CS5487 Program Assignment

Clustering

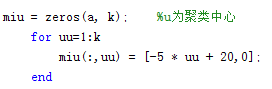
Ruyu Yan

55427256

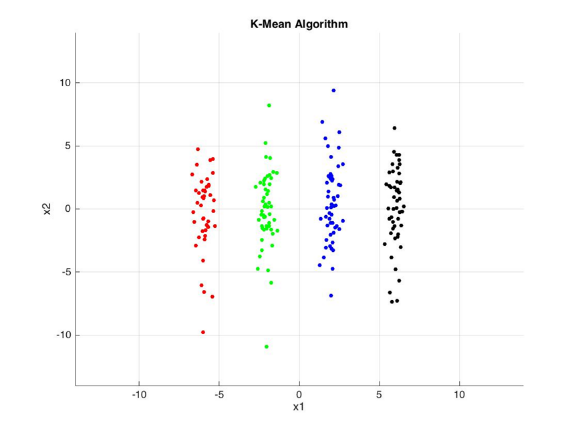
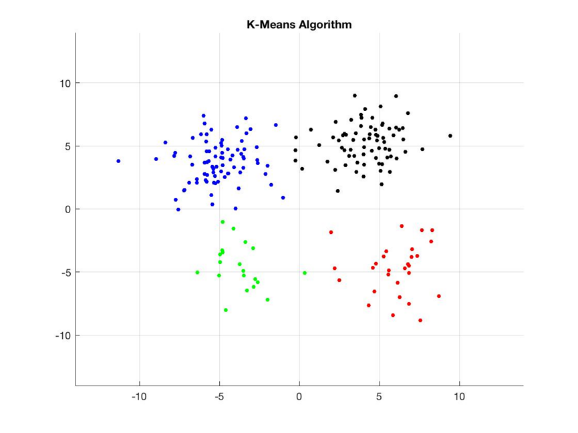
Part Ⅰ Clustering synthetic

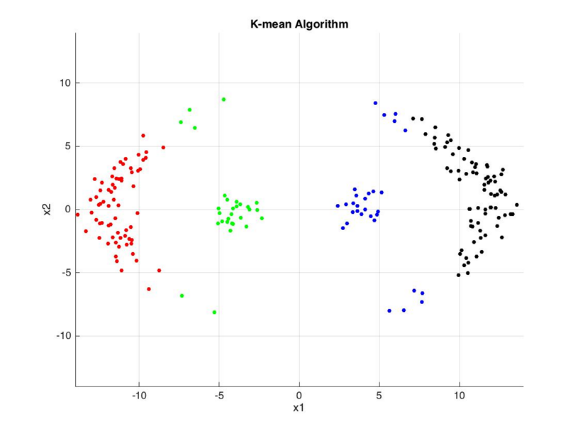
(a)&(b) Implement 3 clustering algorithms and run the algorithms on the three synthetic dataset.

1. K-means algorithm is written in the K-mean.m document which connects the iteration function K\_meanloop. In the initialization phase, according to the given cluster ,setting the uniformly distributed data points which far apart from each other.

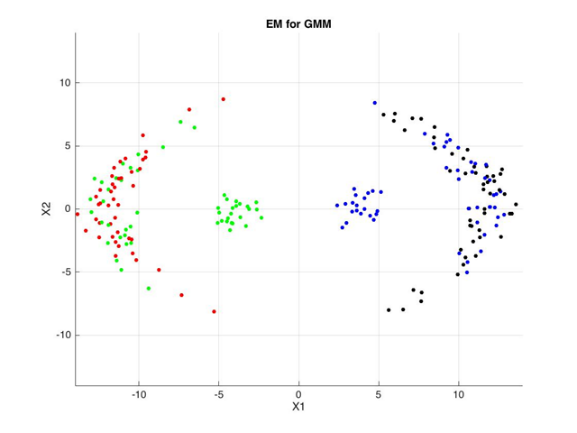
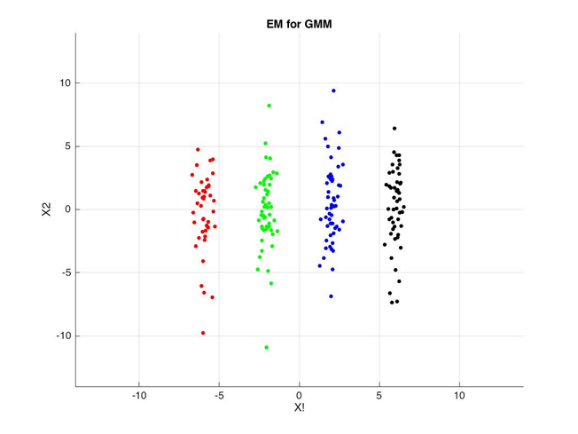
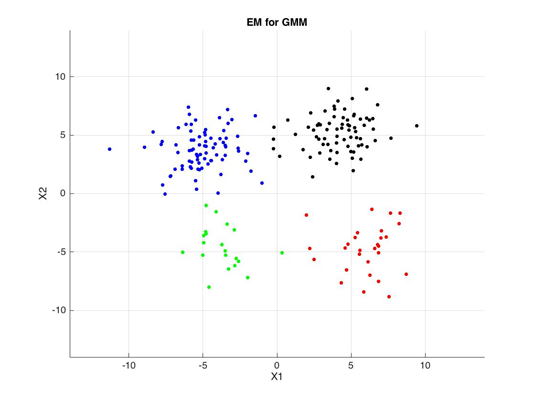


And the result of the clustering synthetic datasets is shown in following figure:

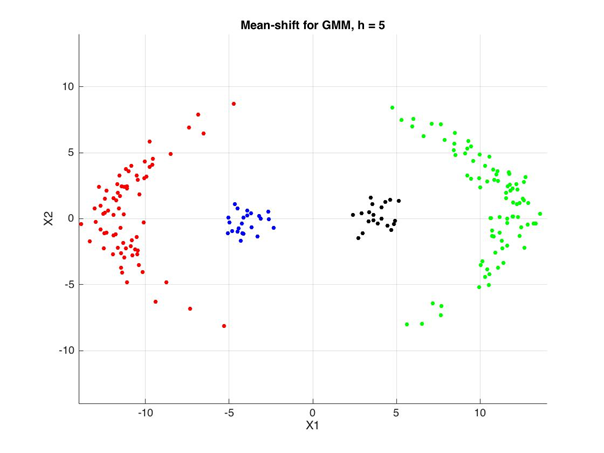
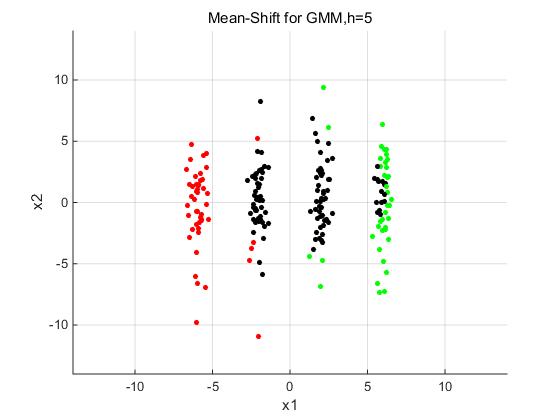
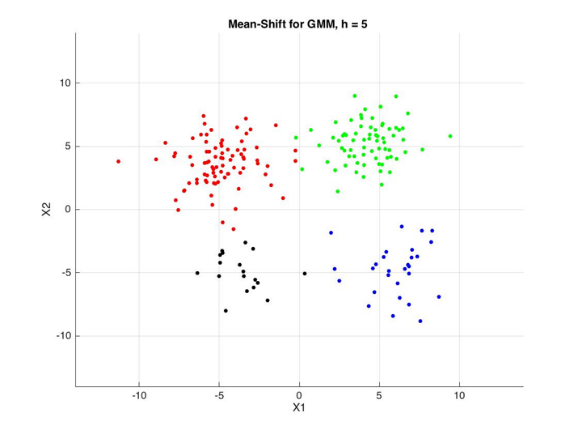




1. EM algorithm for GMM is written in the document EMmain.m, and I use the same initialization method as the K-mean, and the figure will shown as following:



1. Mean-shift use gradient ascent with an adaptive step-size to find local peaks in a kernel density estimate. There is no need to initial the cluster center, instead of setting bandwidth(free choose a number as 5),and the result can be shown as following:



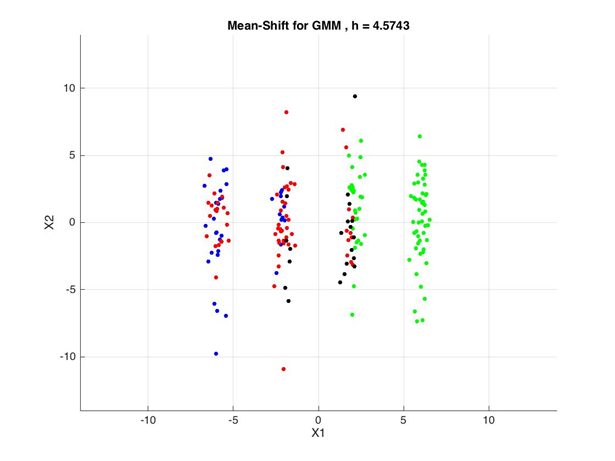
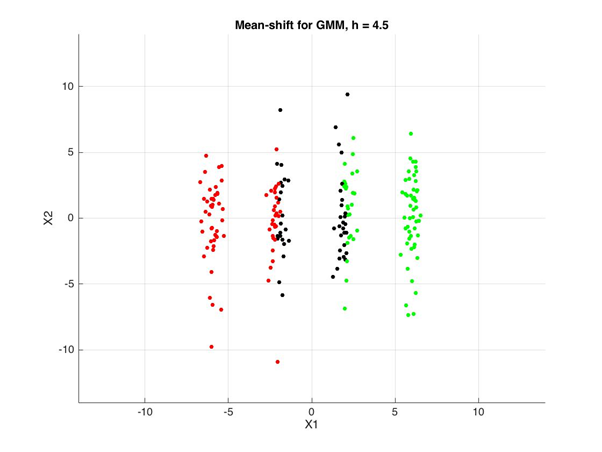
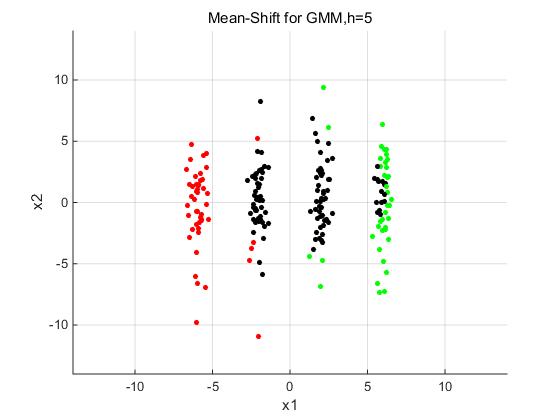
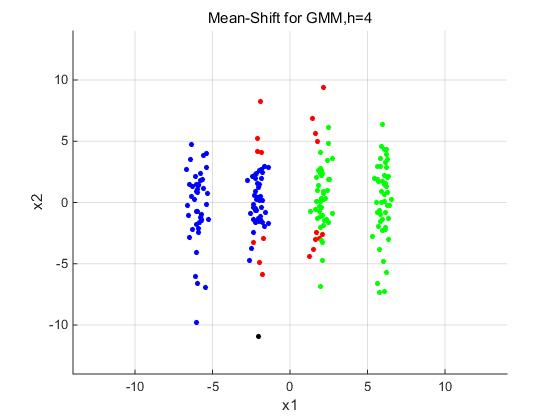
Summery of clustering algorithm:

Firstly for K-means algorithm, because we have to randomly select several cluster centers , the random data is effect the prediction very much so there are some error in the dataC clustering. As for the data set with distribution which is relative uniform and does not cross each other, the result will perform well.

Secondly ,the EM-GMM algorithm have the same time-consuming, they are both fast with operation is no more than about one minute. The difference comes to that the covariance is not fixed in EM-GMM，that means the data have a better model of prior probability and using covariance on calculating the posterior probability to cluster the data in proportion.

Finally we talk about the algorithm Mean-shift. From the mathematical formula, they make the every data points as one cluster center so that there are not cause error. Through the figure we get we can see, it solved the problem that the above two function have, which means that if there are some data not being together with a line distribution, the function also can take them in one cluster well. The disadvantage is that Mean-shift is too slow.

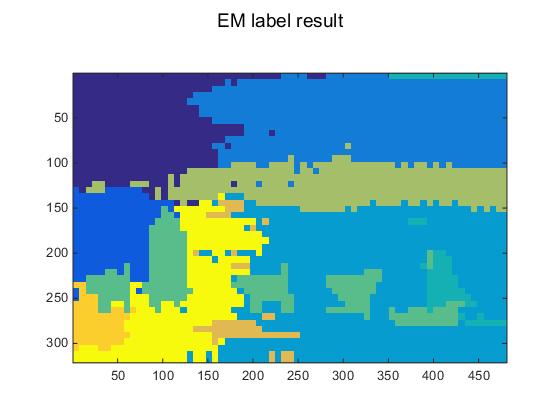
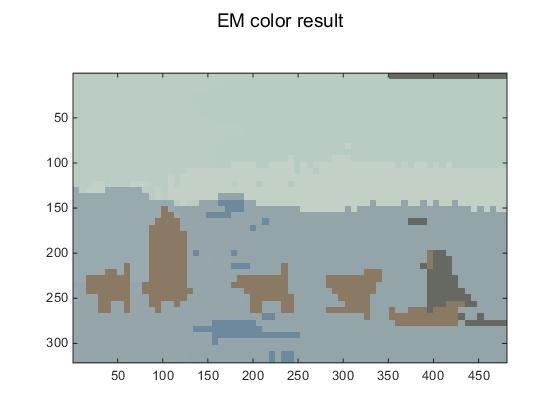
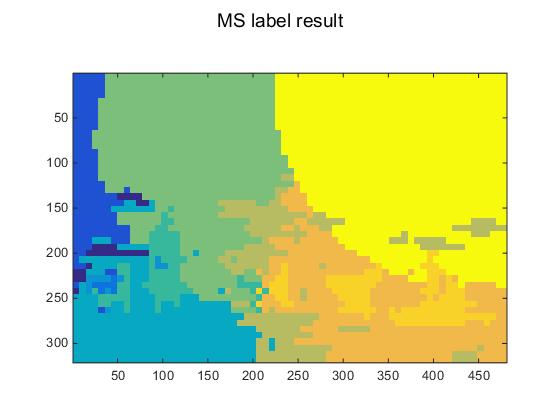
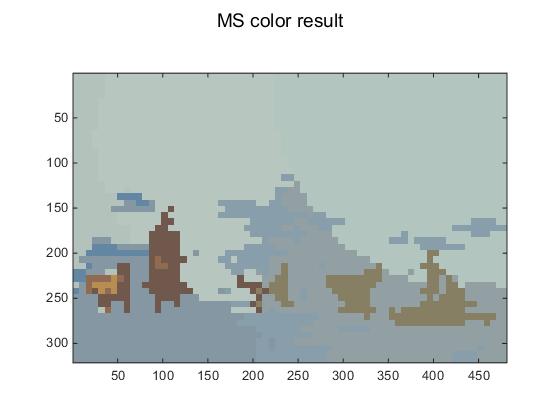
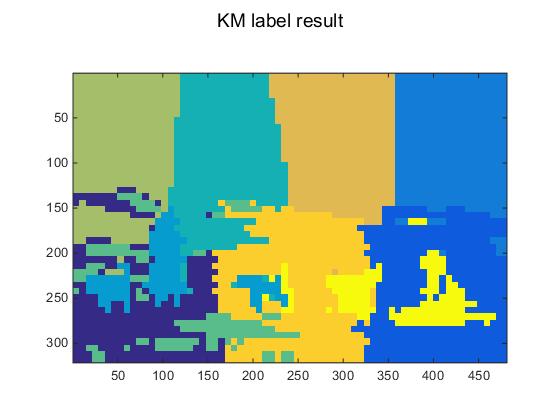
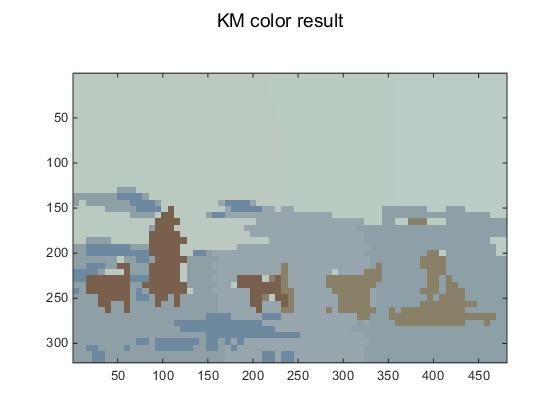
(c)The sensitive of h can be seen at the fact (we choose a dataset which has the most obvious effection), from which we can find the sensitive can at least accurate to three decimal places:



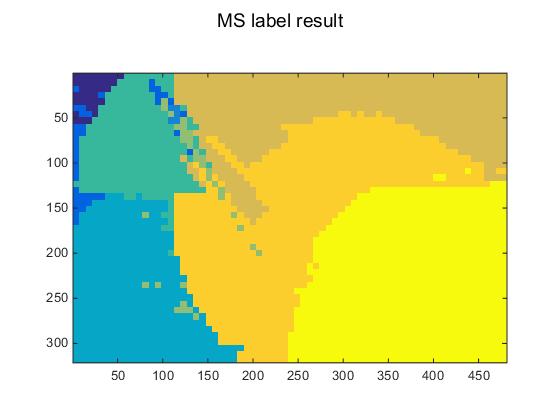
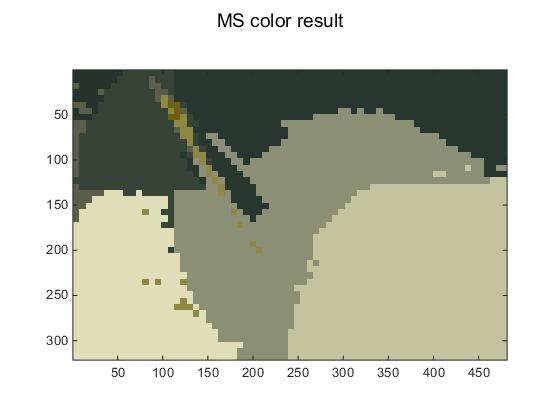
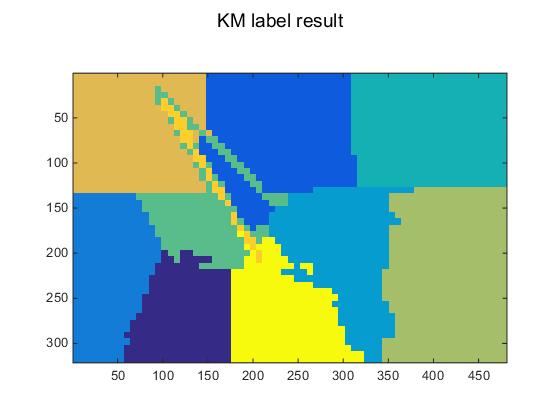
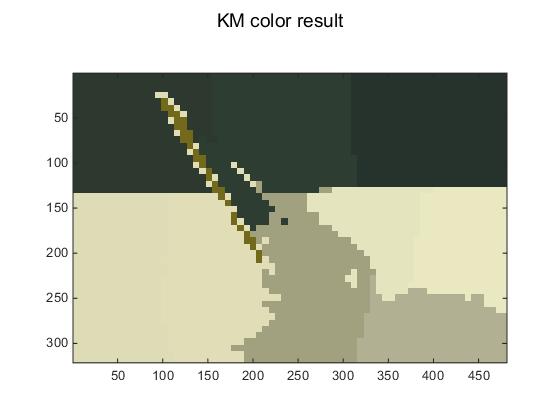
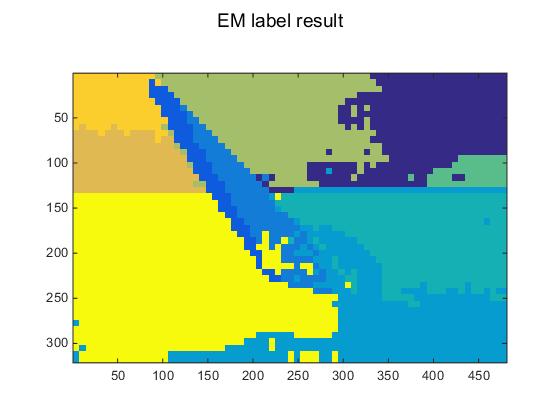
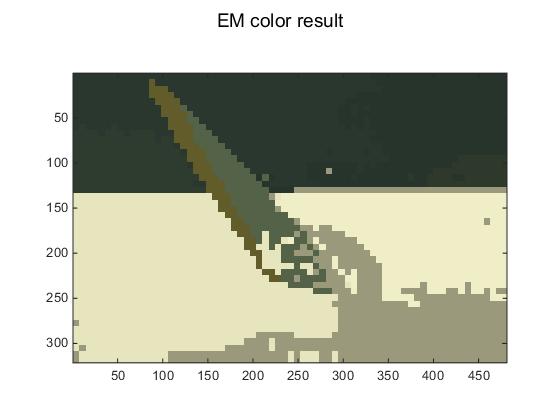
Part Ⅱ Image segmentation

1. Use the 3 clustering algorithm to segment a few of the provided image. I choose three image to do cluster, in which Mean-shift algorithm fixes the h at 7:

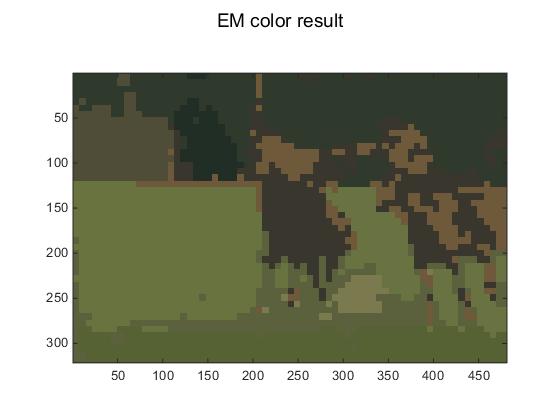
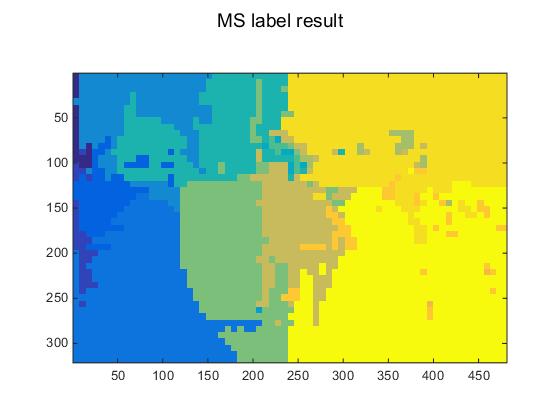
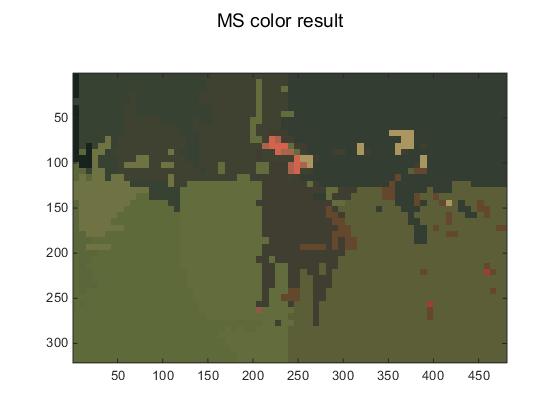
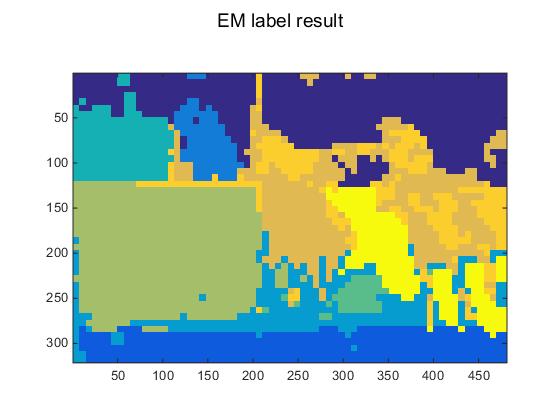
First picture: 

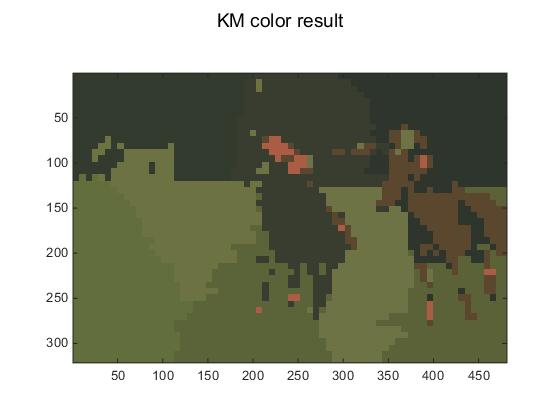
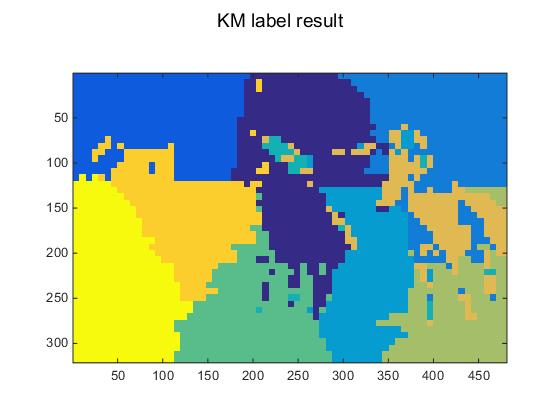


Second picture: 



Third picture:

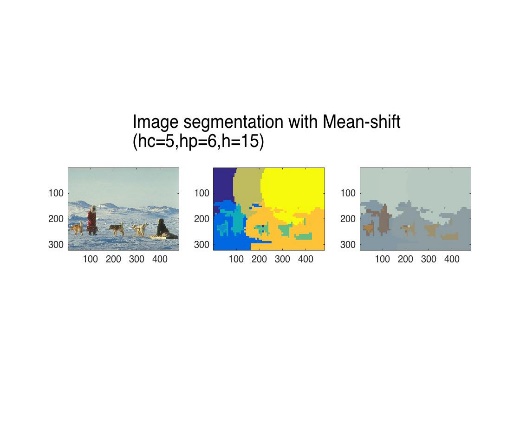
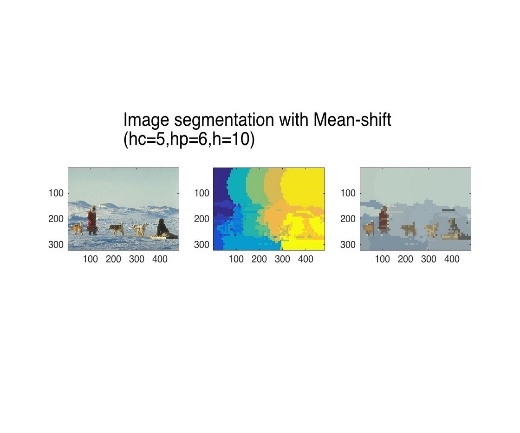




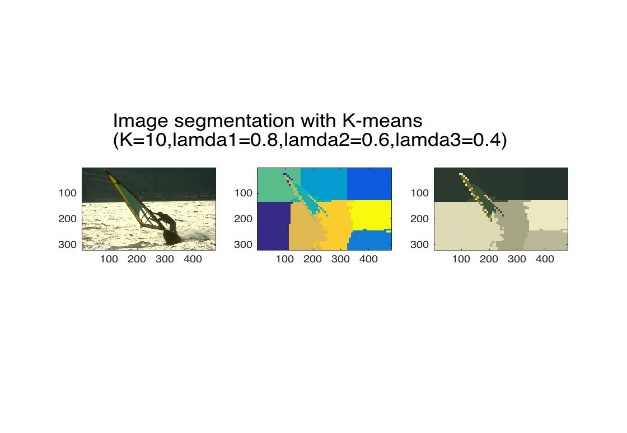
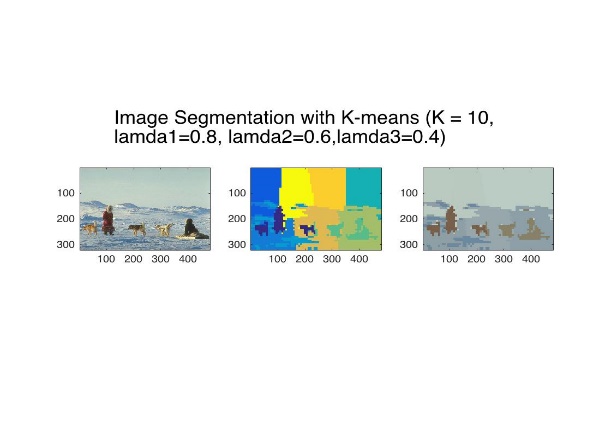
-- From the three sets of picture, we can conclude the EM-GMM is the best, and K-mean’s effect is worse than other .Mean-shift is unstable, which is bad for some sample like picture 2,and also can do well just like the last picture.

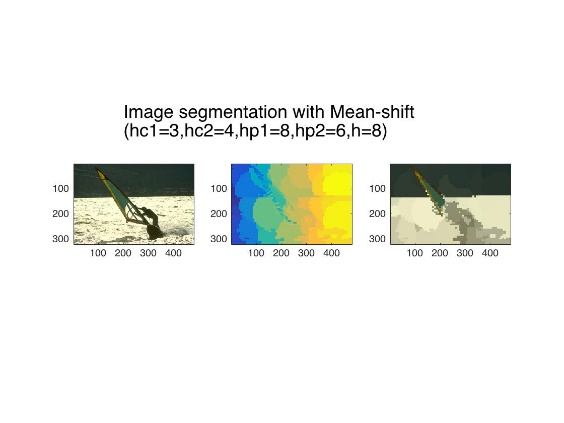
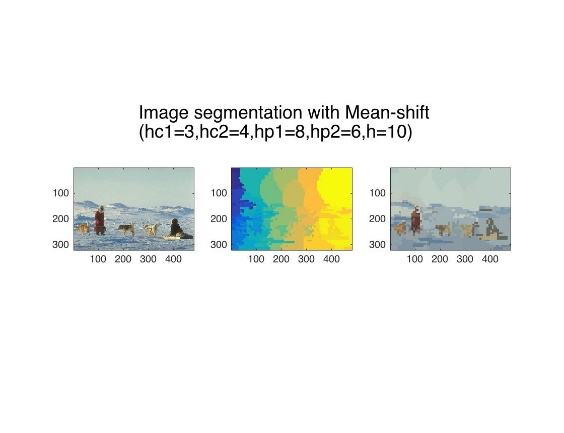
-- Change the parameters K there is not any obvious improvement in the picture cluster, but the time to run the algorithm become longer.

-- Change the parameters h to be bigger ,the predicted picture will be more blurry, just like:

1. Modify the K-means and mean-shift implementations to allow different feature scaling.



[Reference material] https://github.com/royukira/Cluster\_Segmenetation