UNIVERSITÀ DEGLI STUDI DI PADOVA

Segmentation by clustering, k-means

Stefano Ghidoni

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

INTELLIGENT AUTONOMOUS SYSTEMS LAB

Agenda

- The k-means clustering
  - Objective function
  - Initialization
  - Iterative solution
  - Examples

[1]Shai Shalev-Shwartz, Shai Ben-David, "Understanding machine learning", Cambridge University Press

Segmentation techniques

- Segmentation by thresholding (histogram-based)
- Region growing methods
- Watershed transformation
- Clustering-based methods
- Model-based segmentation
- Edge-based methods
- Graph partitioning methods
- Multi-scale segmentation
- Many others…

Clustering

- Clustering: the task of grouping a collection of heterogeneous elements into sets (clusters) of similar elements
- Two questions:
  - How are *elements* described in the context of computer vision?
  - What does similar mean?

Image representation

- How to provide an image representation that is compact and expressive?

Image representation

- How to provide an image representation that is compact and expressive?
- We represent each pixel with a feature vector
  - This representation depends on the goal of the image analysis process we are implementing
  - A multi-dimensional vector
- One feature vector for each pixel!

Image representation

- How to provide an image representation that is compact and expressive?
- We represent each pixel with a feature vector
  - The vector contains all the measurements that may be relevant to describe a pixel
    - Spatial position (coordinates)
    - Intensity/brightness (grayscale images)
    - Color information (RGB/YUV/CieLAB)
    - … (including a combination of the above)

Segmentation by clustering

- Segmentation by clustering: segment an image using a clustering technique
  - Provide the vector representation previously discussed
  - Apply a suitable clustering algorithm
    - Pixels grouped based on their vectors

Distance functions

- Clustering techniques often evaluate how similar two pixels are
  - This means comparing the corresponding feature vectors
- We need a distance function to compare vectors
- Distance is critical when multiple types of data are involved
  - E.g., spatial + brightness

Distance functions

- Some typical distance functions – D is the dimension of the feature vector
- Absolute value / Manhattan

$$d_a(\overline{x_i}, \overline{x_j}) = \sum_{k=1}^{D} |x_{i,k} - x_{j,k}|$$

- Euclidean

$$d_e(\overline{x_i}, \overline{x_j}) = \sqrt{\sum_{k=1}^{D} \left(x_{i,k} - x_{j,k}\right)^2}$$

- Minkowski

$$d_m(\overline{x_i}, \overline{x_j}) = \left[\sum_{k=1}^{D} \left(x_{i,k} - x_{j,k}\right)^p\right]^{\frac{1}{p}}$$

Clustering techniques

- Two basic approaches to clustering

1. **Divisive clustering**
   - Starting point: the entire dataset is considered as a cluster
   - Recursively split each cluster to yield a good clustering
     - Some form of cluster quality measurement is needed

Clustering techniques

- Two basic approaches to clustering

2. **Agglomerative clustering**
   - Starting point: every single pixel is considered as a cluster
   - Recursively merge each cluster to yield a good clustering
     - Some form of cluster quality measurement is needed

Clustering techniques

- Several clustering techniques are available:
  - K-means
  - Mean shift
  - Spectral clustering
  - Hierarchical clustering
  - Density-based approach
  - …

K-means

K-means

- A simple clustering algorithm
- Based on a fixed number of clusters (k)
  - It shall be provided to the algorithm
  - Is it a good or a bad element?

K-means

- A simple clustering algorithm
- Based on a fixed number of clusters (k)
  - It shall be provided to the algorithm
  - Is it a good or a bad element?
- After the process, each feature vector is associated with one of the k clusters

K-means

- The k clusters are disjoint sets $C_1, \dots, C_k$
  - Each C_i has a centroid $\boldsymbol{\mu}_i$
- The k-means objective function measures the distance between each data point and the centroid of its cluster

K-means

- The k clusters are disjoint sets $C_1, \dots, C_k$
  - Each C_i has a centroid $\boldsymbol{\mu}_i$
- The goal is to minimize the error made by approximating the points with the center of the cluster it belongs to

$$\min\left(\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i} d(\boldsymbol{x},\boldsymbol{\mu}_i)\right)$$

where $d(\cdot)$ is an appropriate distance

K-means

- Exhaustive search is computationally unfeasible (too many combinations)
  - We need an euristic approach
- Commonly used iterative algorithm
- Can be applied to vectors containing any set of features

Iterative clustering

Lloyd's algorithm (AKA k-means algorithm)

1. Get k initial centroids (see next slides)
2. Associate each point to the "closest" centroid

$$C_i = \{\boldsymbol{x} : i = \mathrm{argmin}_j d(\boldsymbol{x}, \boldsymbol{\mu}_j)\}$$

for $i = 1, \dots, k$

3. Compute the new centroids (center of mass of the associated points)

$$\mu_i = \frac{1}{C_i} \sum_{x \in C_i} \boldsymbol{x}$$

4. Repeat steps 2 and 3 until (the centroids do not sensibly move) or (max number of steps)

Initialization

- Cluster centroids should be initialized
- An initialization method is needed

1. Forgy method: k points randomly chosen among the data points
   - Centroids spread among the dataset

Initialization

- Cluster centroids should be initialized
- An initialization method is needed

2. Random partition: build the k clusters randomly assigning all the points to clusters, then computing the centroids
   - Centroids concentrated towards the dataset center of mass

K-means: pros & cons

- What are K-means pros & cons?

- Anti-spoiler ☺

K-means: pros & cons

- K-means pros & cons
- Pros
  - Light and simple
  - Computational complexity can be reduced using euristics
  - Fast convergence
- Cons
  - Optimality is not guaranteed
  - Solution found depends on initialization
  - The number of clusters, k, needs to be known in advance
  - Forces spherical symmetry of clusters (in the N-dimensional space)

Features for image segmentation

- K-means clustering can work considering:
  - The histogram (AKA gray levels, faster)
  - Pixel vectors (better results, tunable)
- Possible distance measures
  - Intensity level difference (grayscale)
  - Color channel difference (color image, depends on color space)
  - Combinations of position, color, texture descriptor, …

Color and intensity clustering

- Segmented pixels: mean intensity/color of its cluster
  - Focus on spatial distribution of clusters

Original

Intensity clustering

Color clustering

Color clustering

- Color clustering, increasing k

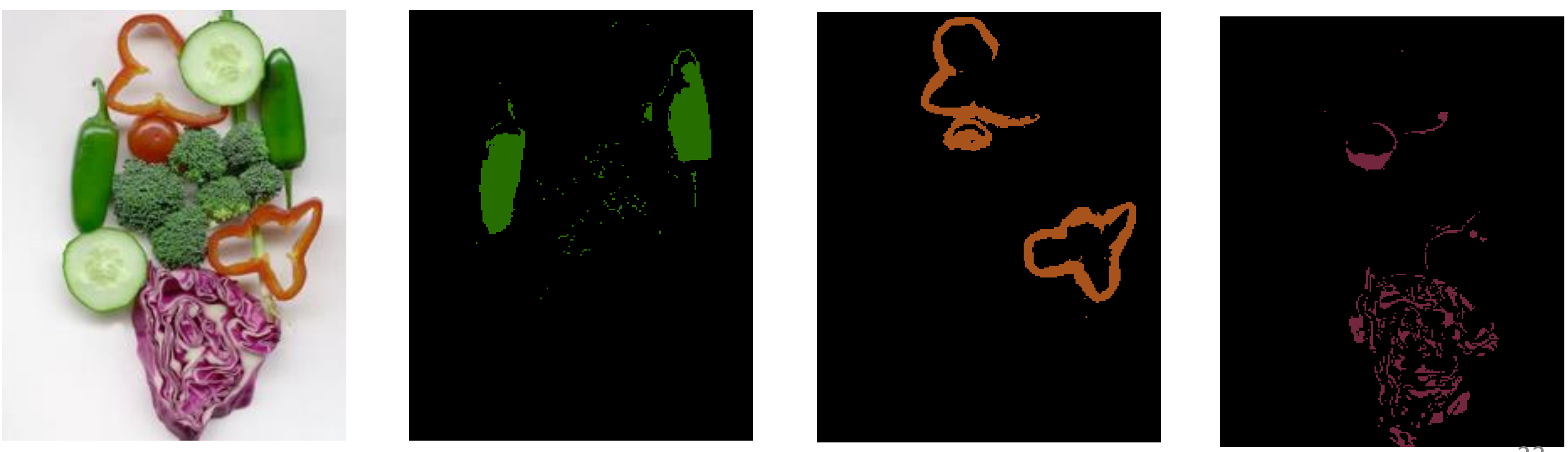
Original

Color clustering, k=5

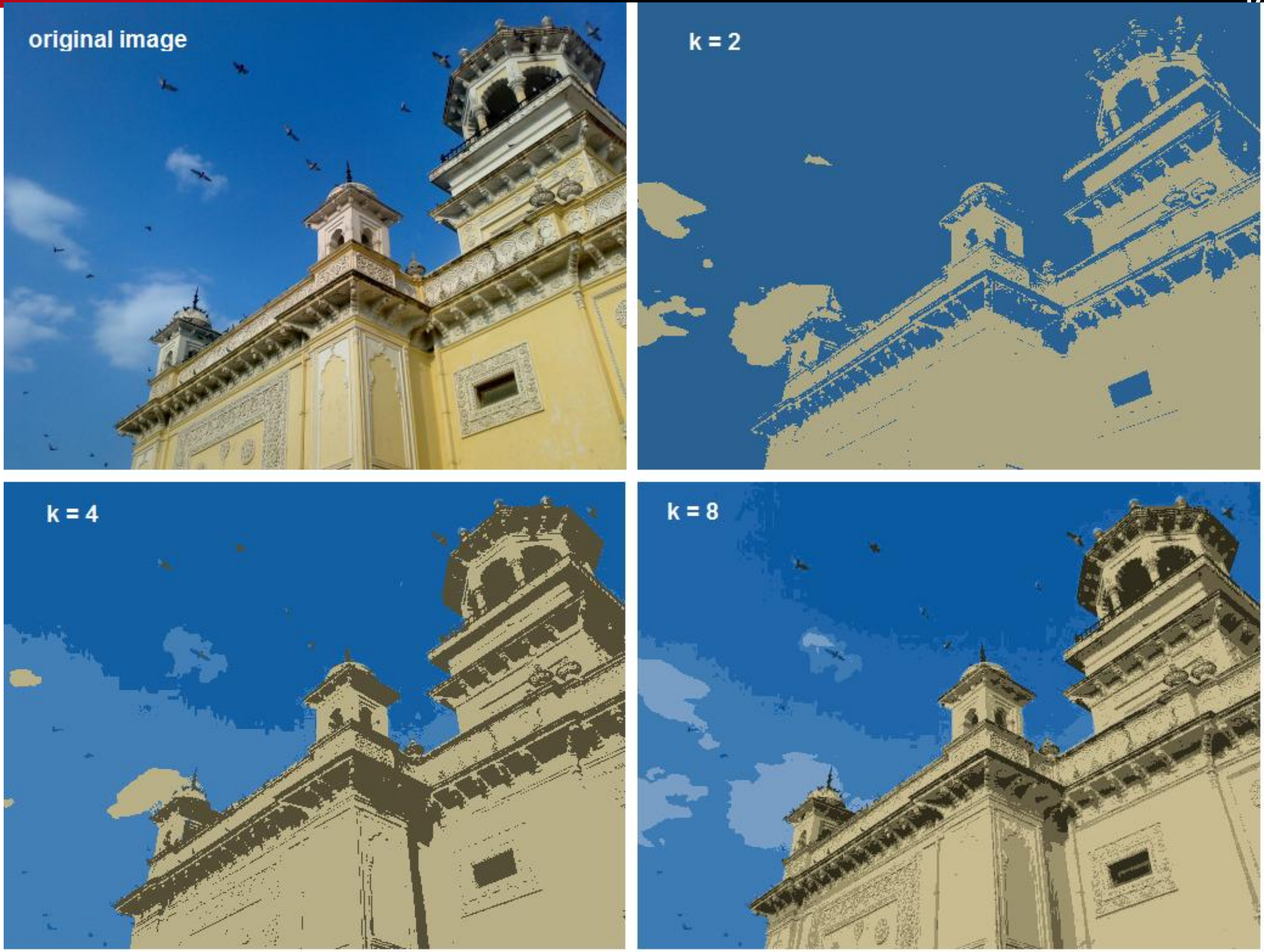
Color clustering, k=11

Clusters

- Some segments shown – not necessarily connected
- Some clusters associated with objects
  - Similar objects in the same cluster
- Some clusters are meaningless
- Observe spatial distribution
- Problems with textured objects (e.g., the cabbage)

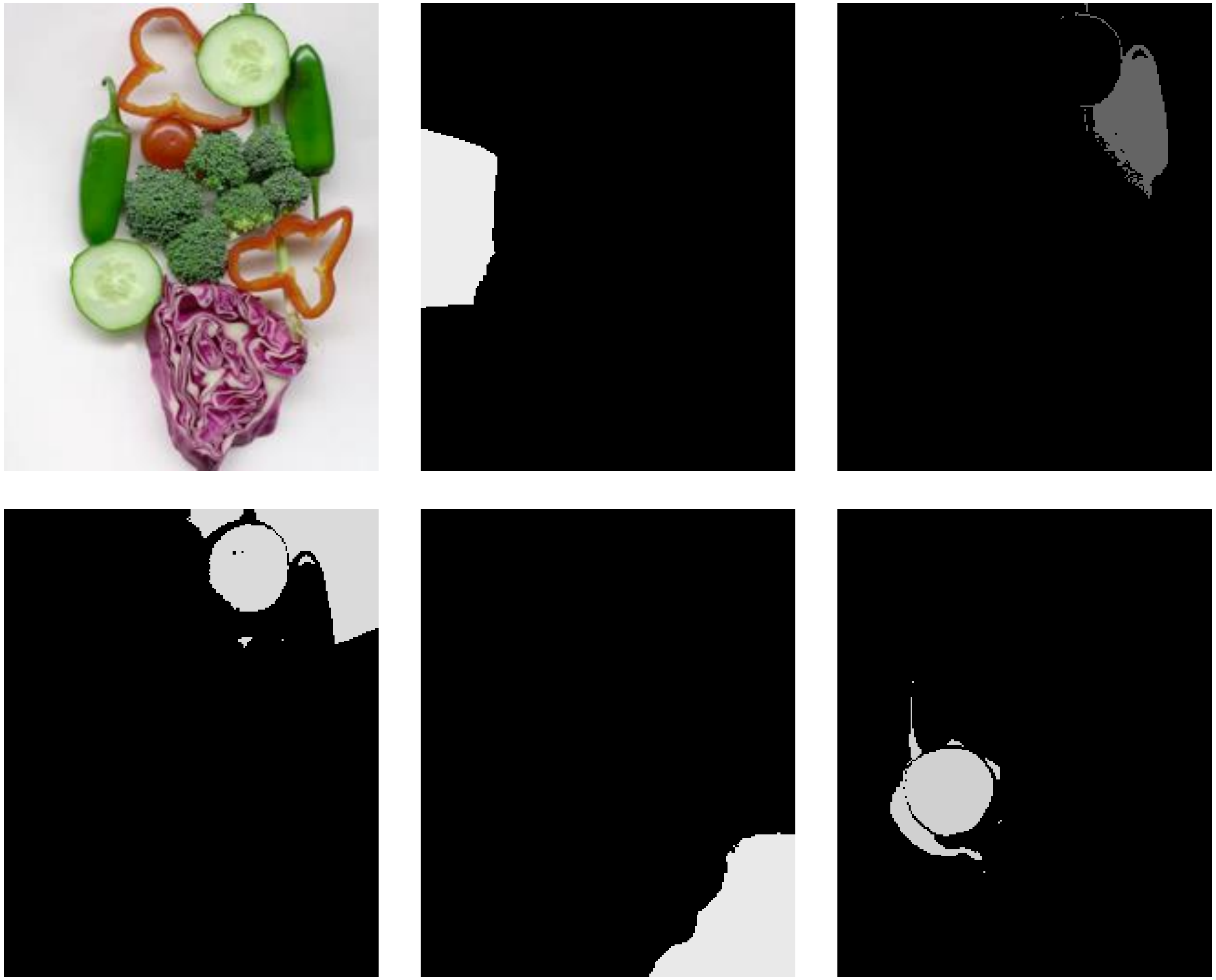


K-means using color alone



K-means using color and position

- Now using vectors including **color and position**
- K=20
- Improved object separation
- Background split among clusters: centroids too far away



UNIVERSITÀ DEGLI STUDI DI PADOVA

Segmentation by clustering, k-means

Stefano Ghidoni