist function, n: number of dimensions, α: density threshold.

```
for p ∈ Points do
    p.class = UNCLASSIFIED;
    [p.ε, p.neighbors] = kNNDistVal(D, p, k);
end
Points.sort();                                   /* Sort on ε */
ClusterID = 1;
for p ∈ Points do
    if p.class == UNCLASSIFIED and localMax(p) then
        ExpandCluster(p, ClusterID, n, α);
        ClusterID = ClusterID + 1;
    end
end
```

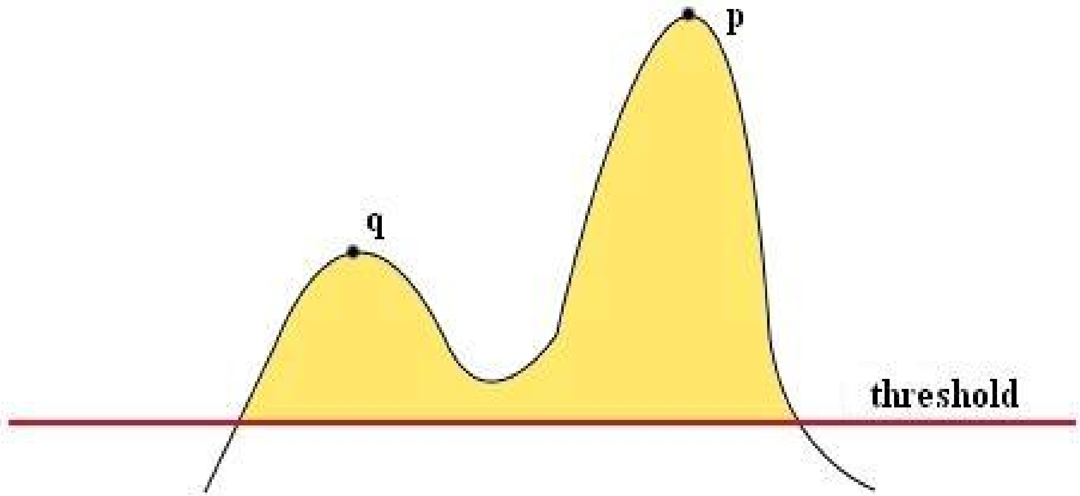
Algorithm 1: LSDBC: Locally Scaled Density Based Clustering

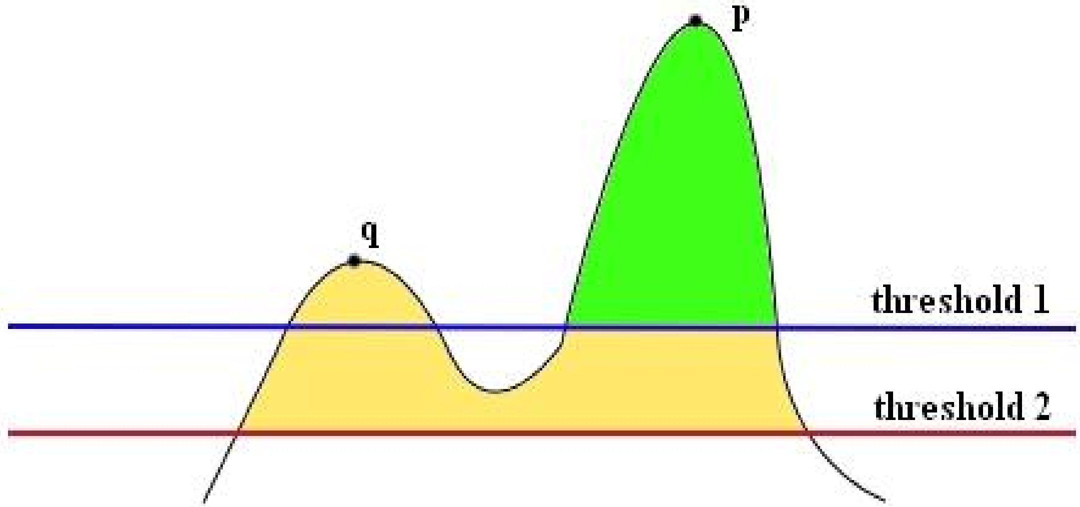
$ExpandCluster$: Expands the cluster of a given point

Input: $point$, $ClusterID$, n, α.

$point.class = ClusterID$;
$Seeds = point.neighbors$; /* Remove clustered points */
while $Seeds.length > 0$ **do**
    $currentP = Seeds.first()$; /* density $\geq \frac{\texttt{density}(core)}{2^\alpha}$ */
    if $currentP.Eps \leq 2^{\alpha/n} \times point.Eps$ **then**
        $Neighbors = currentP.neighbors$;
        for $neighborP \in Neighbors$ **do**
            if $neighborP.class == UNCLASSIFIED$ **then**
                $Seeds.append(neighborP)$;
                $neighborP.class = ClusterID$;
            end
        end
    end
    $Seeds.delete(currentP)$;
end

ExpandCluster: Expands the cluster of a given point





Experiments and Results

We compared our algorithm, LSDBC with:

1. Original density based clustering algorithm, DBSCAN,

2. Spectral clustering with local scaling,

3. k-means clustering.

Robustness of LSDBC

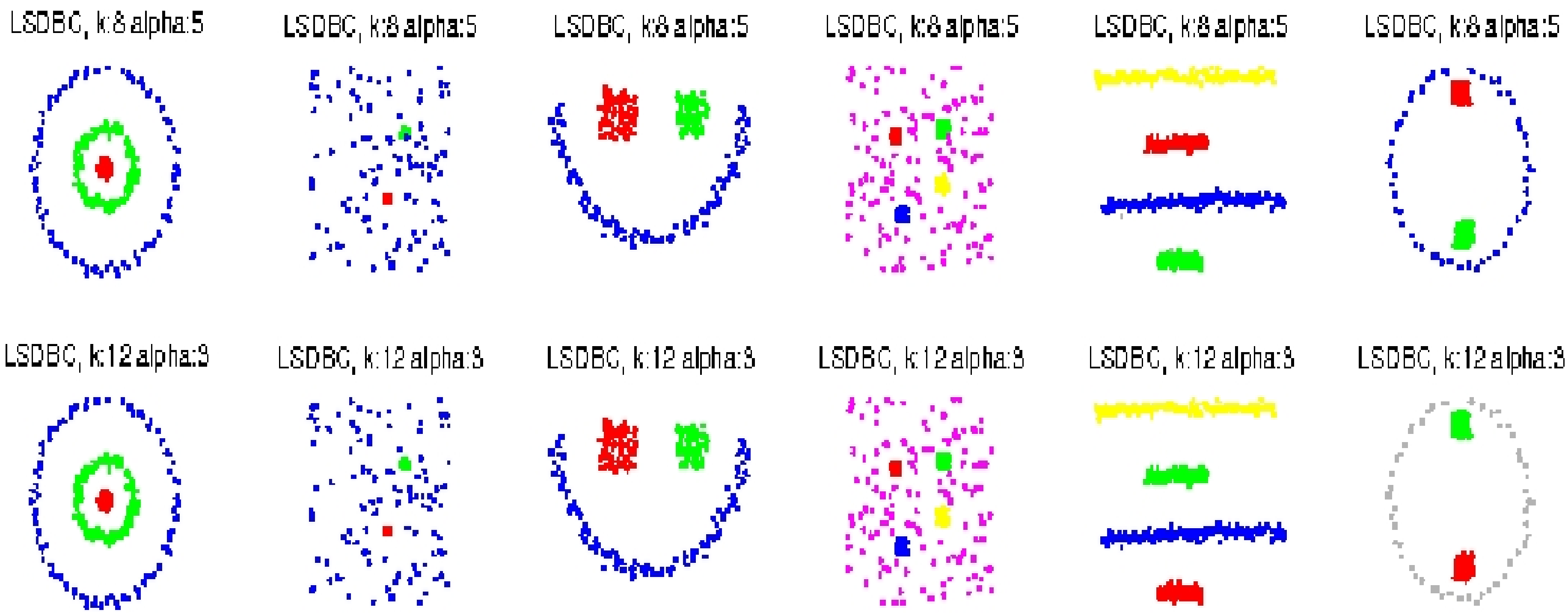


Figure 2: Robustness of LSDBC for different values of k and α

Comparative Results: DBSCAN

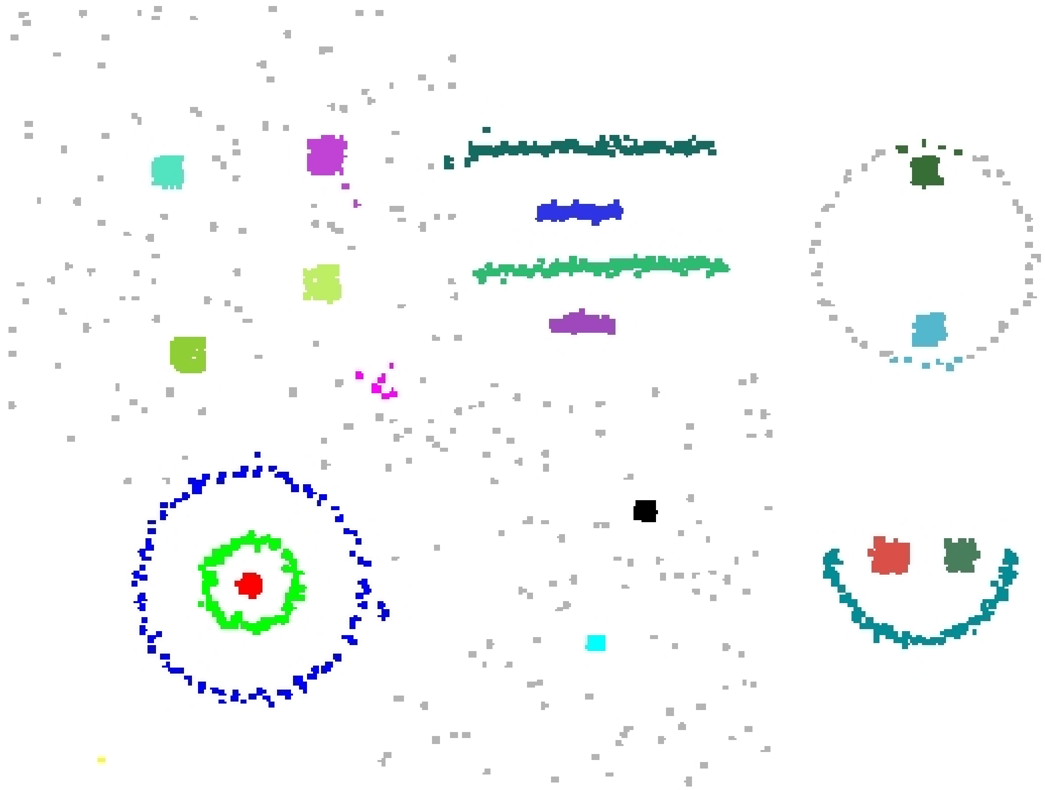


Figure 3: With best parameter settings: $r{=}0.04$, $k{=}5$: Only highly dense regions are recognized.

Comparative Results: k-means

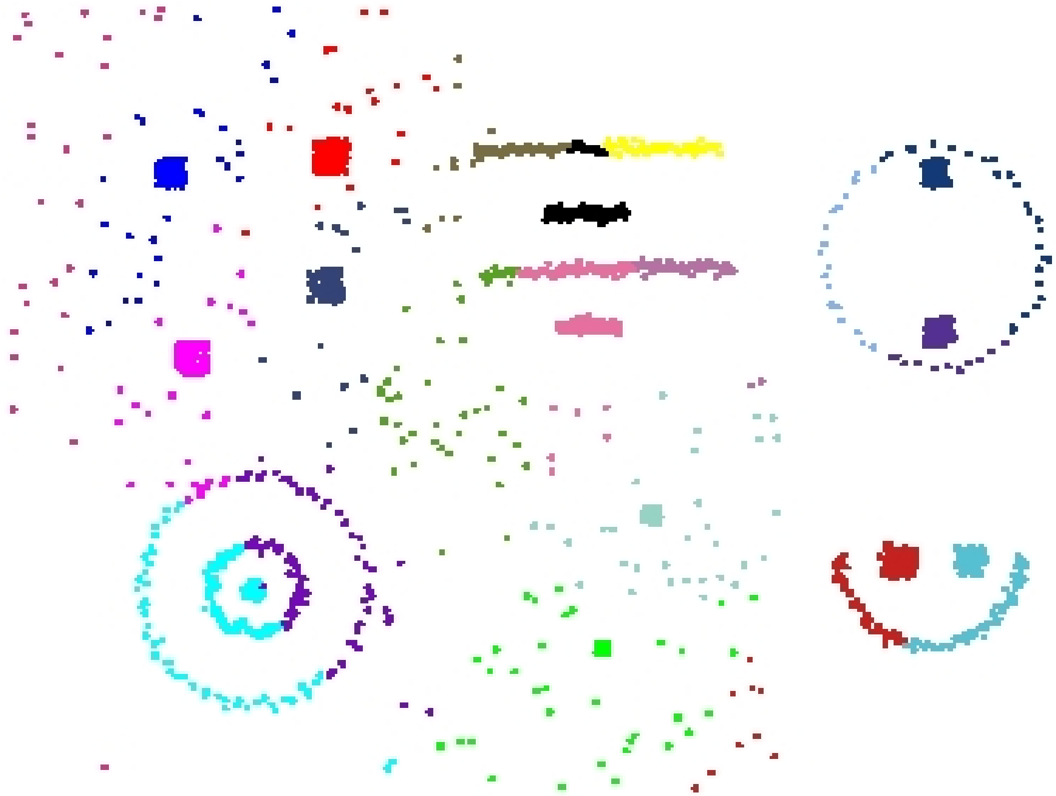


Figure 4: With best parameter settings: k=20: Densely populated regions are divided.

Comparative Results: Spectral with local scaling

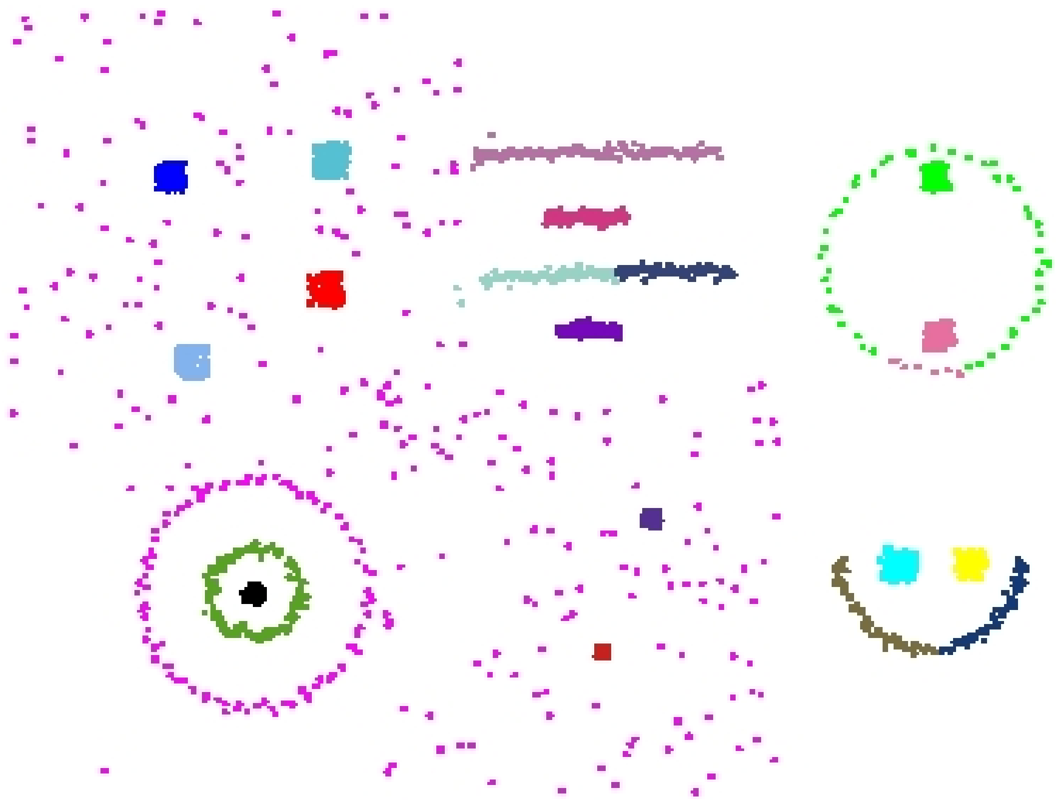


Figure 5: With best parameter settings: k=20: Densely populated regions are divided, diversely populated regions merged.

Comparative Results: LSDBC

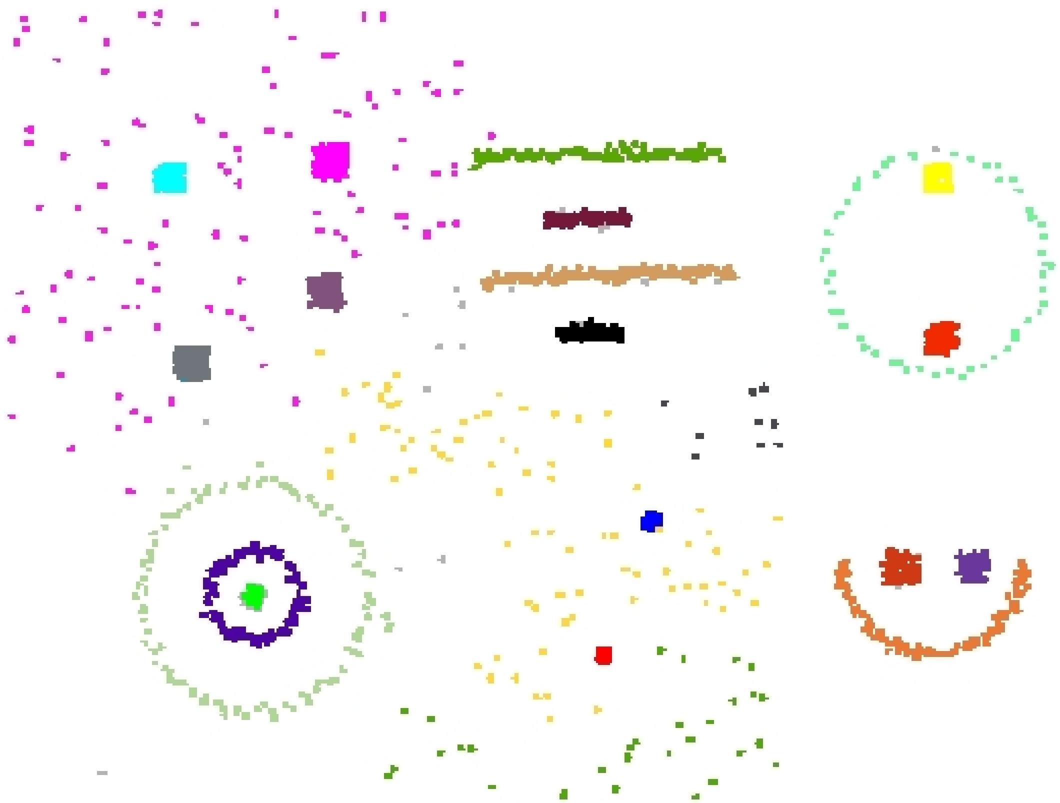


Figure 6: k=6, α=3: Background clutter is divided into 3 clusters.

Comparative Results: LSDBC

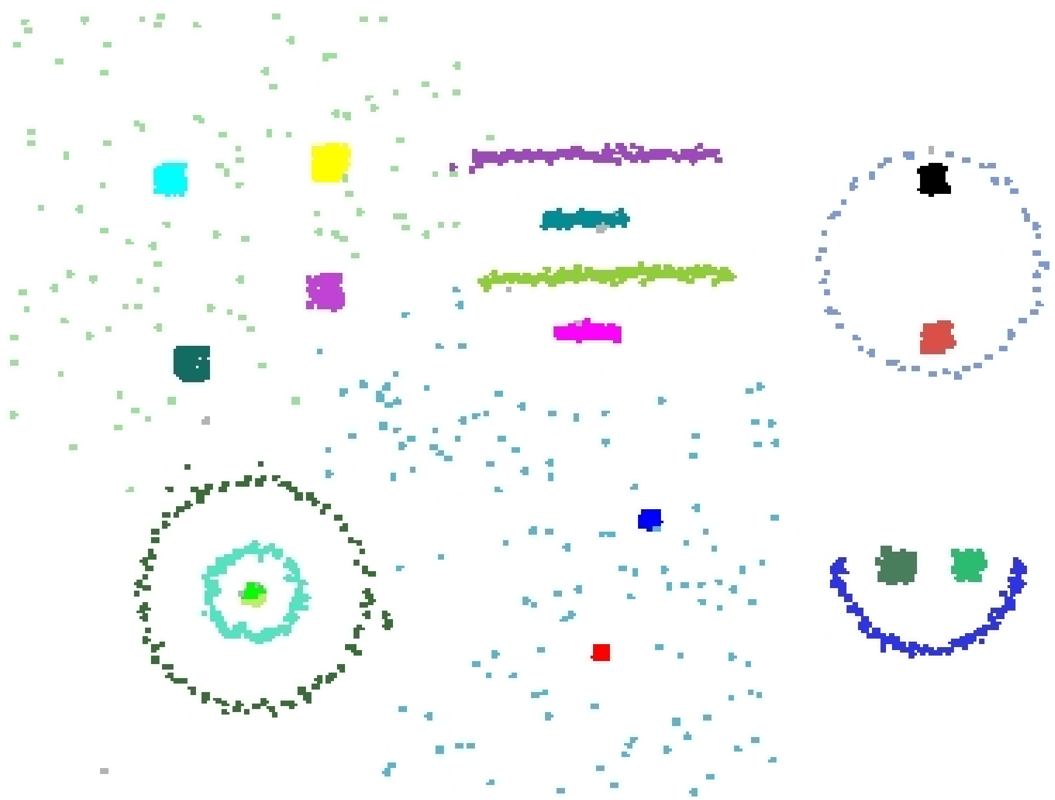


Figure 7: $k=7$, $\alpha=3$: Background clutter is divided into 2 clusters.

Comparative Results: LSDBC

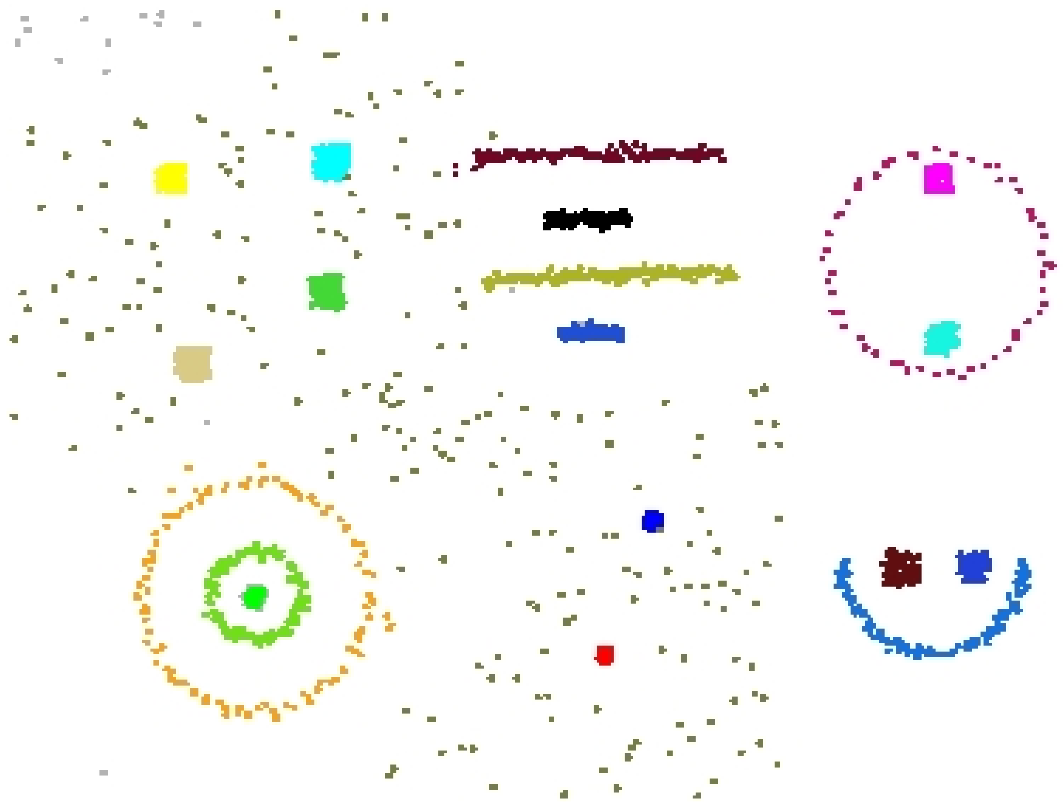
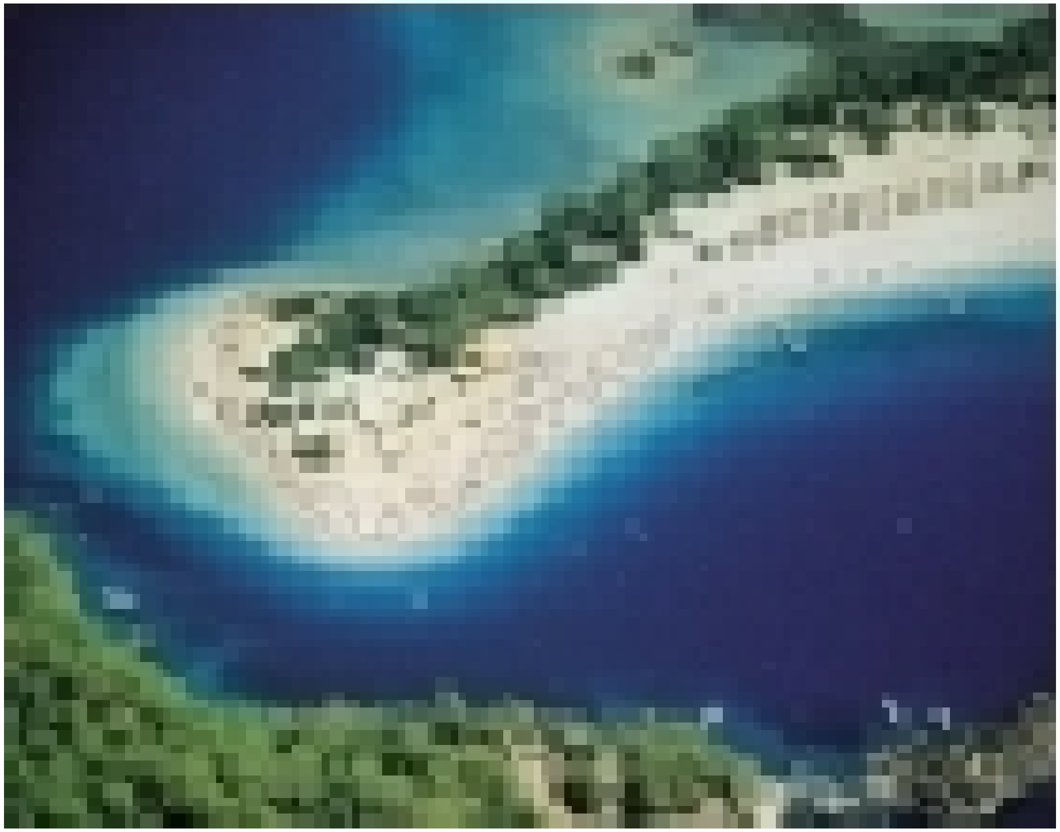


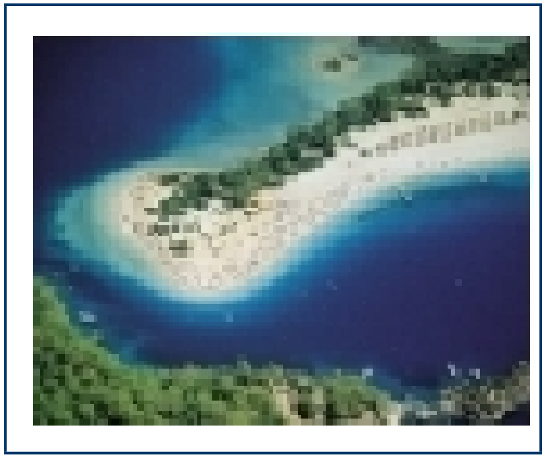
Figure 8: $k=8$, $\alpha=3$: Background clutter is classified as a single cluster.

Image Segmentation Results

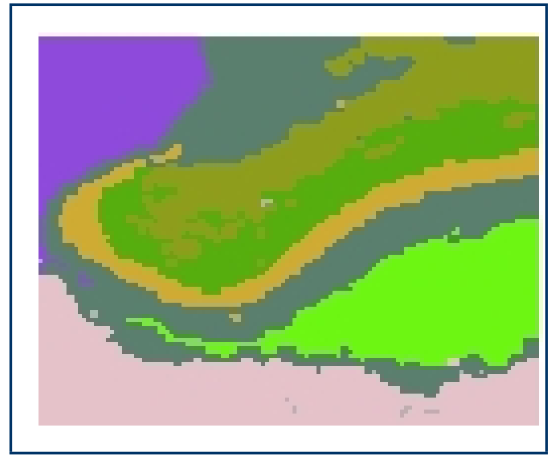


Segmentation of an image of a seaside, Oludeniz, Fethiye, Turkey

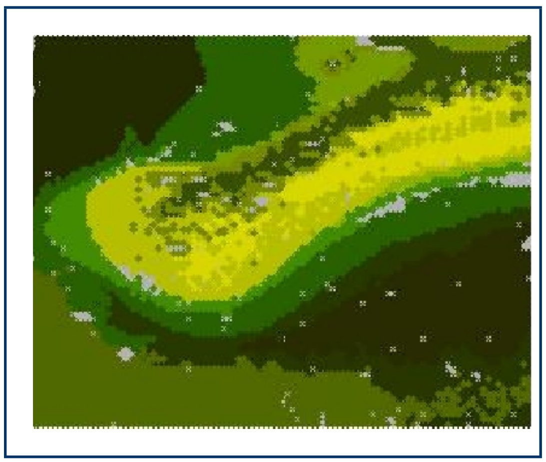
Image Segmentation Results



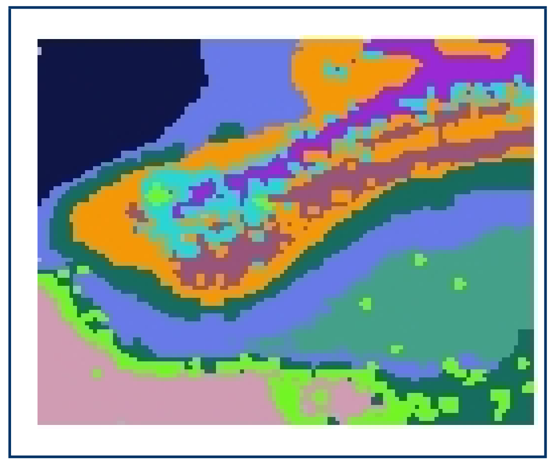
Oludeniz (original)



LSDBC: k=12, α=5



k=10, α=6



k=13, α=6

Conclusion

- Locally scaled density based clustering is introduced.
- Clusters are discovered via a k-NN density estimation method and grown until the density falls below a pre-specified ratio of the center point's density.
- LSDBC is able to identify clusters of arbitrary shape on noisy backgrounds that contain significant density gradients.
- The performance is demonstrated on a number of synthetic datasets and real images for a broad range of its parameters.
- LSDBC can also be used to summarize and segment images into meaningful regions.

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[2] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Eighteenth Annual Conference on Neural Information Processing Systems*, 2004.

Thank you!

Acknowledgments: The research reported here was supported in part by TUBITAK.