IMAGE SEGMENTATION

Lecturer: Ondřej Drbohlav

Slide Authors: Václav Hlaváč, Ondřej Drbohlav

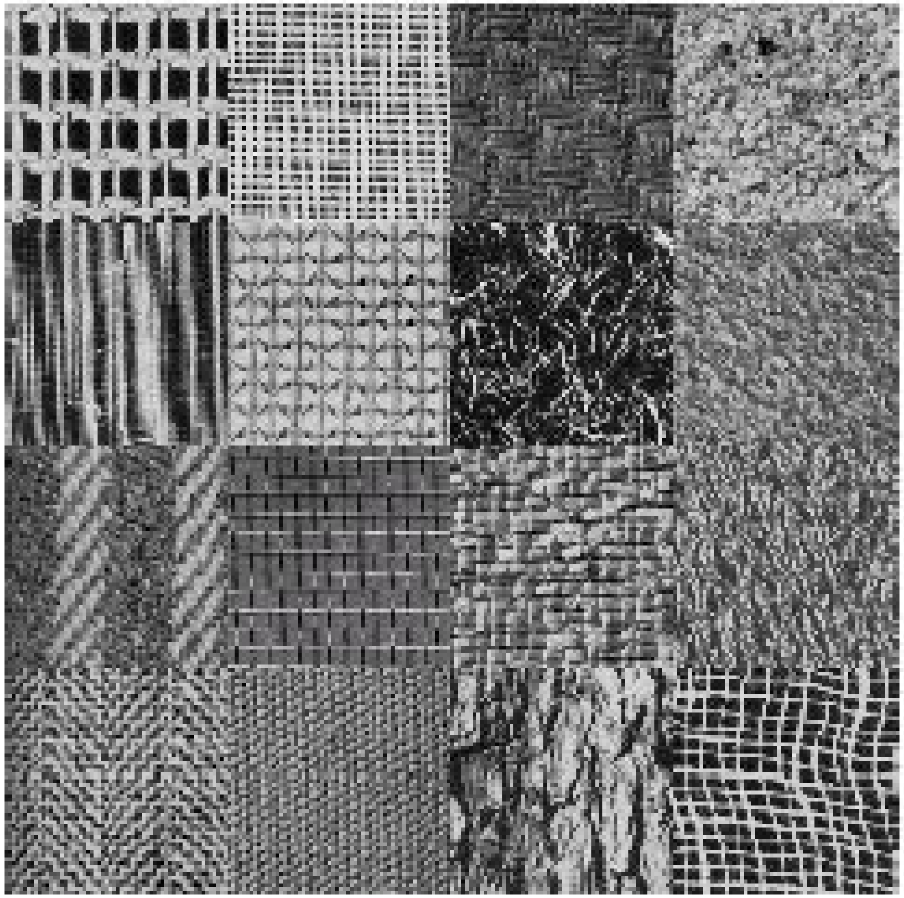
Dec 3, 2014 (with post-lecture corrections)

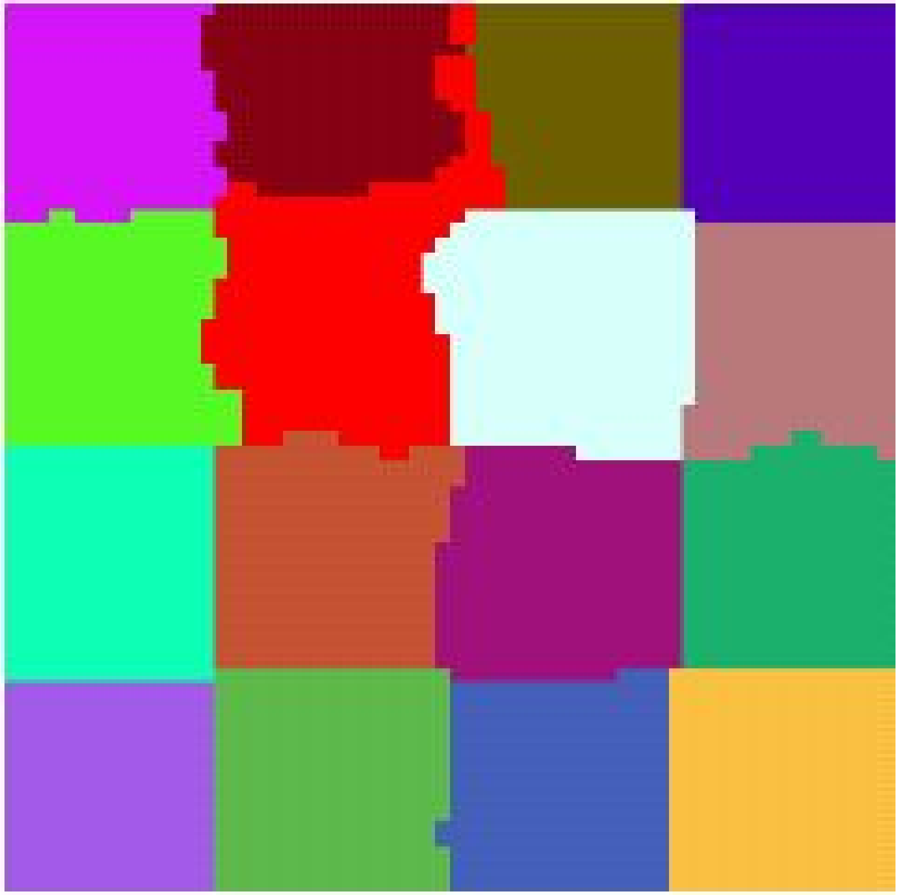
Outline of the talk:

- What is image segmentation?
- Thresholding, K-means, EM algorithm.
- . . . *other methods: next lecture.*

Segmentation, Example 1

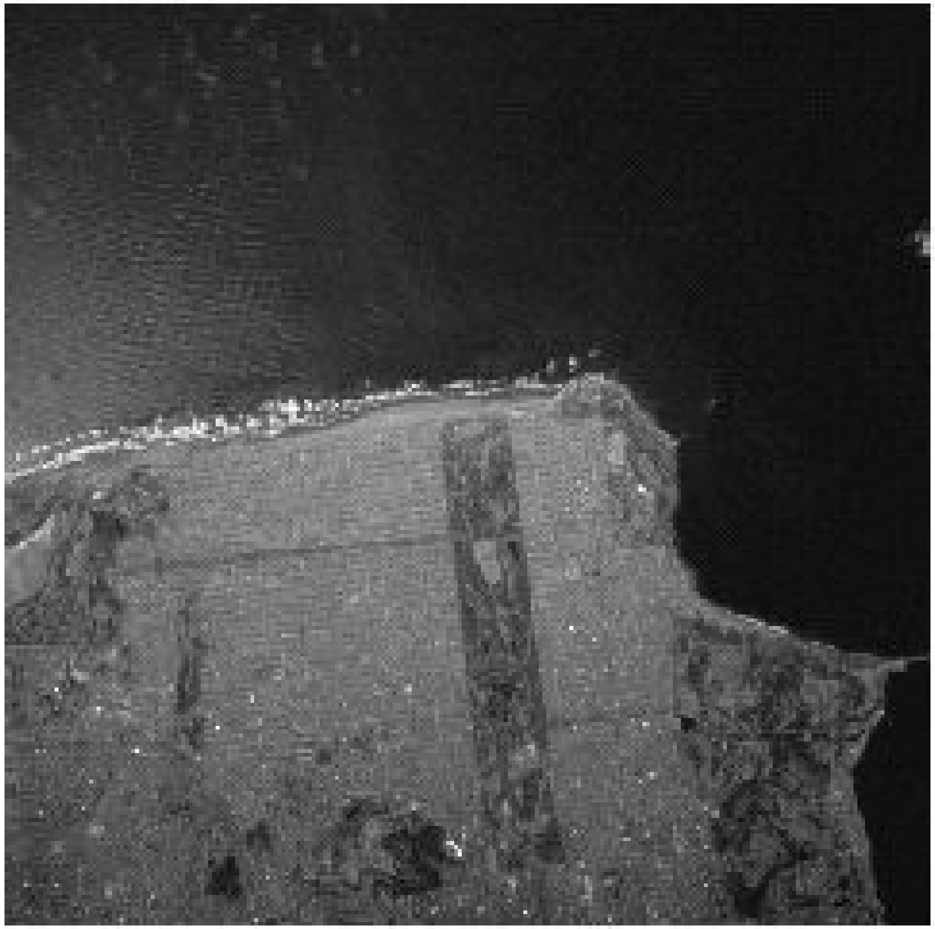
Task: Group pixels to regions with similar texture.

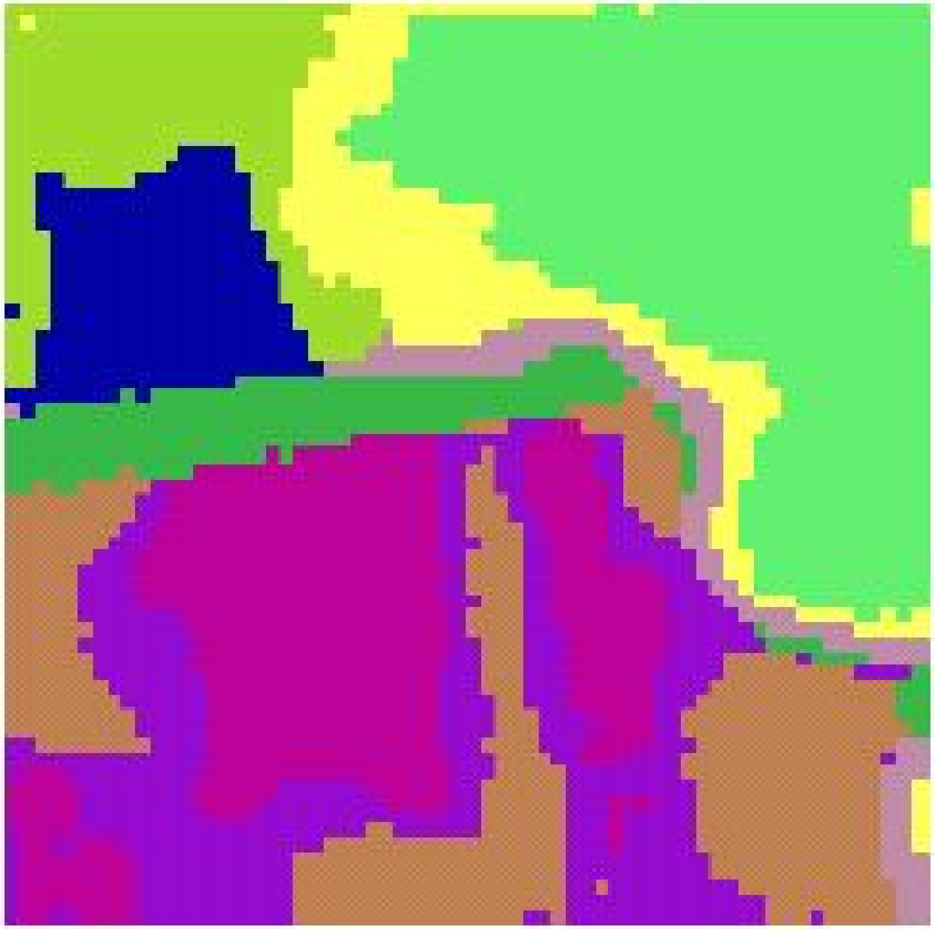




Segmentation, Example 2

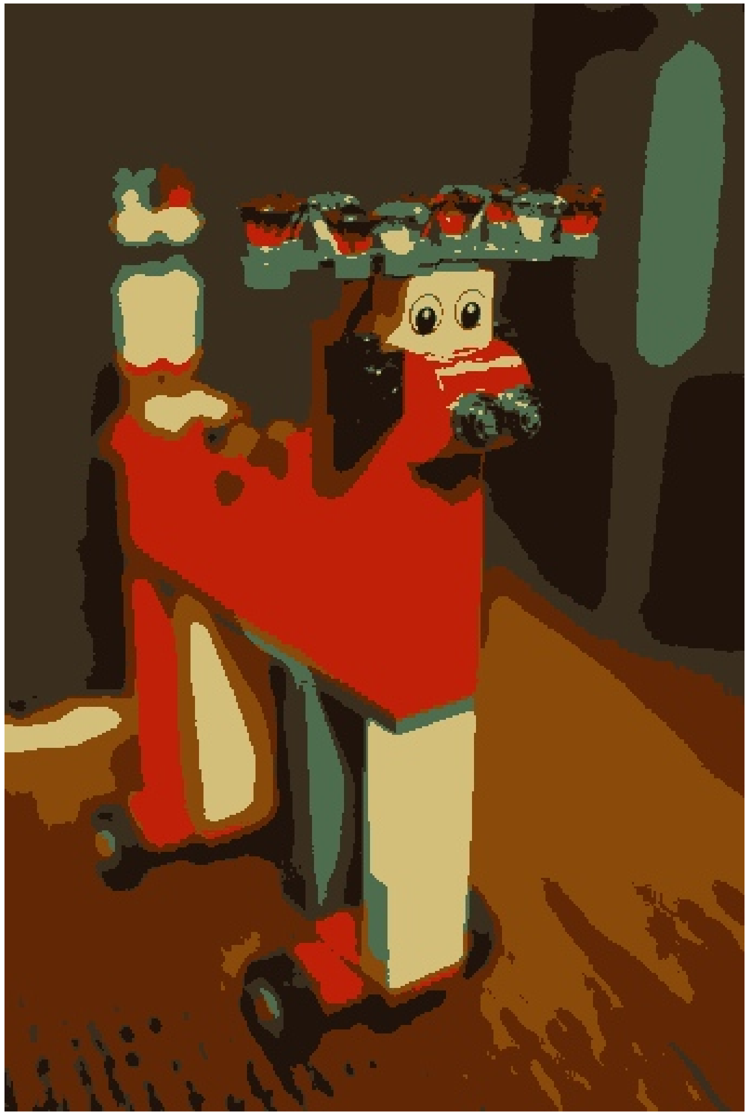
Task: Group pixels to regions corresponding to different types of land.
(aerial image of a sea coast)





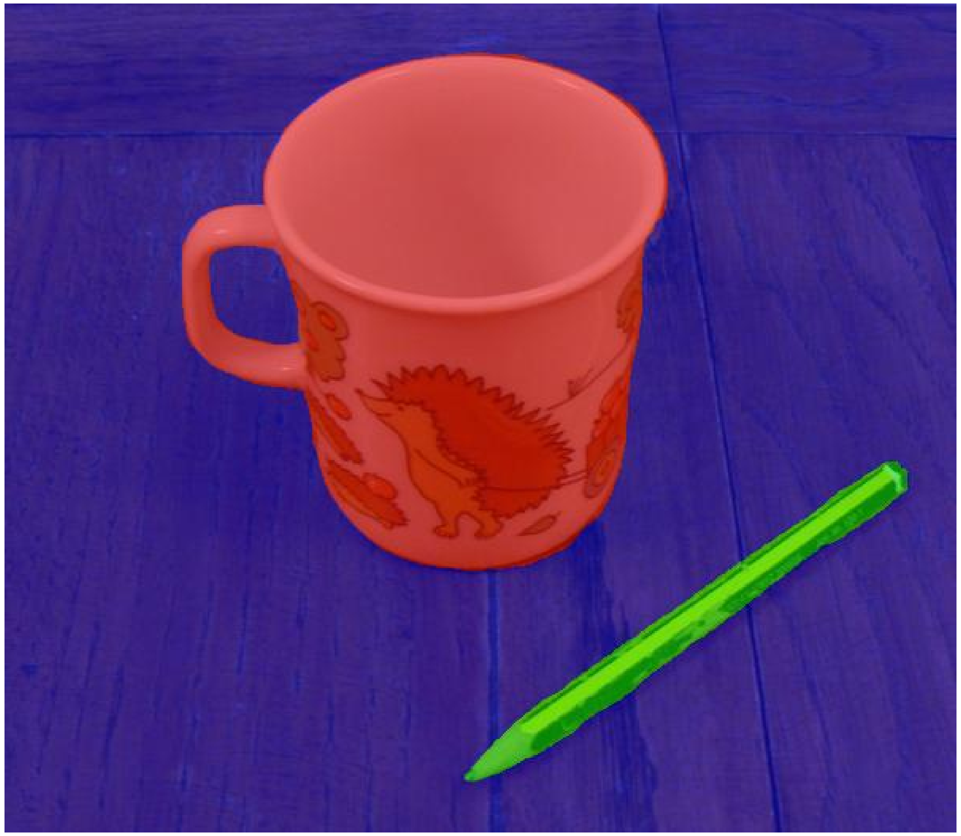
Segmentation, Example 3

Task: Reduce the number of colors in an image to $K = 7$ (segmentation by color, other conditions apply)

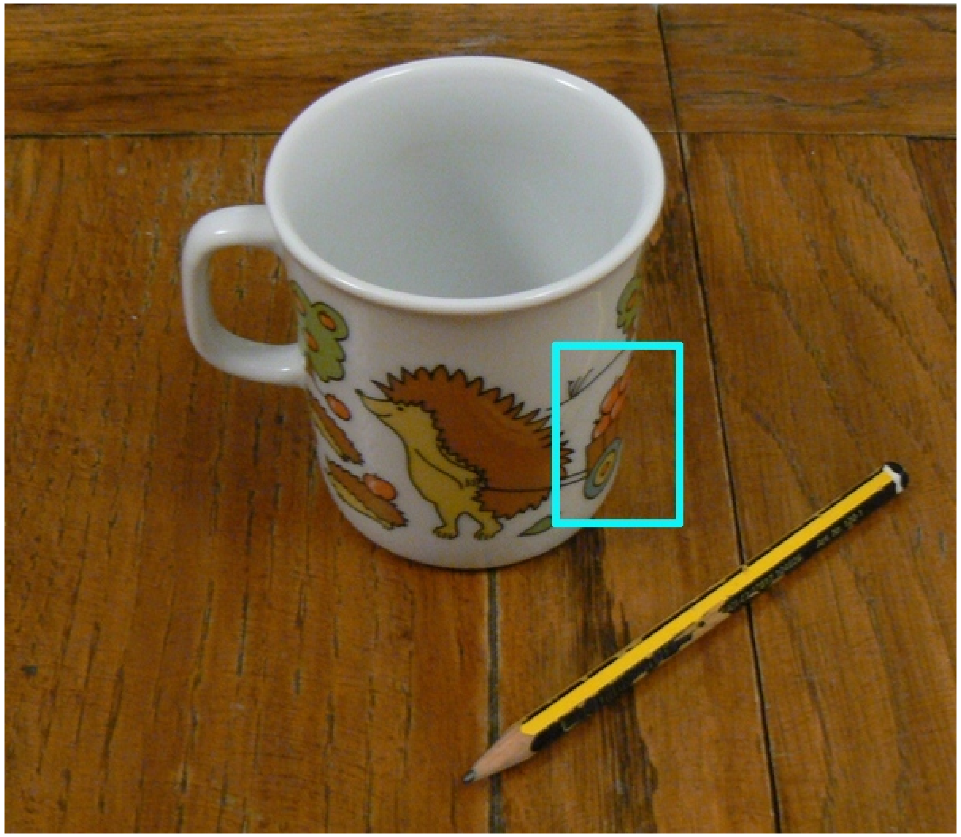


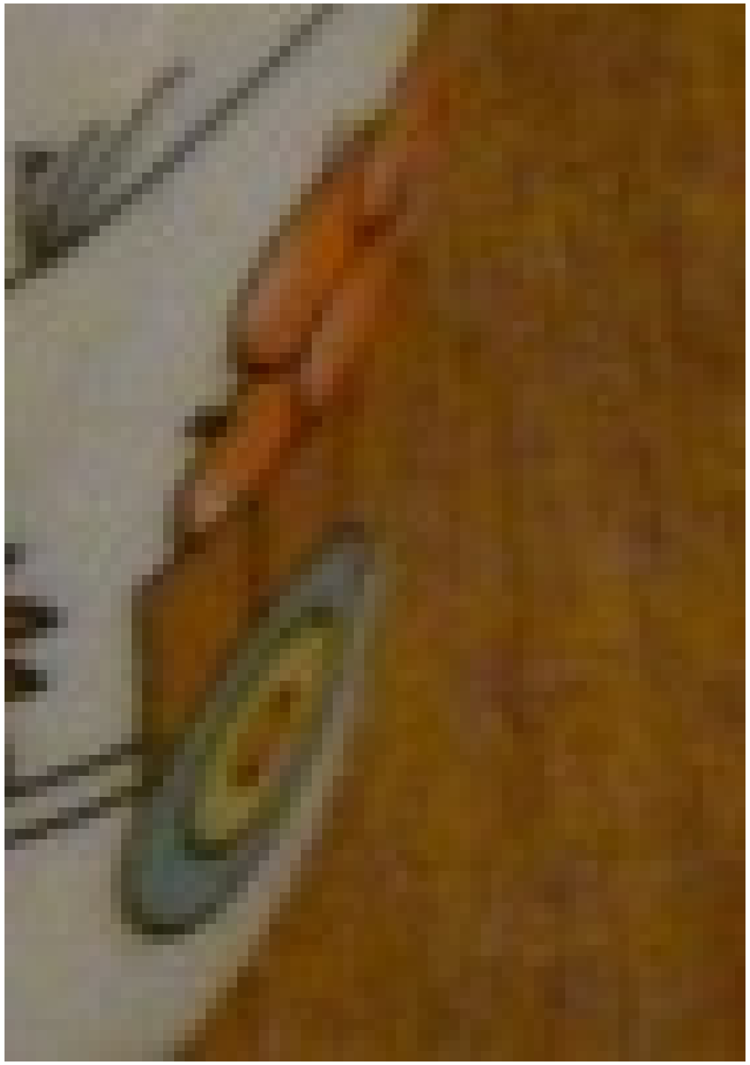
Segmentation, Example 4

Task: Group pixels to regions corresponding to objects.



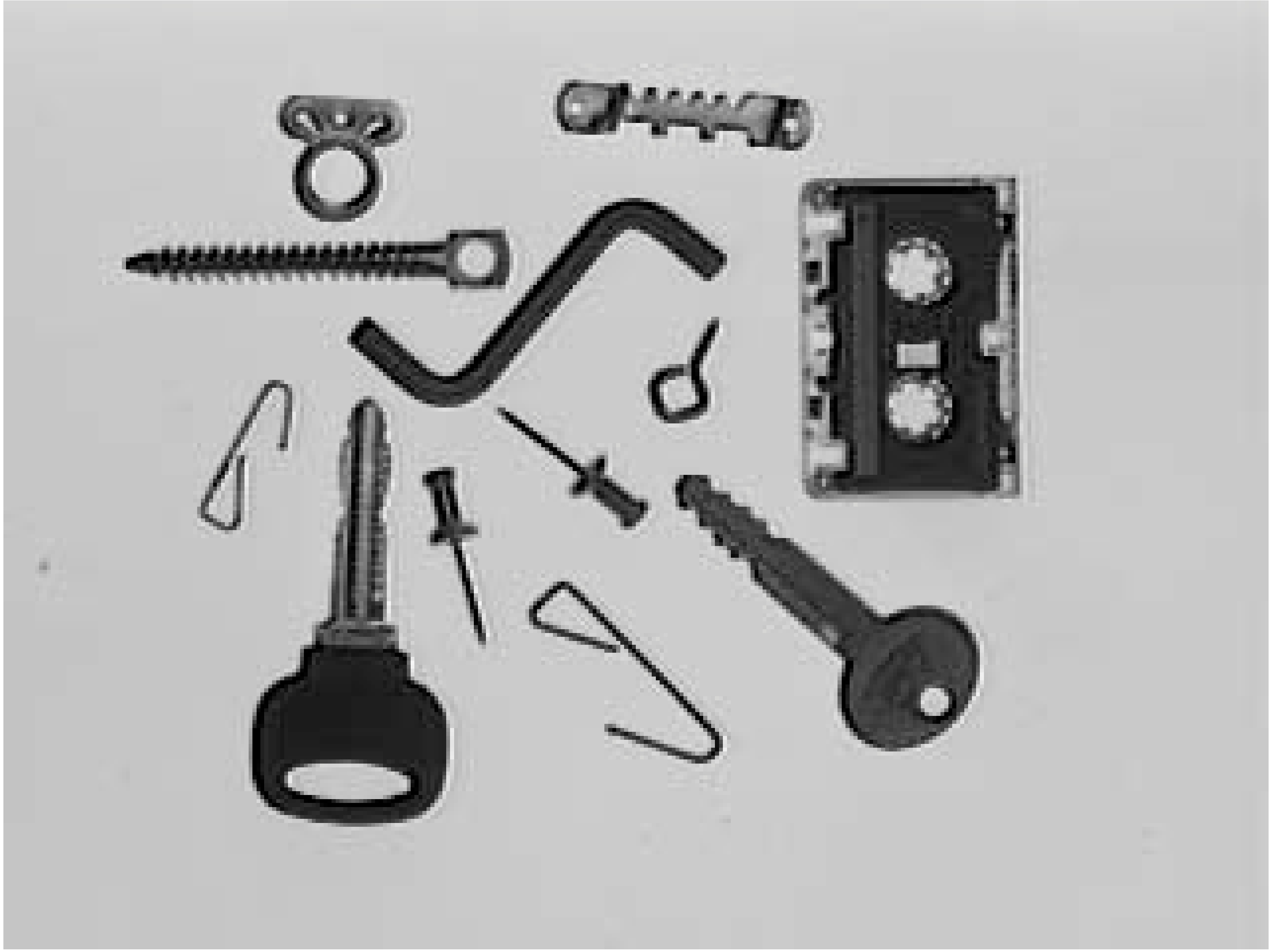
Segmentation, Example 4: Easy or Not? (2/2)





Where is the border between the cup and the background?

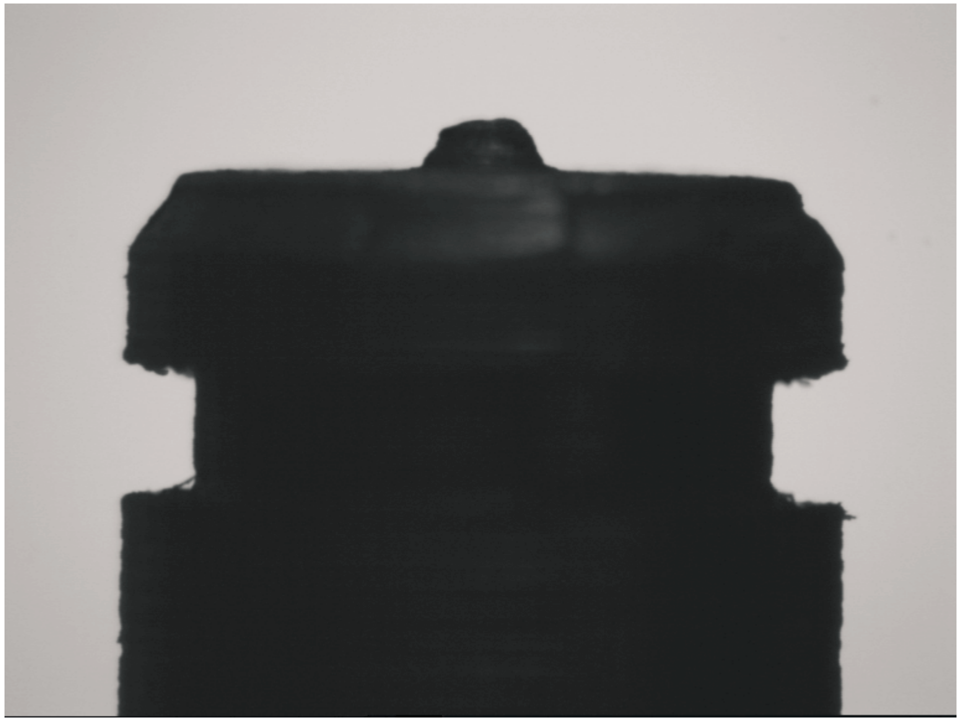
Segmentation, Example 4: Easy or Not? (2/2)

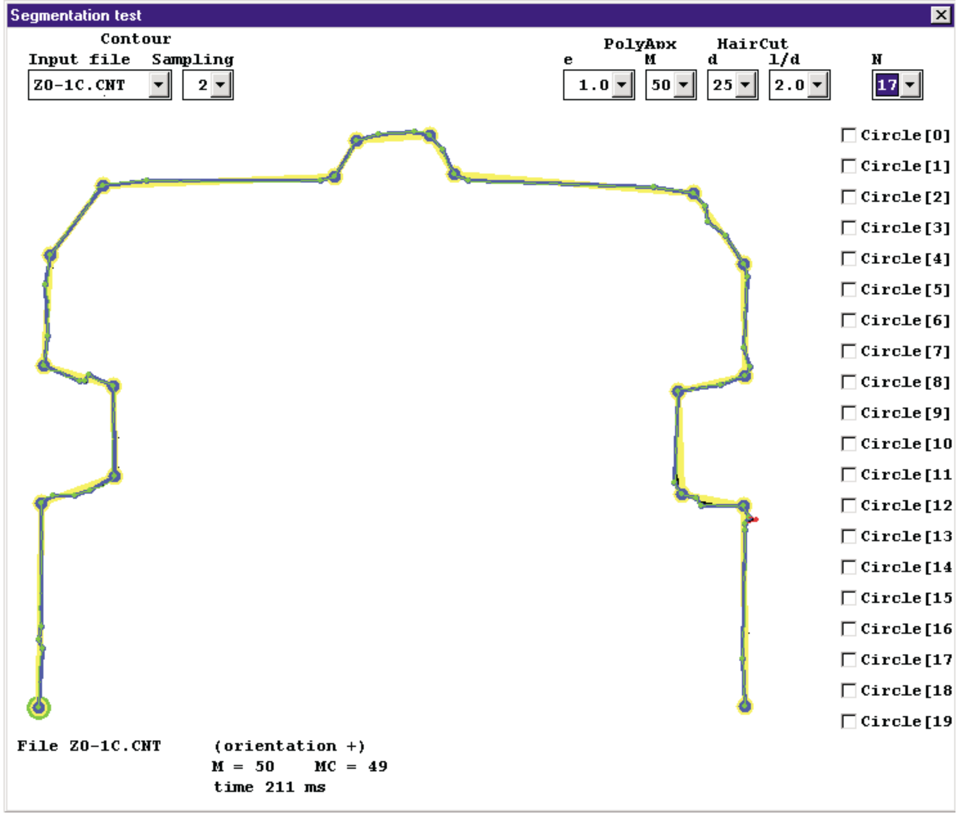


Segmentation, Example 5

Back-illuminated object. Example from the lathe-turning: *Courtesy: Neovision s.r.o.*

- Left image – a the back-illuminated detail showing imperfection and dust.
- Right image – automatically approximated shape allowing for automatic gauge check against a technical drawing.





Segmentation test

Contour
Input file: ZO-1C.CNT
Sampling: 2

PolyApx
e: 1.0
M: 50

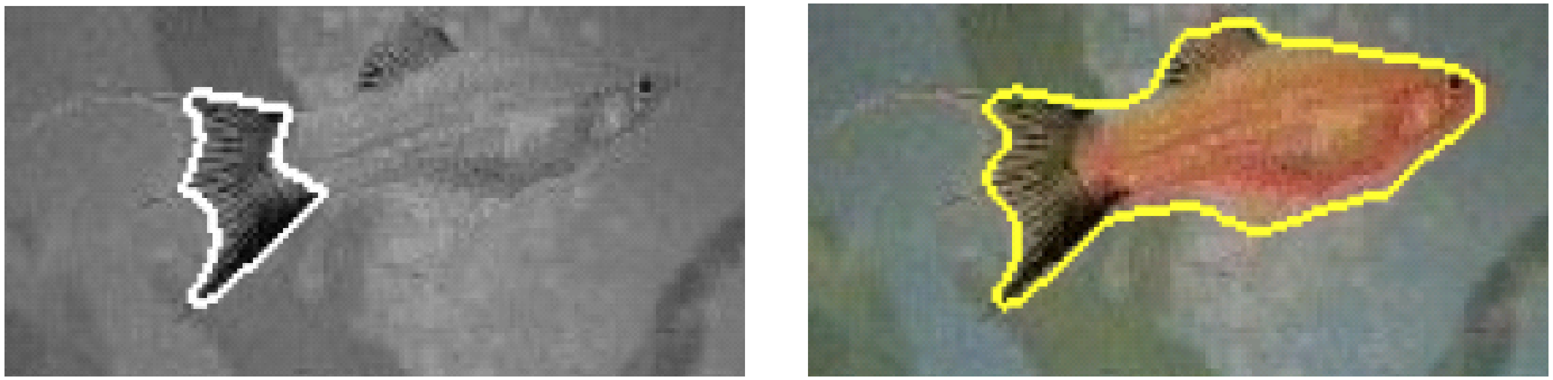
HairCut
d: 25
l/d: 2.0

N: 17

```
File ZO-1C.CNT      (orientation +)
                    M = 50    MC = 49
                    time 211 ms
```

Image Segmentation

- Collection of methods for “grouping pixels to larger sets”.
- Often: binary classification (e.g., foreground/background, object/the rest)
- Easy for humans
- No “general” ground truth. The desired result is always application-dependent.



Courtesy, images: Thomas Brox, TU Dresden, 2008

Complete Segmentation

The “Holy Grail”.

Complete segmentation divides an image into non-overlapping regions that match to the real world objects.

Complete segmentation is a partition. It divides an image R into the finite number S of regions $R_1, \dots, R_S$

$$R = \bigcup_{i=1}^{S} R_i \, , \qquad R_i \cap R_j = \emptyset \, , \qquad i \neq j \, .$$

Use of a priori information

Examples:

- Assumed shape of the region.
- Required position, orientation.
- Known initial and final point of the boundary *(e.g., in the application analyzing shape of a polymer drop, where the polymer sample comes out a tube of known position).*
- Relation of the region considered to other regions with required properties *(e.g., above, inside).*

Examples, two application areas:

Remote sensing: Look for ships in the water. Typical properties of railway lines, highways (minimal curvatures). Rivers do not cross.

Medical: Blood vessels ar roughly parallel. Relate to anatomic atlas (model-based approach).

Simple approaches for segmentation

- Local, use local property (intensity, color, texture), the global knowledge is represented by its distribution.
- Spatial coherence (≈ clustering of 'tokens'). Proximity of pixels is taken into account.
  - Connecting, e.g., edgels because edges bear often an important information about objects.
  - Region merging/splitting. Regions come from aggregating pixels with similar properties (homogeneity criterion)
- Template matching – detection and fitting tokens in the image to a known template.
  - Parametric model detection, e.g., straight line, circle, ellipse, . . .

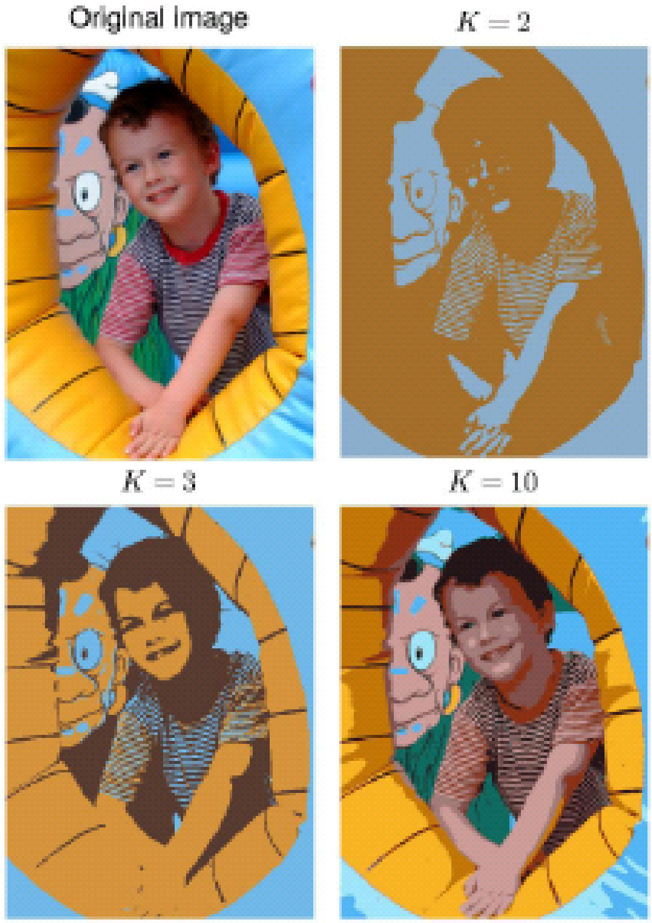
Segmentation as a Clustering Problem

- Some formulation of image segmentation lead are equivalent to clustering methods which are known in statistics and machine learning.
- The approach is to treat a pixel along with its properties as a point in a D-dimensional space, and then find representation of the distribution of these points.
- In this lecture: K-means, EM algorithm.

K-means clustering

- Input: K – the required number of clusters, $\{x_i, i = 1, 2, ..., N\}$ – the data points
- Objective: Find cluster centers $\{\mu_k, k = 1, 2, ..., K\}$ and pixel partition $\{\mathcal{T}_k, k = 1, 2, ..., K\}$ which minimize
$$J = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in \mathcal{T}_k} \|x_i - \mu_k\|^2 ,$$
- Initialize each point as belonging to a random cluster out of the K clusters.

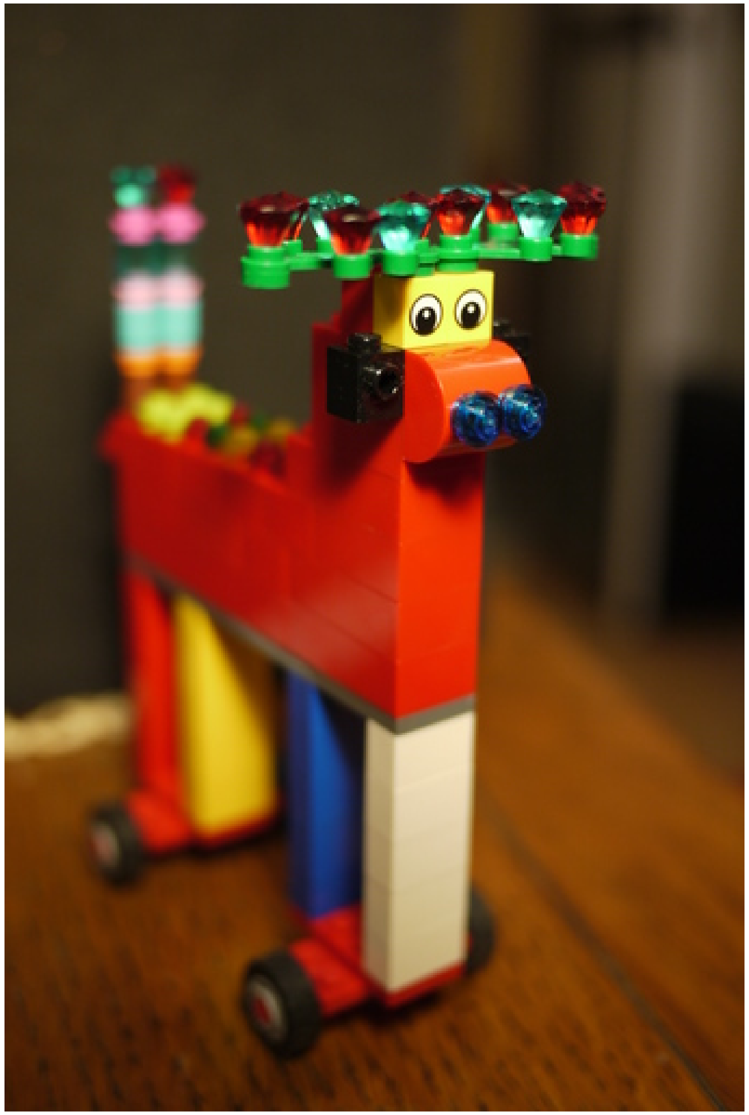
  1. Compute the mean feature vector, μ_k, for each cluster, $k = 1, 2, ..., K$
  2. For each point: assign the point to its closest cluster.
  3. If pixel assignments changed, go to 1., otherwise finish.

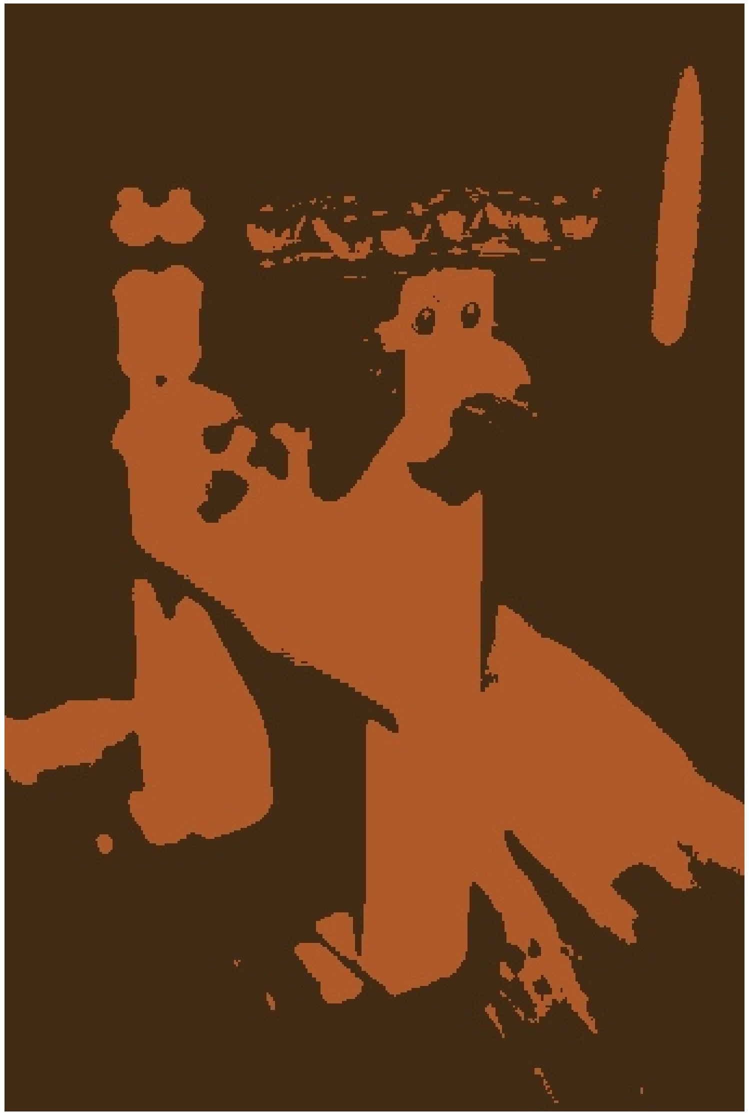


Author: Christopher Bishop

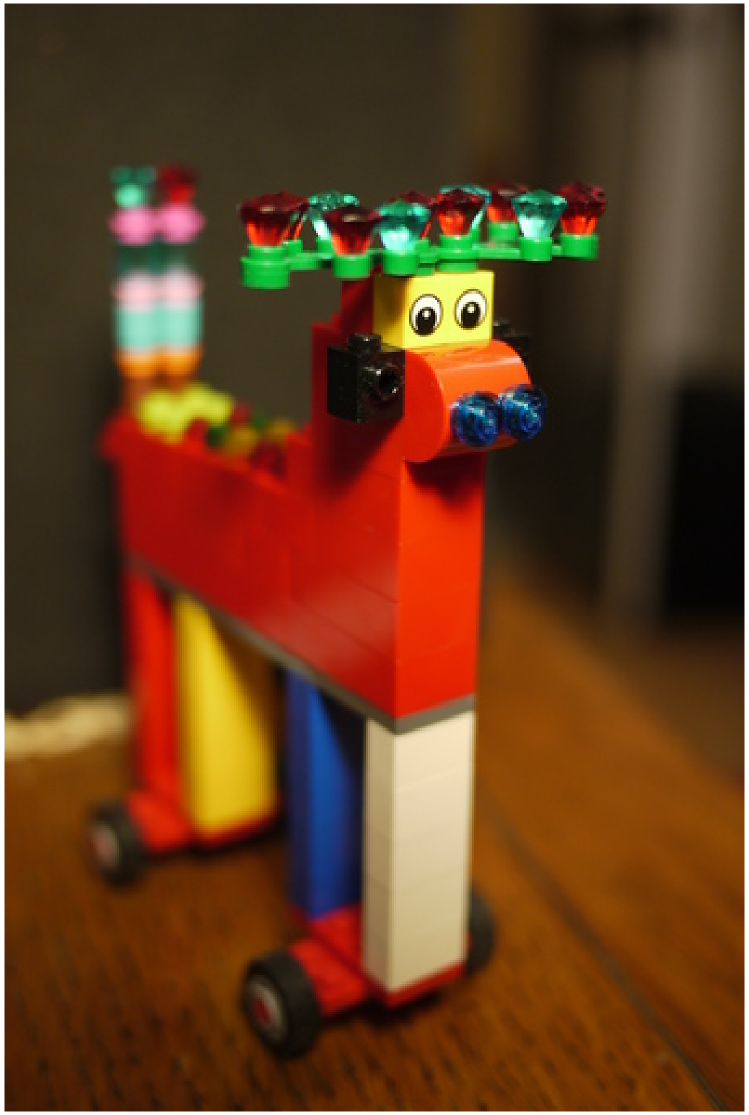
S. Lloyd, Last square quantization in PCM's. Bell Telephone Laboratories Paper (1957). In the journal much later: S. P. Lloyd. Least squares quantization in PCM. Special issue on quantization, IEEE Trans. Inform. Theory, 28:129–137, 1982.

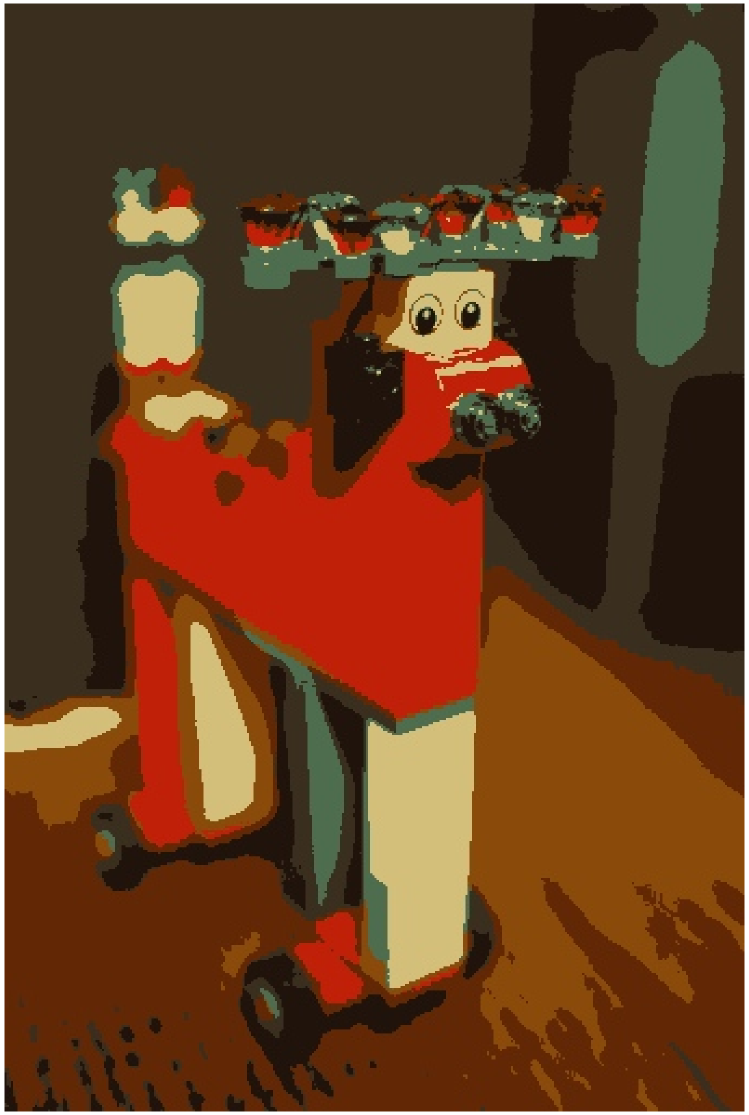
K-means, example 1, K=2



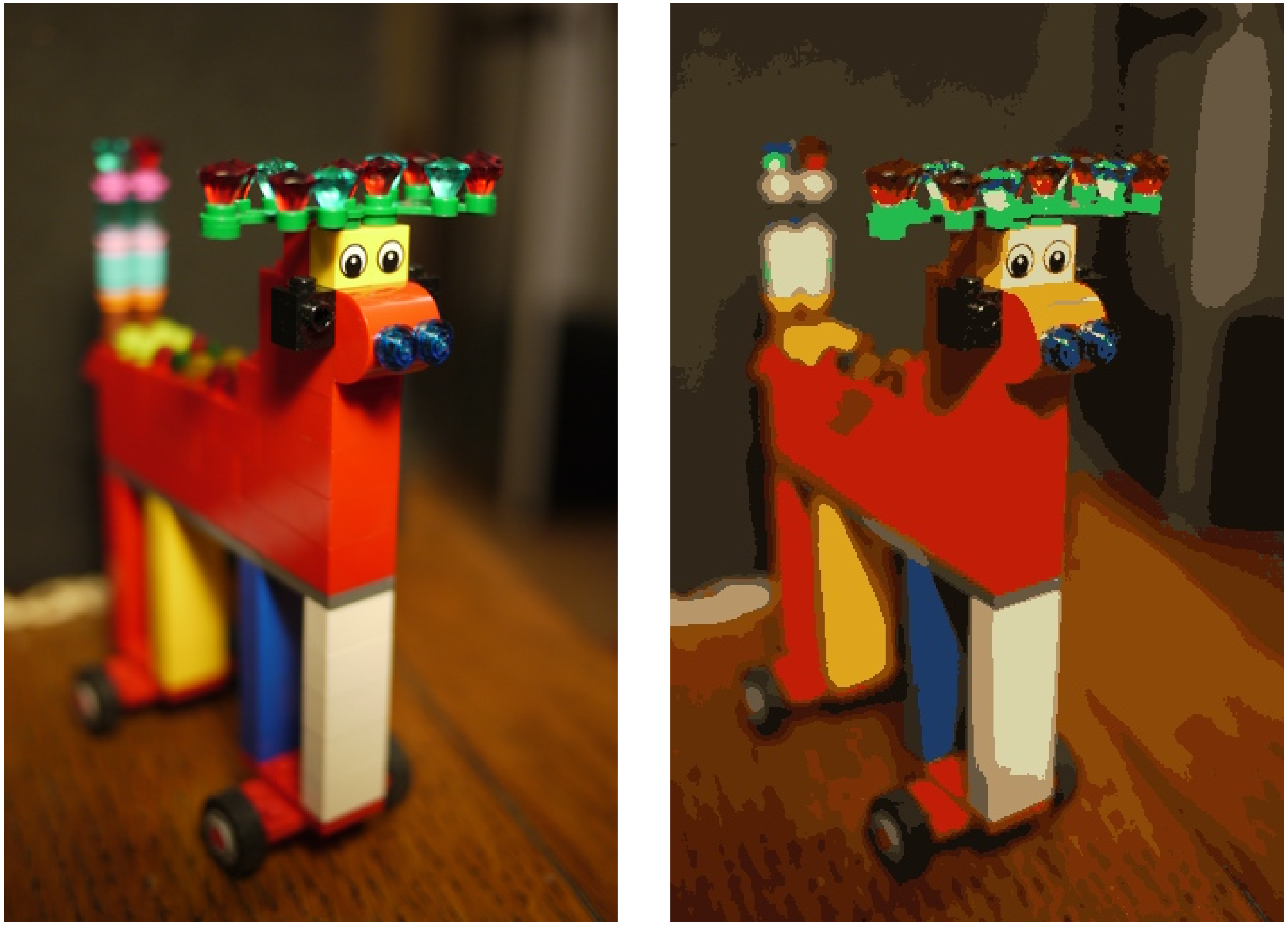


K-means, example 1, K=7

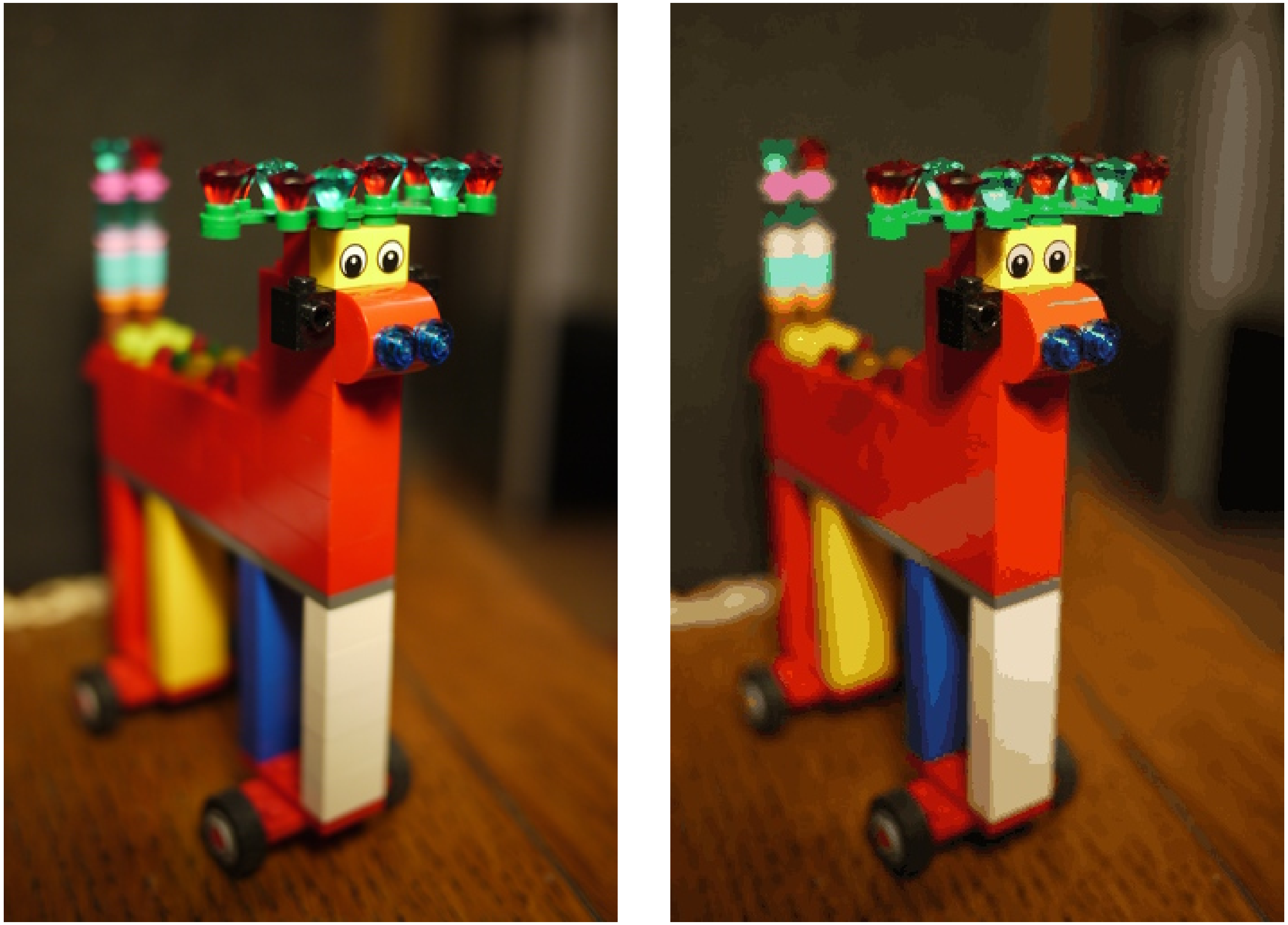




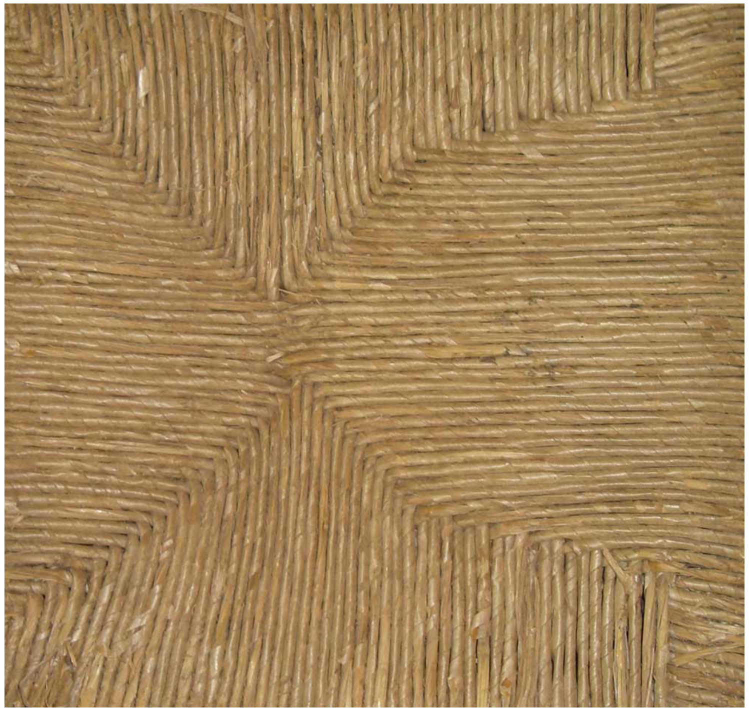
K-means, example 1, K=15

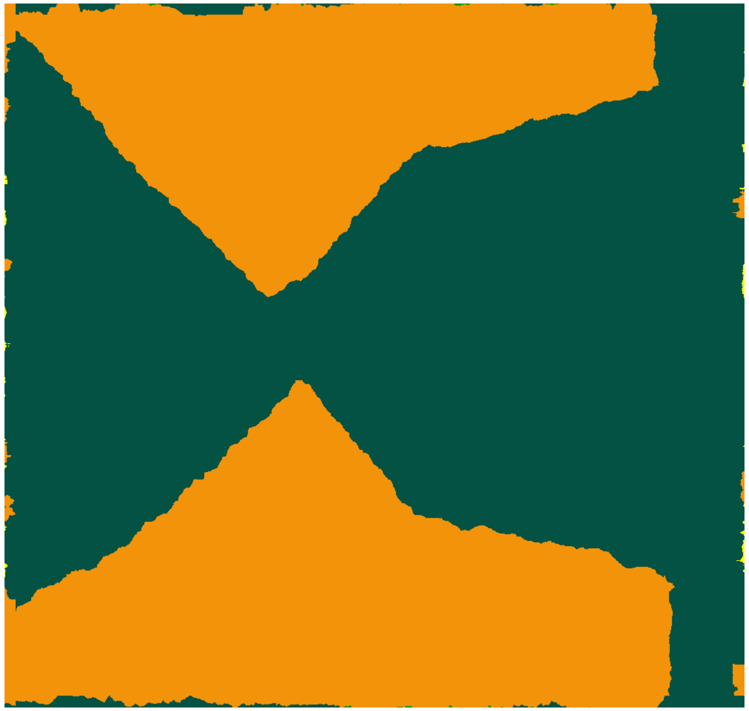


K-means, example 1, K=64



K-means, Example 2





- feature: absolute value of partial derivatives, $(|\frac{\partial I}{\partial x}|, |\frac{\partial I}{\partial y}|)$
- $K = 2$

K-means, Notes

- Local method, thus finding the global minimum of $J(\cdot)$ is not guaranteed.
- The objective function $J(\cdot)$ decreases or stays the same in each step.
- The algorithm finishes after a finite number of steps.
- For $K = 2$ and data dimensionality 1, K-means essentially finds a threshold which divides the data into two clusters.

EM algorithm with Gaussian Mixtures (1/2)

- Distribution of points $p(x)$ will approximated by a Gaussian Mixture:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \right\}$$

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k,\Sigma_k)$$

such that the log-likelihood of the dataset is minimised:

$$\log p(\{x_1..x_N\}|\{\pi_k,\mu_k,\Sigma_k,k=1..K\}) = \sum_{i=1}^{N} \log\left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k,\Sigma_k) \right\}$$

- The optimization parameters are μ_k, Σ_k, π_k for each Gaussian node ($\sum_k \pi_k = 1$).

EM algorithm with Gaussian Mixtures (2/2)

- Note that in K-means, point assignments $e_k(x_i)$ of x_i to cluster K are hard ($e_k(x_i) \in \{0, 1\}$, $\sum_k e_k(x_i) = 1$)
- Here, assignment of points to Gaussian nodes are soft:

$$e_k(x_i) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}$$

- They are used as weights for computing the optimized parameters:

$$\mu_k = \frac{\sum_{i=1}^{N} e_k(x_i) x_i}{\sum_{i=1}^{N} e_k(x_i)}, \qquad \Sigma_k = \frac{\sum_{i=1}^{N} e_k(x_i)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} e_k(x_i)},$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} e_k(x_i)$$

- The algorithm iterates in a loop, computing the weights (E-step) and the Gaussian nodes parameters (M-step)

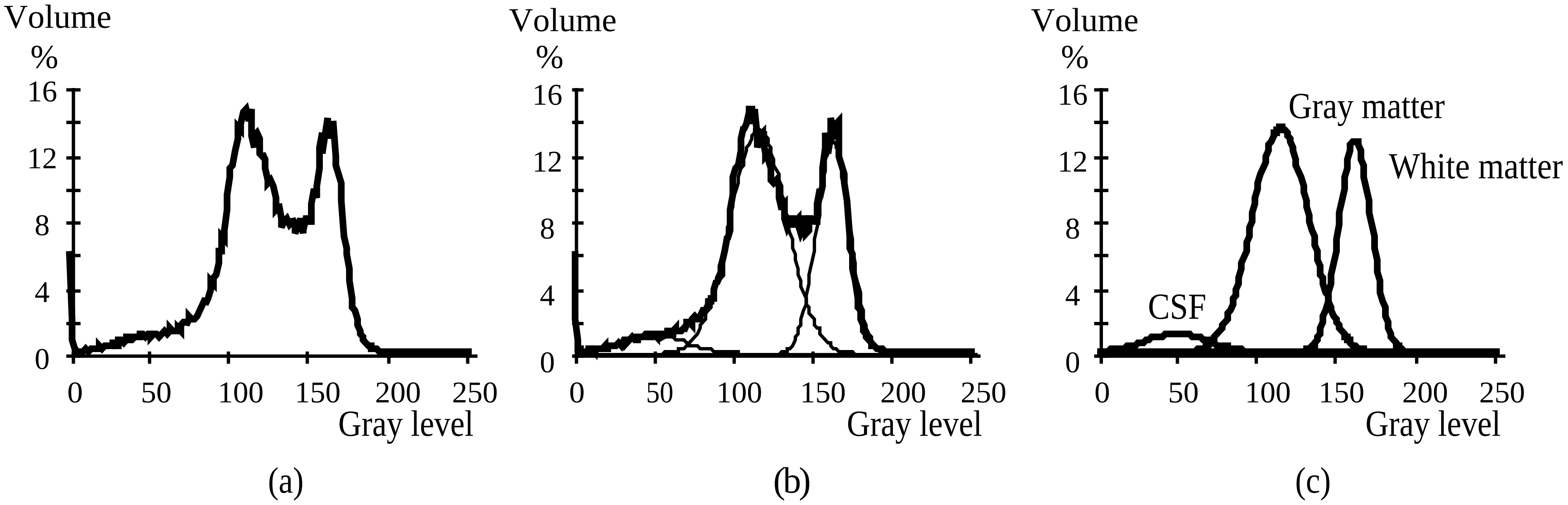
Example, Segmentation of the brain MR

- data = image intensities

$$p(x) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} = \sum_{k=1}^{K} \pi_k N(x, \mu_k, \sigma_k) \, .$$

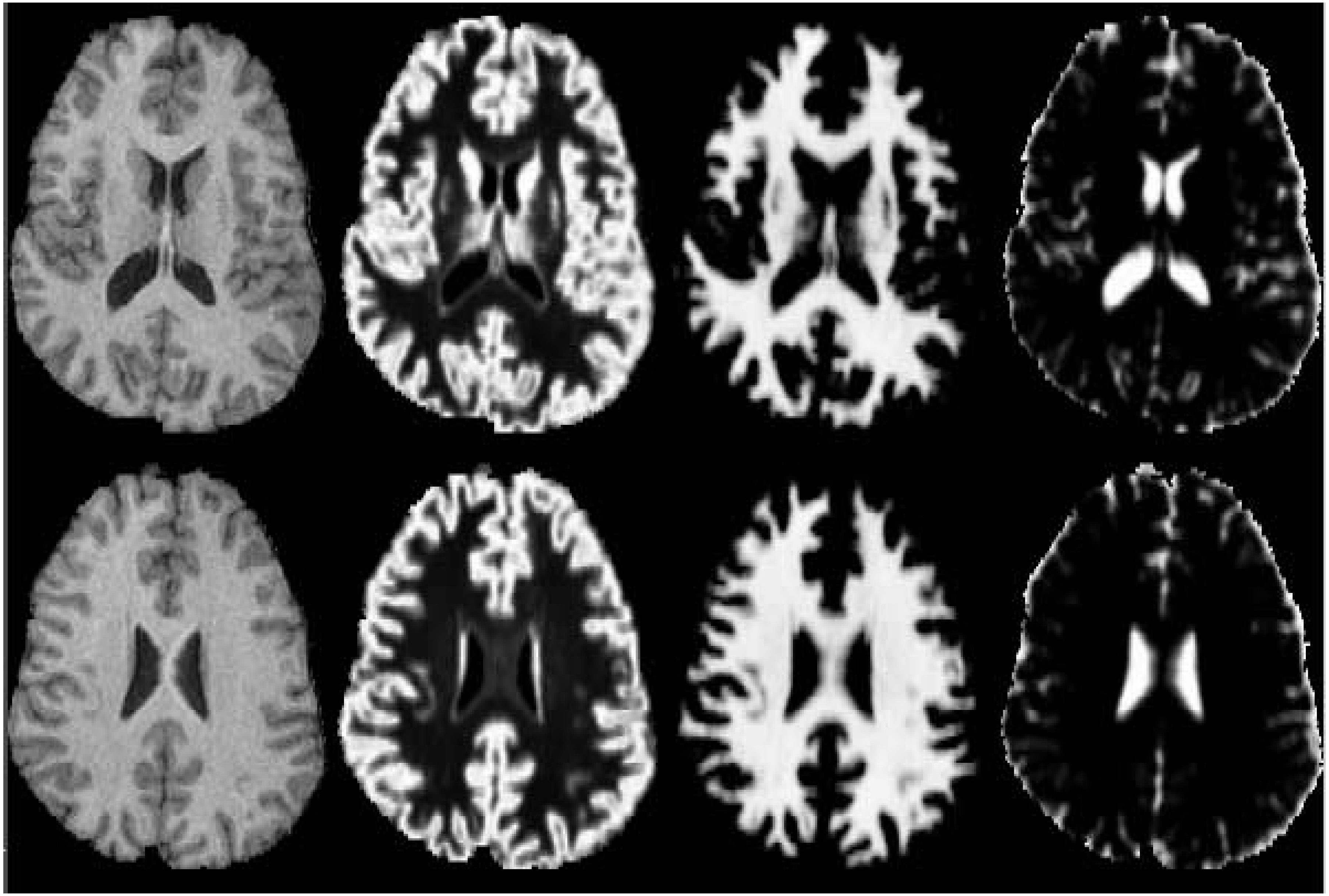
Example, Segmentation of the brain MR

- Input: T1-weighted NMR images.
- Desired classes: white matter, grey matter, celebro-spinal fluid (CSF)



Courtesy: Milan Šonka, University of Iowa.

Brain MR, Segmentation result



original gray matter white matter CSF

Note

- K-means clustering can be thought of as a simplified EM algorithm with Gaussian mixtures
- all σ_i equal, all π_i equal $\Rightarrow$ only distance to μ_i plays a role
- the assignments are hard (binary)

Thresholding

Input image $f(i,j)$, output image $g(i,j)$.

For each pixel (i,j)

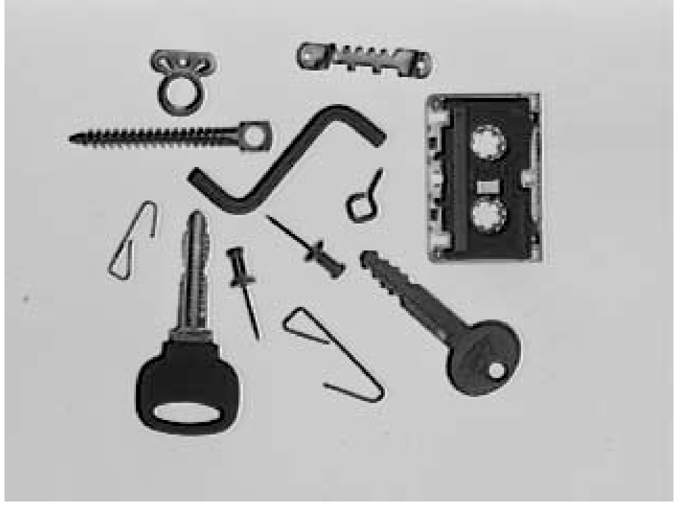
$$g(i,j) = \begin{cases} 1 & \text{for} \quad f(i,j) \geq \mathrm{Threshold}\,, \\ 0 & \text{for} \quad f(i,j) < \mathrm{Threshold}\,. \end{cases}$$

+ Simple technique, long time and more often used.

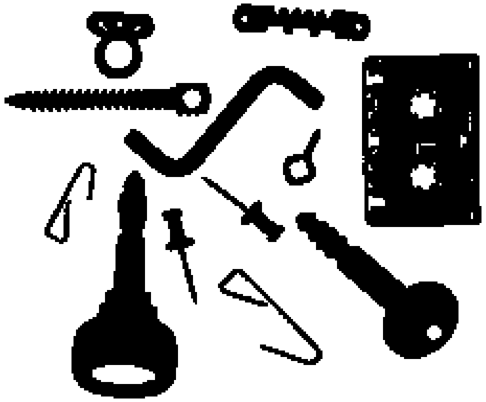
+ Easy in hardware, intrinsically parallel.

– Works only for subclass of images in which objects are distinct from background in intensity.

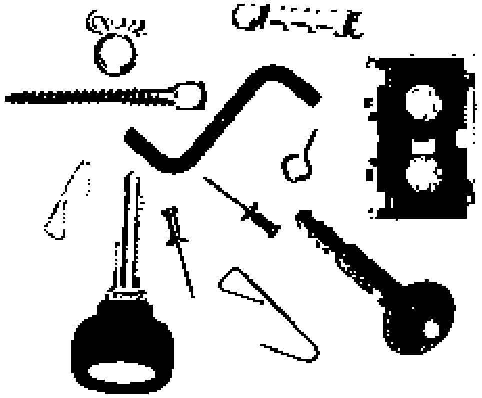
Example, dependence on threshold



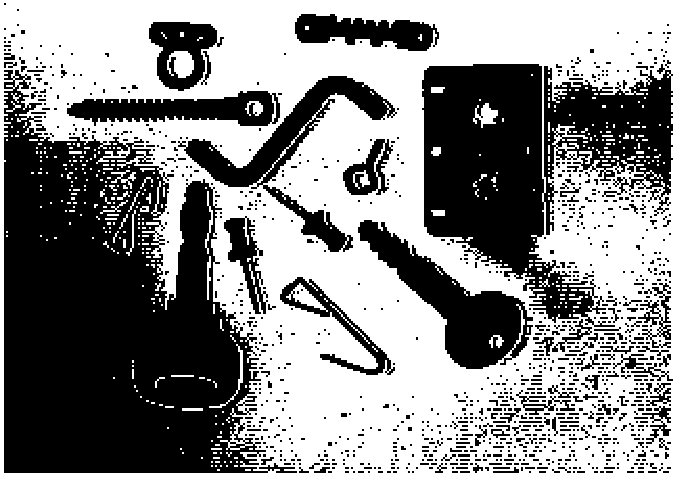
Original image.



Threshold segmentation.



Threshold too low.



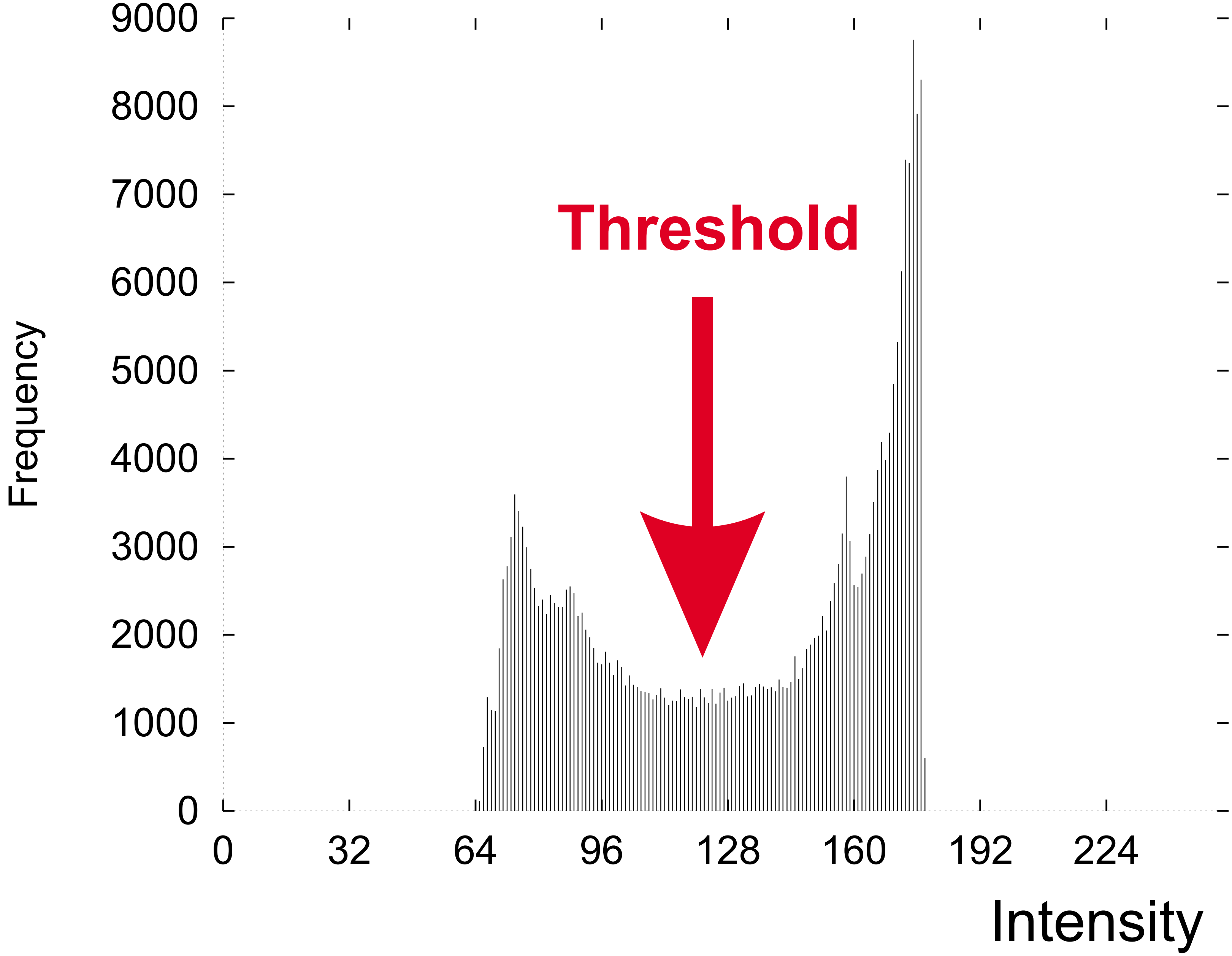
Threshold too high.

Threshold detection, use of the histogram

p-tile thresholding, if we know that the objects cover $1/p$ of the image, e.g. printed characters on a sheet $\Longrightarrow 1/p$ area of a histogram.

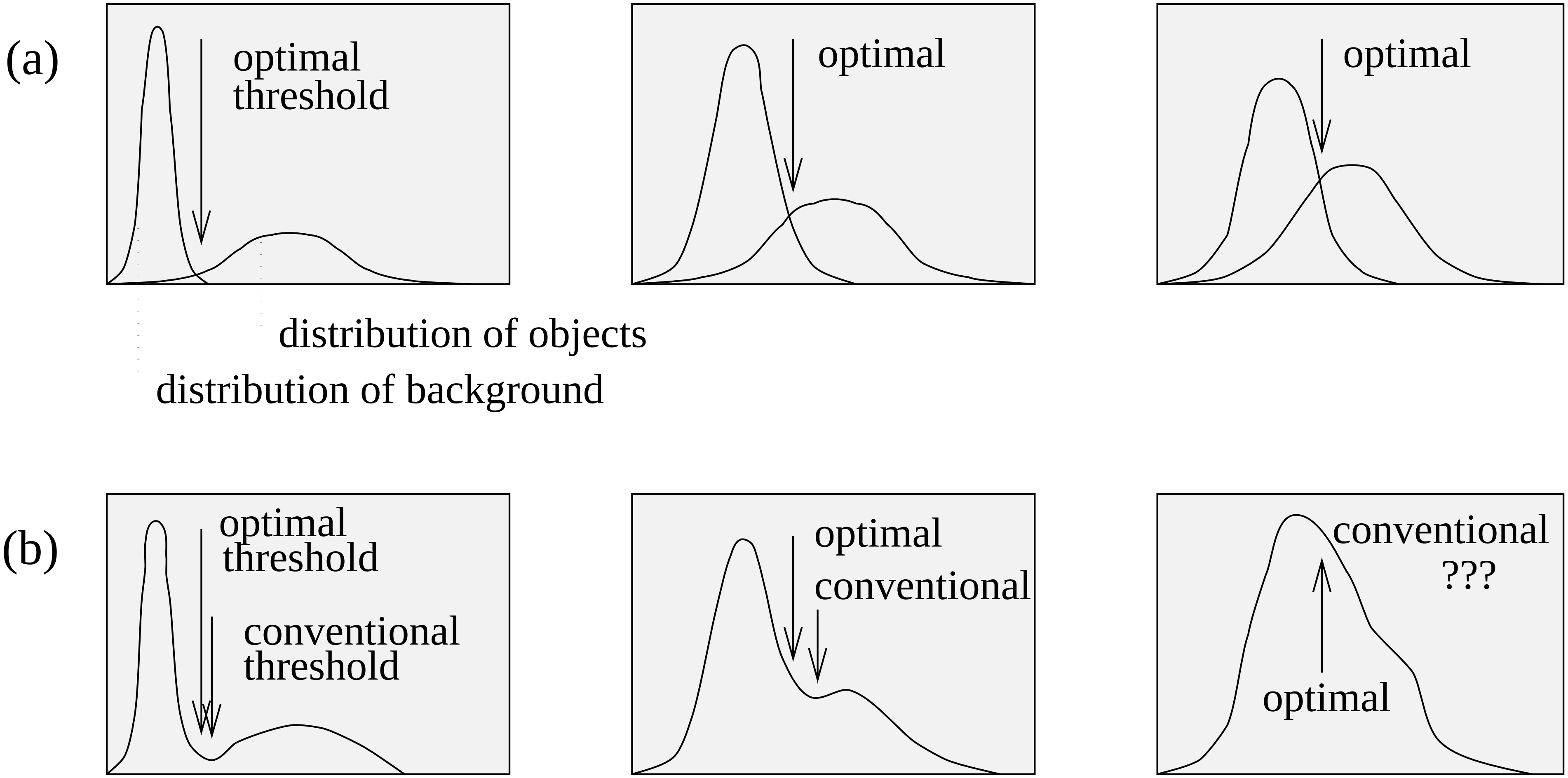
Histogram shape analysis, distinct objects on background correspond to a bi-modal histogram. Find the separating point between the modes.

Automatically found threshold according to a bi-modal histogram



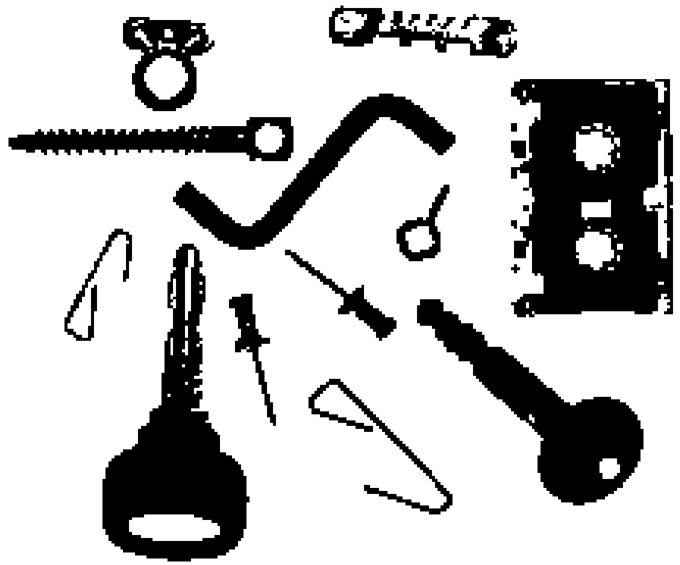
Optimal thresholding by a mixture of Gaussians

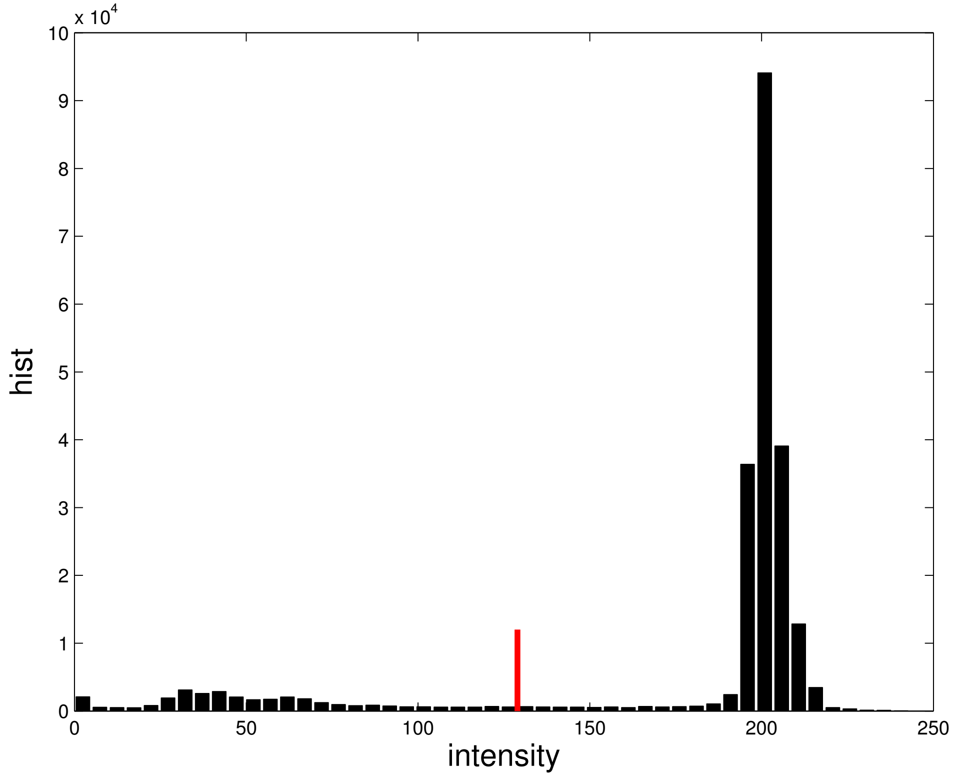
Motivation:



Finding the threshold by K-means

- Data are 1D, $K = 2$
- Result:



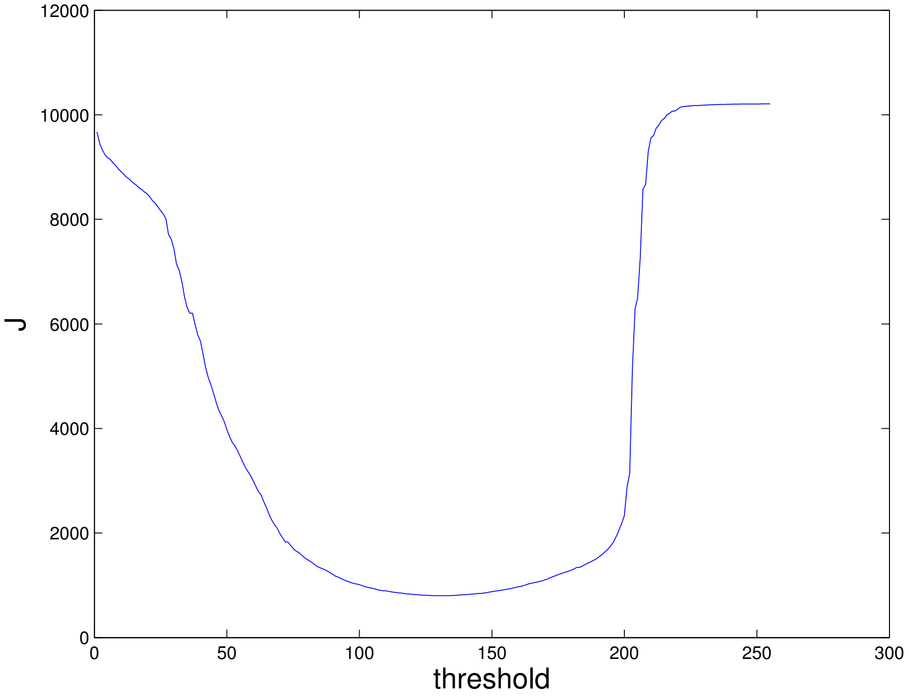


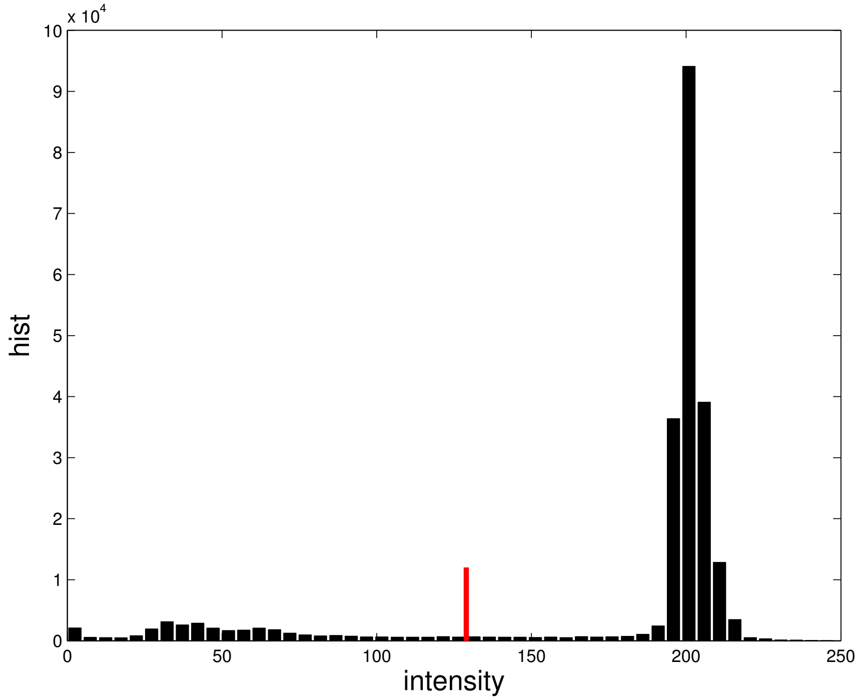
$K = 2$, 1D data

- Recall that the minimization criterion is

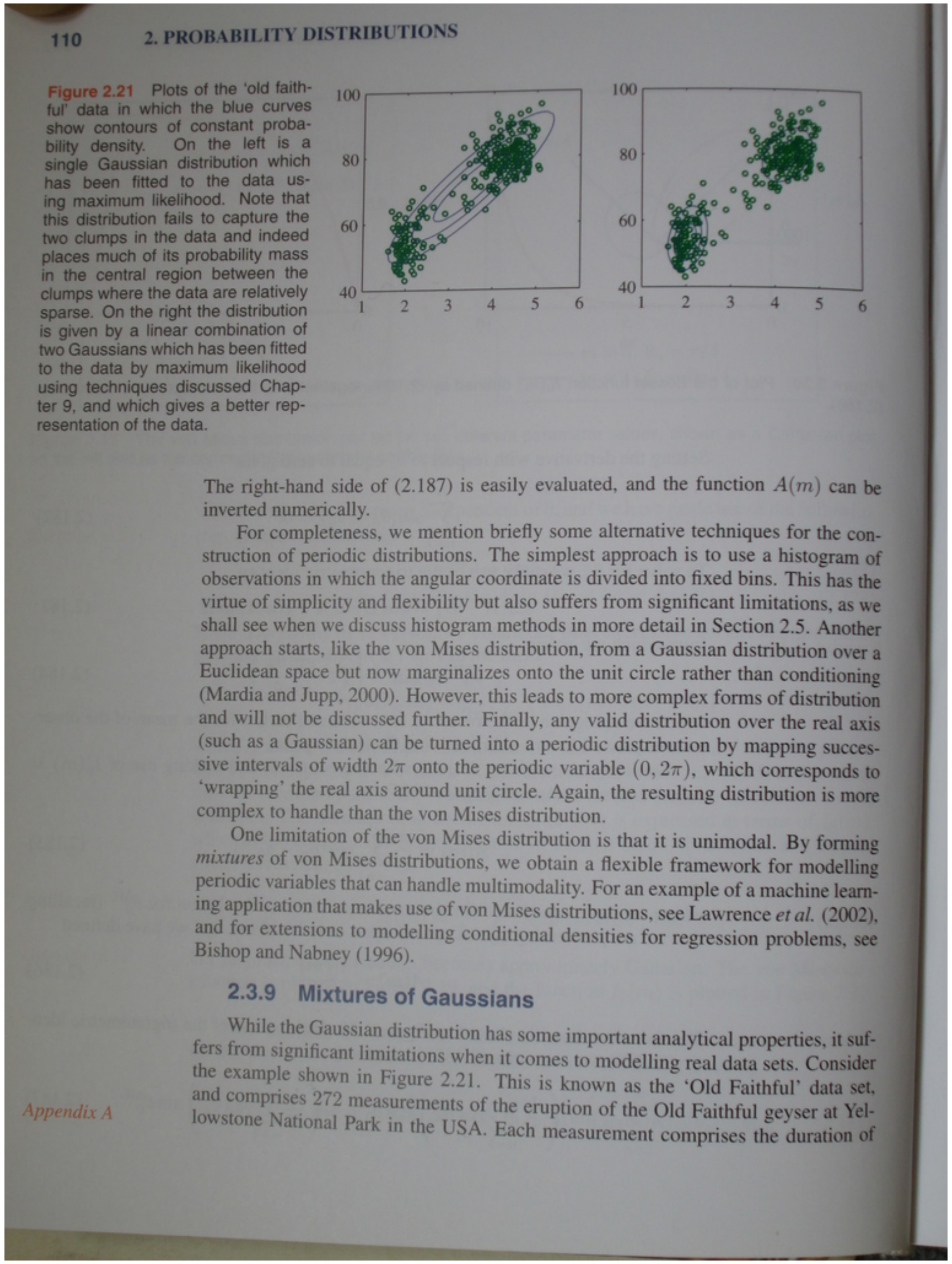
$$J = \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in \mathcal{T}_k} \|x_i - \mu_k\|^2 \,,$$

- We can easily evaluate this for all possible thresholds (finite number)
- This guarantees to find the *global* minimum.





Real world example


100
80
60
40
1
2
3
4
5
6
100
80
60
40
1
2
3
4
5
6

110 2. PROBABILITY DISTRIBUTIONS

Figure 2.21 Plots of the 'old faithful' data in which the blue curves show contours of constant probability density. On the left is a single Gaussian distribution which has been fitted to the data using maximum likelihood. Note that this distribution fails to capture the two clumps in the data and indeed places much of its probability mass in the central region between the clumps where the data are relatively sparse. On the right the distribution is given by a linear combination of two Gaussians which has been fitted to the data by maximum likelihood using techniques discussed Chapter 9, and which gives a better representation of the data.

The right-hand side of (2.187) is easily evaluated, and the function $A(m)$ can be inverted numerically.

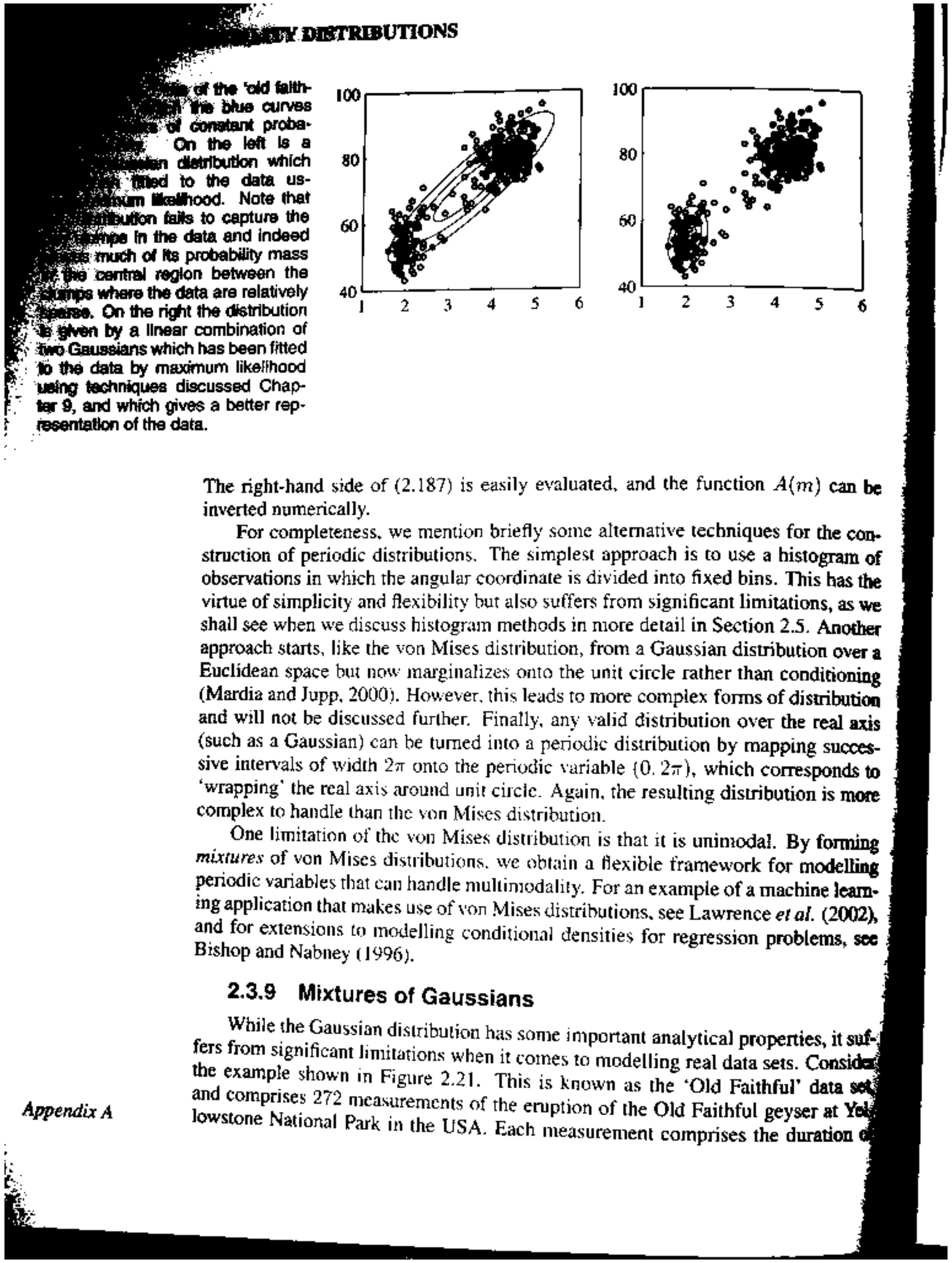
For completeness, we mention briefly some alternative techniques for the construction of periodic distributions. The simplest approach is to use a histogram of observations in which the angular coordinate is divided into fixed bins. This has the virtue of simplicity and flexibility but also suffers from significant limitations, as we shall see when we discuss histogram methods in more detail in Section 2.5. Another approach starts, like the von Mises distribution, from a Gaussian distribution over a Euclidean space but now marginalizes onto the unit circle rather than conditioning (Mardia and Jupp, 2000). However, this leads to more complex forms of distribution and will not be discussed further. Finally, any valid distribution over the real axis (such as a Gaussian) can be turned into a periodic distribution by mapping successive intervals of width 2π onto the periodic variable $(0, 2\pi)$, which corresponds to 'wrapping' the real axis around unit circle. Again, the resulting distribution is more complex to handle than the von Mises distribution.

One limitation of the von Mises distribution is that it is unimodal. By forming *mixtures* of von Mises distributions, we obtain a flexible framework for modelling periodic variables that can handle multimodality. For an example of a machine learning application that makes use of von Mises distributions, see Lawrence *et al.* (2002), and for extensions to modelling conditional densities for regression problems, see Bishop and Nabney (1996).

2.3.9 Mixtures of Gaussians

Appendix A

While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modelling real data sets. Consider the example shown in Figure 2.21. This is known as the 'Old Faithful' data set, and comprises 272 measurements of the eruption of the Old Faithful geyser at Yellowstone National Park in the USA. Each measurement comprises the duration of


100
80
60
40
1
2
3
4
5
6
100
80
60
40
1
2
3
4
5
6

DISTRIBUTIONS

of the 'old faithful' data in which the blue curves show contours of constant probability density. On the left is a single Gaussian distribution which has been fitted to the data using maximum likelihood. Note that this distribution fails to capture the two clumps in the data and indeed places much of its probability mass in the central region between the clumps where the data are relatively sparse. On the right the distribution is given by a linear combination of two Gaussians which has been fitted to the data by maximum likelihood using techniques discussed Chapter 9, and which gives a better representation of the data.

The right-hand side of (2.187) is easily evaluated, and the function $A(m)$ can be inverted numerically.

For completeness, we mention briefly some alternative techniques for the construction of periodic distributions. The simplest approach is to use a histogram of observations in which the angular coordinate is divided into fixed bins. This has the virtue of simplicity and flexibility but also suffers from significant limitations, as we shall see when we discuss histogram methods in more detail in Section 2.5. Another approach starts, like the von Mises distribution, from a Gaussian distribution over a Euclidean space but now marginalizes onto the unit circle rather than conditioning (Mardia and Jupp, 2000). However, this leads to more complex forms of distribution and will not be discussed further. Finally, any valid distribution over the real axis (such as a Gaussian) can be turned into a periodic distribution by mapping successive intervals of width 2π onto the periodic variable $(0, 2\pi)$, which corresponds to 'wrapping' the real axis around unit circle. Again, the resulting distribution is more complex to handle than the von Mises distribution.

One limitation of the von Mises distribution is that it is unimodal. By forming *mixtures* of von Mises distributions, we obtain a flexible framework for modelling periodic variables that can handle multimodality. For an example of a machine learning application that makes use of von Mises distributions, see Lawrence *et al.* (2002), and for extensions to modelling conditional densities for regression problems, see Bishop and Nabney (1996).

2.3.9 Mixtures of Gaussians

Appendix A

While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modelling real data sets. Consider the example shown in Figure 2.21. This is known as the 'Old Faithful' data set, and comprises 272 measurements of the eruption of the Old Faithful geyser at Yellowstone National Park in the USA. Each measurement comprises the duration

Conclusion

- Real world example, possible approaches.
- Recap.