CS5487 Programming Report 2
Clustering

Zhiyuan Yang
Department of Computer Science
City University of Hong Kong

Implement and test several clustering algorithms on both synthetic and real data.

Problem 1 Clustering synthetic data

(a) Implement the three clustering algorithms K-means, EM-GMM, Mean-shift: see the code.

(b) Run the algorithms on the three synthetic datasets
on dataA : see Figure 1, mean-shift divided into 7 categories
on dataB : see Figure 2, mean-shift divided into 4 categories
on dataC : see Figure 3, mean-shift divided into 8 categories

To judge each clustering algorithm performance on each dataset, I firstly find a point in each of the four types to determine the classification category and secondly rename their categories to make they are same with ground truth, and finally make a quantitative judgment, the result is be seen in Table 1 2 3.

on dataA: K-mean $>$ EM-GMM $>$ Mean-shfit
on dataB: EM-GMM $=$ Mean-shfit $>$ K-mean
on dataC: EM-GMM $>$ Mean-shfit $>$ K-mean

(c) How sensitive is mean-shift to the bandwidth parameter h: see Figure 4 and see Table 4
It's easy to find that when h=2, mean-shift perform best, there is no wrong cluster.

Table 1: dataA cluster performance

clustering algorithms	wrong cluster sum
K-means	2
EM-GMM	6
Mean-shift	14

Table 2: dataB cluster performance

clustering algorithms	wrong cluster sum
K-means	120
EM-GMM	0
Mean-shift	0

Figure 1: dataA

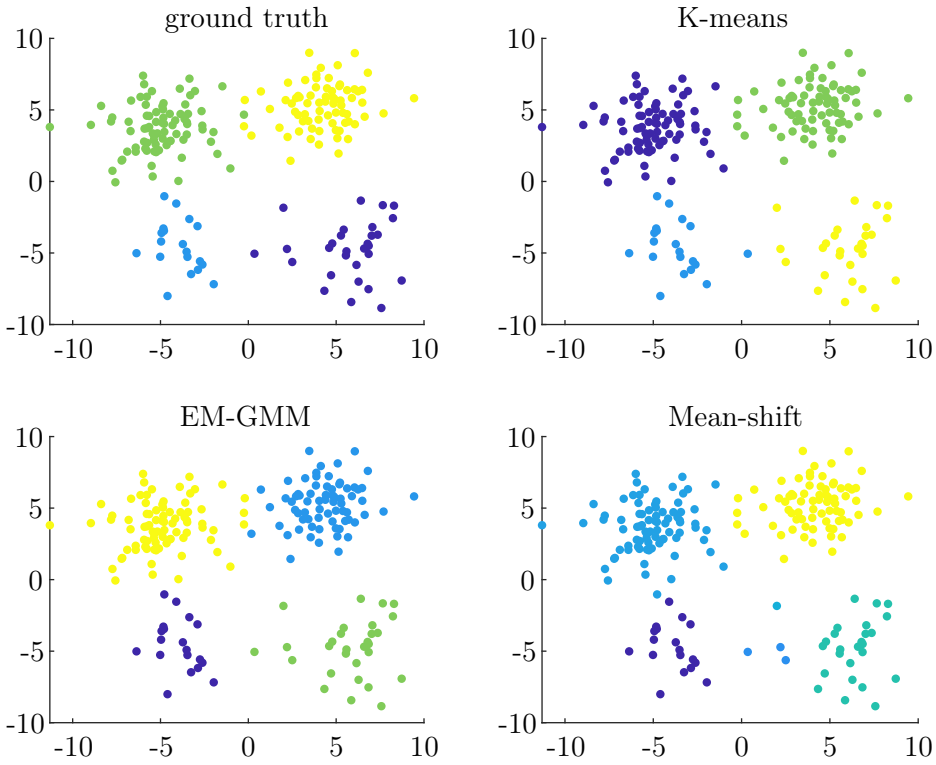


Figure 2: dataB

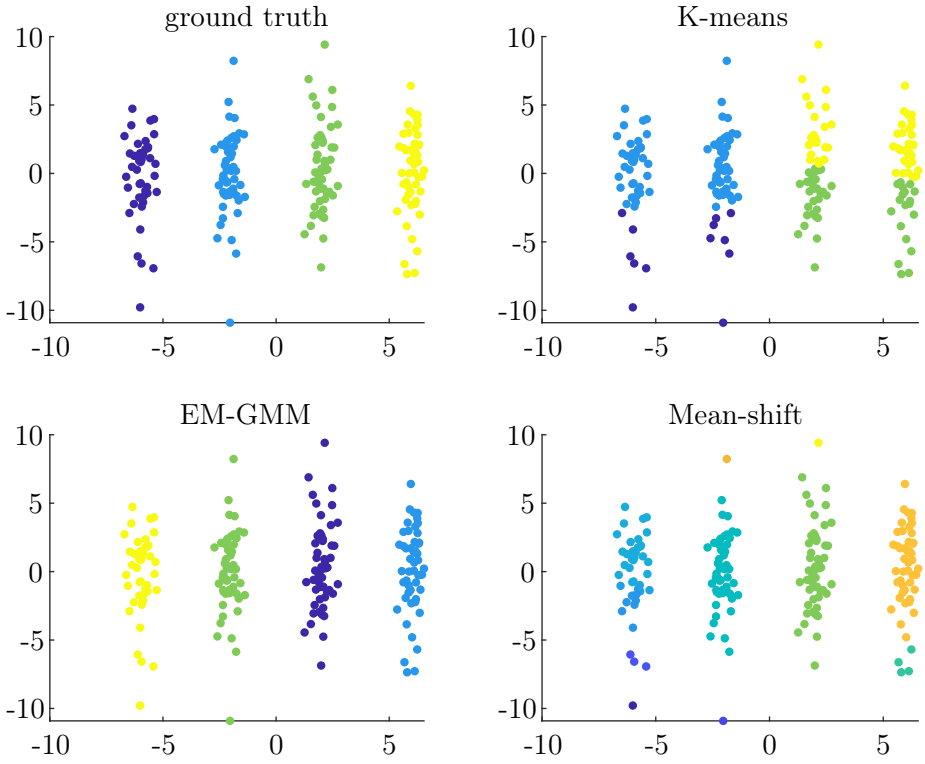


Figure 3: dataC

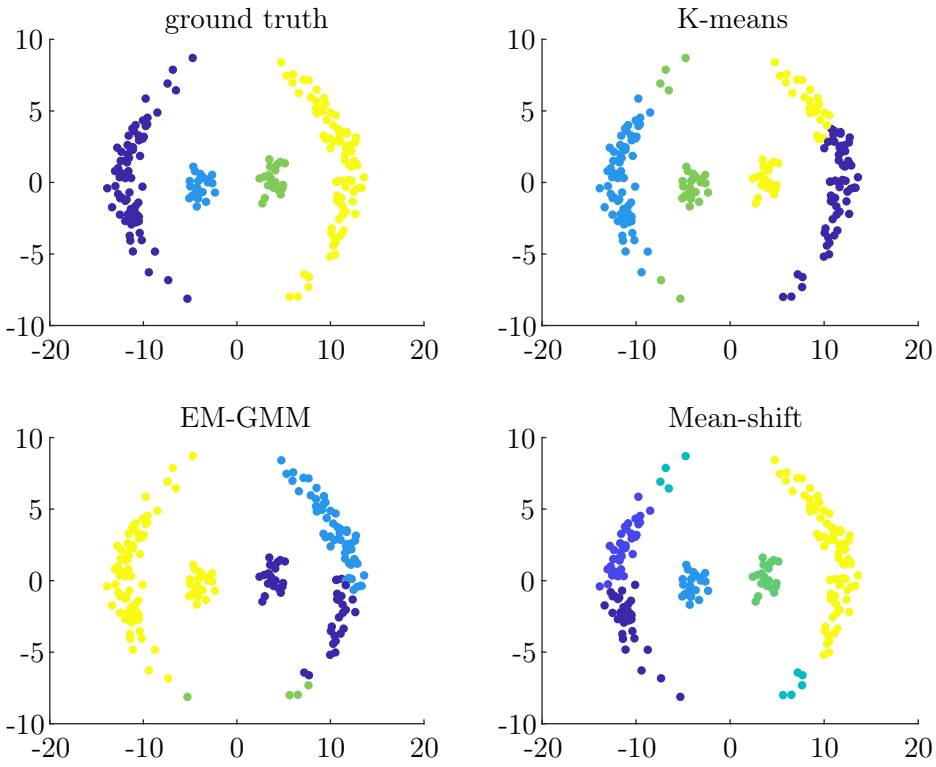


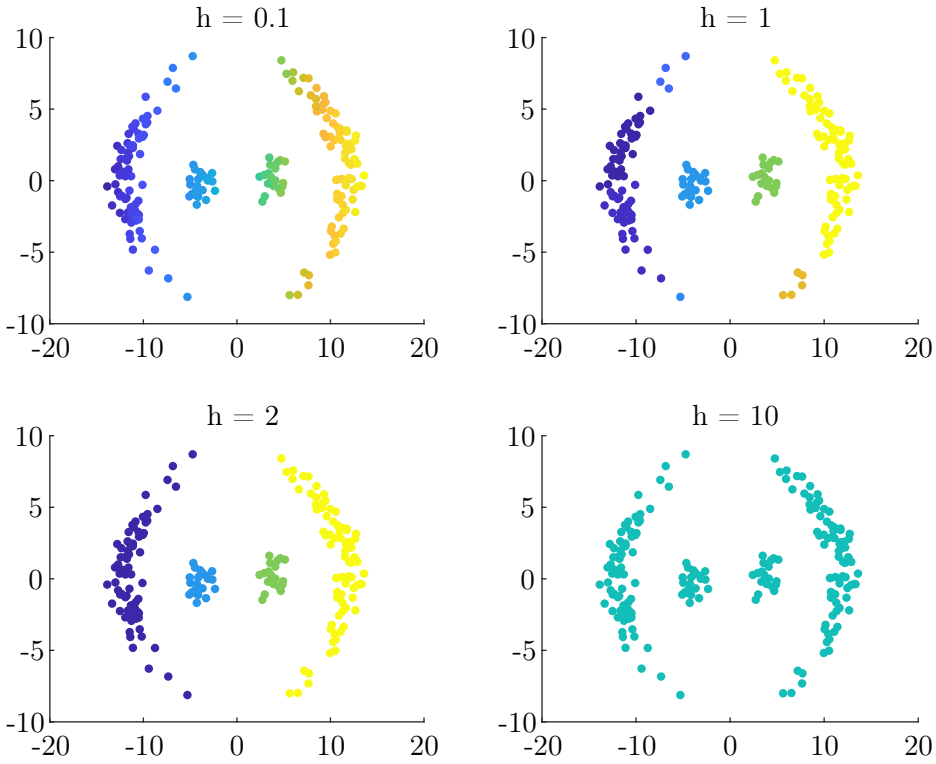
Table 3: dataC cluster performance

clustering algorithms	wrong cluster sum
K-means	52
EM-GMM	25
Mean-shift	40

Table 4: mean-shift sensitive to the bandwidth h

h	wrong cluster sum
0.1	160
1	40
2	0
10	115

Figure 4: mean-shift sensitive to the bandwidth h



A real world clustering problem image segmentation

(a) Use the three clustering algorithms to segment a few of the provided images. I choose two picture from the folder to verify my algorithm, see Figure 5 6 7. Change another picture and see Figure 8 9 10

By these two image segmentations, I can get a solution that the algorithm of mean-shift divide the picture into more than 2 parts, it is not easy to compare it with other two algorithm, and the algorithm of EM-GMM is better than K-means.

When I change different K in KM and GMM, see Figure 11 and Figure 12

When I change different h in Mean-shift, see Figure 13

I can see that the image will be closer to the original picture accompany with the increase of K values but it lost the property of abstraction of segmentation. However, just see the K-means, when I increase the number of K, it still divides the image into two parts, maybe this is its best divide result. When it comes to h, larger h values make the output picture more abstract and are divided into less clusters. Small h values produce a blurry version of the original one and may considered to be an anti-noise processing.So MS is more sensitive.

(b) Modify your K-means and mean-shift implementations to allow different feature scaling. I have combine these two questions into one, you can also see question (a) figure. By compare each algorithm, I find that new distance perform better than the old one.

Figure 5: kmeans

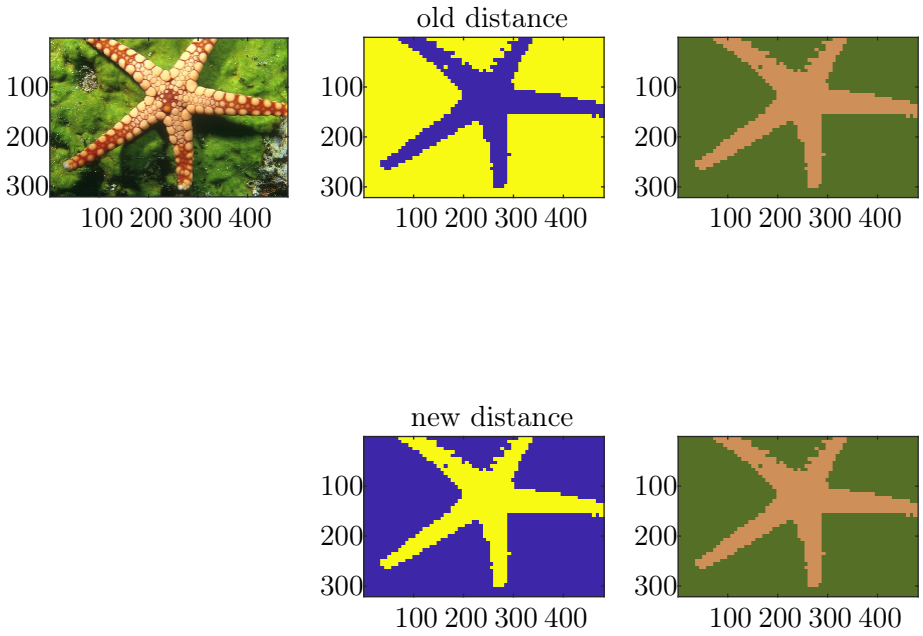


Figure 6: emgmm

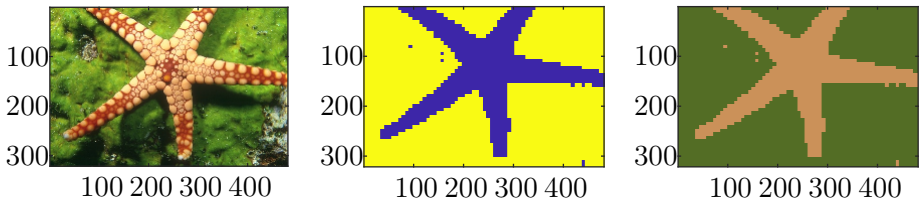


Figure 7: mean-shift

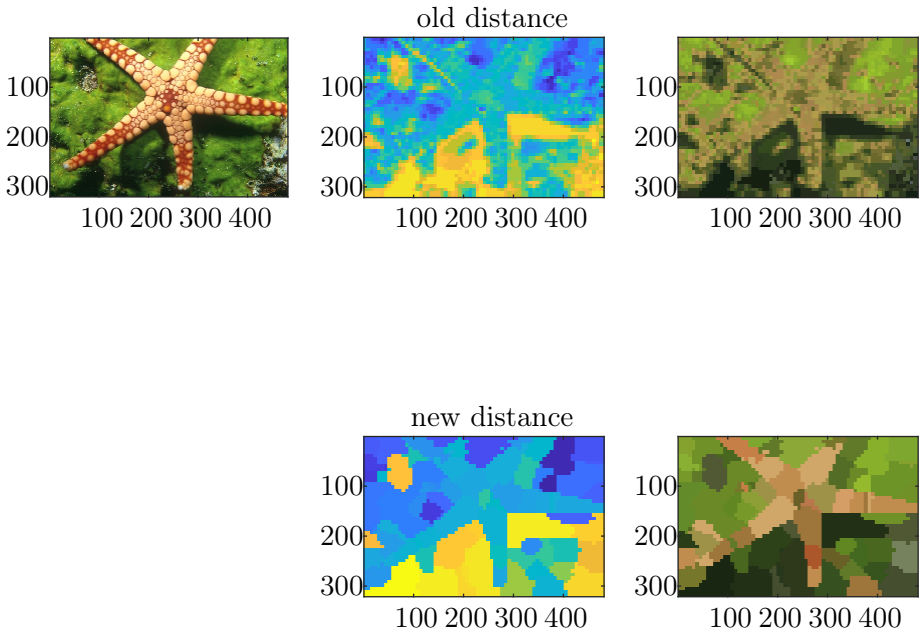


Figure 8: kmeans

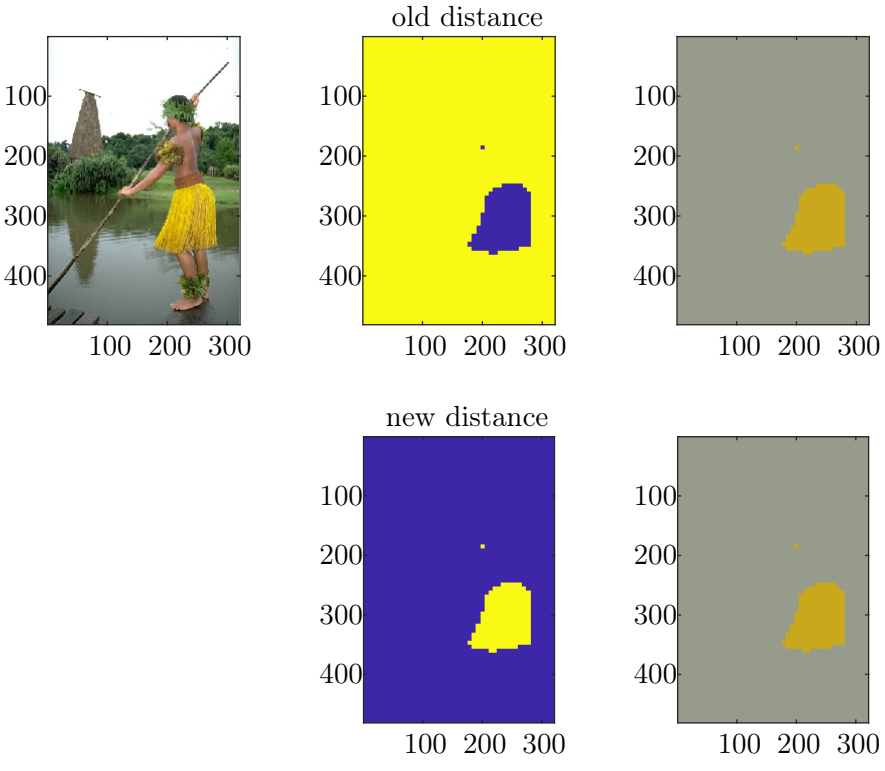


Figure 9: emgmm

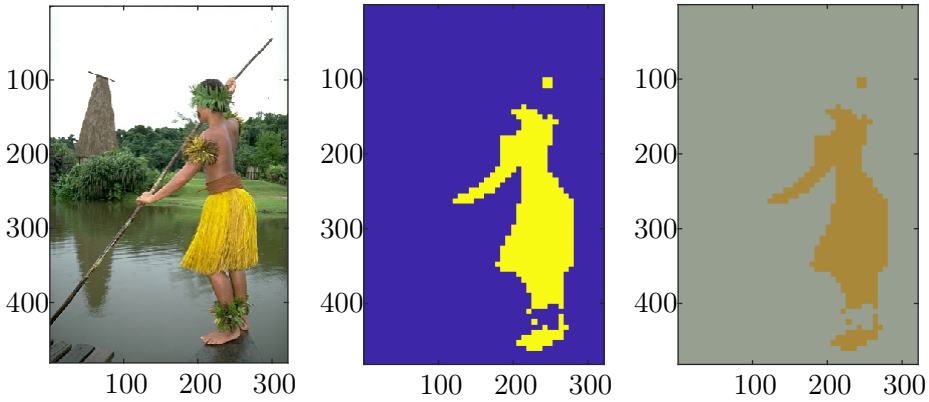


Figure 10: mean-shift

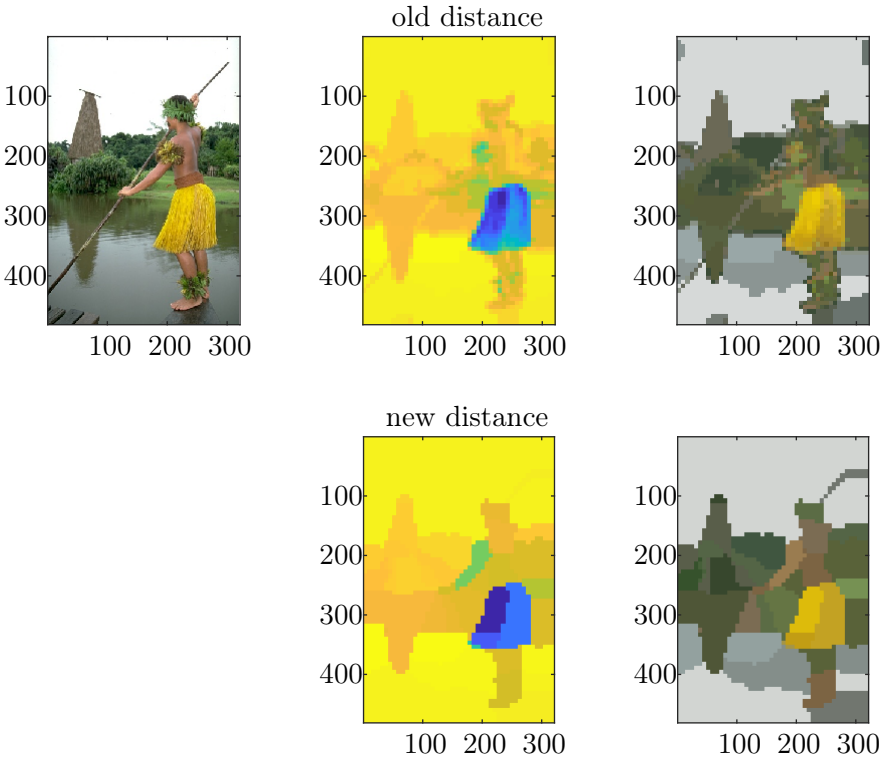


Figure 11: K-means with different k

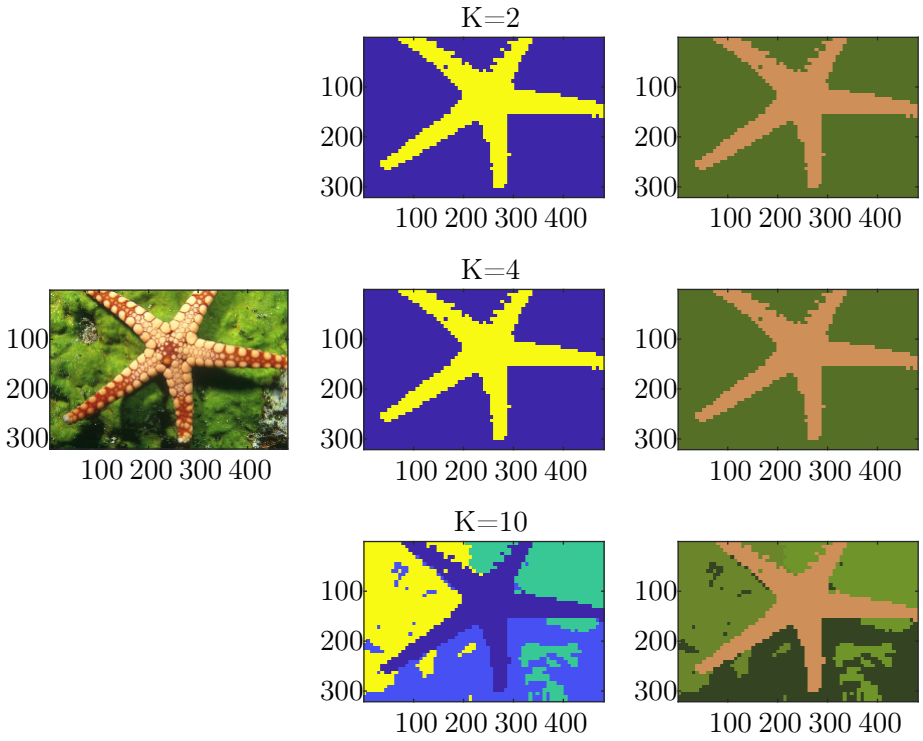


Figure 12: EM-GMM with different k

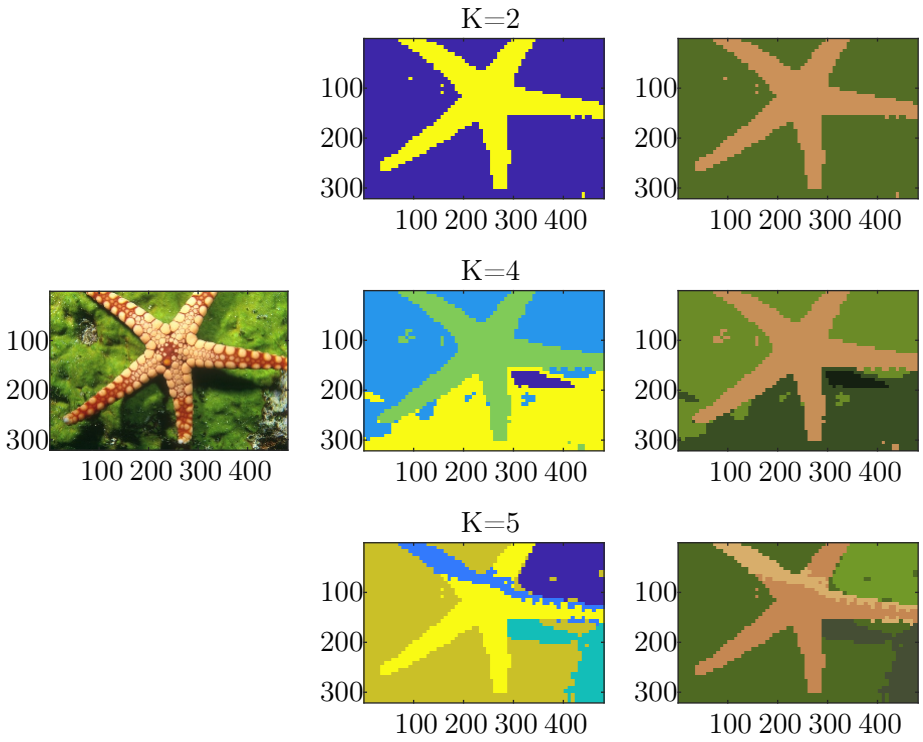


Figure 13: Mean-shift with different h