Name: Hazem Taha Student ID: 400550393

Assignment 5 Report

I. K-Means Algorithm Implementation

I implemented the traditional unsupervised K-means algorithm with some minor changes to accommodate the problem at hand as well as the assignment requirements:

Centroid Initialization:

I used two methods as required in the assignment to do the initialization of the k centroids:

- Randomly pick the centers from the data points.
- Pick the centers such that they have a sufficiently large distance between them.

The way I did the second method (called it Max Distance initialization) is as follows:

1) Initialize the first centroid randomly.
2) Calculate the distance to the nearest centroid for each point.
3) Use the output to find the farthest point from its centroid.
4) Set this point as a new centroid and go back to step (2) until all k centroids are set.

Normalization/Standardization:

It's worth noting that I didn't normalize/standardize any of the images used. This is because the RGB channels (the features for the k-means model) already have the same scale and therefore there is no need to do normalization.

Centroid Update:

In some edge cases, when re-assigning points clusters, one cluster might not get any points and hence it can't update its mean and will stay like this forever. What I did in such a case is that I randomly re-assigned the cluster centroid again from within the data points.

II. Experimentation

1) Segmenting First Image:

Shown in figure 1 is the original image chosen for the first experiment. Figure 2 shows the effect of increasing the number if clusters (k) on the final reconstruction of the image.

Figure 1: Original version of first image used in experimentation.

Figure 2: Effect of varying the number of clusters K on the final reconstructed Image.

Name: Hazem Taha Student ID: 400550393

I tried K values from 2, 3, 10, 20 and 40. For each k, I did 3 different trainings (runs), 2 of which I applied random centroid initialization and the third I applied the max distance initialization. Every row in figure 2 shows the 3 runs corresponding to one of the k values. Every image shows the reconstruction as well the MSE from the original image and how many iterations it took for the model to fit.

Finally, table 1 shows a summary of the results acquired from this experiment. Figures 3,4 shows a visual plotting of the same data for easier interpretation.

Table 1: Summary of the results of experiment 1

Number of clusters (k)	Run	Initialization Method	MSE	Converge at Iteration
2	0	Random	96.3093	31
	1	Max Distance	96.3093	37
	2	Random	96.309	9
3	0	Random	76.9687	12
	1	Max Distance	76.9687	16
	2	Random	76.9687	16
10	0	Random	53.0788	58
	1	Max Distance	60.0773	80
	2	Random	55.2411	73
20	0	Random	42.327	102
	1	Max Distance	41.0376	163
	2	Random	40.1458	151
40	0	Random	28.4020	126
	1	Max Distance	29.9306	307
	2	Random	30.1603	147

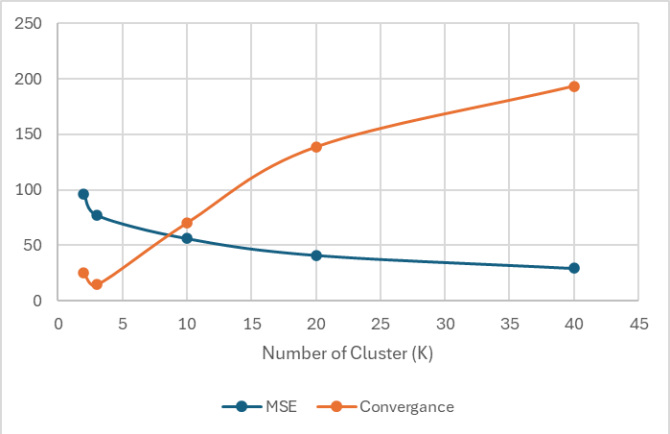

Figure 3: Summary of results (averaged over runs of each K)

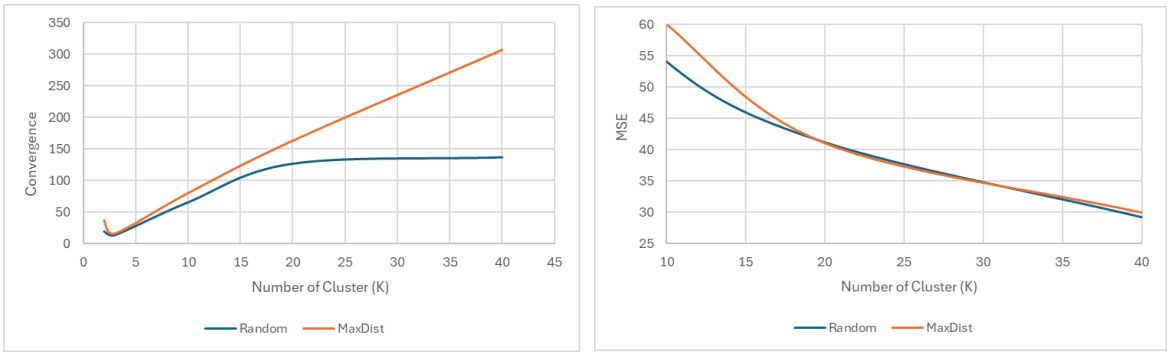

Figure 4: Comparison between different initialization for each K

The following can be noted from the results:

- Figure 3 shows that the MSE decreases nearly logarithmically with increasing the number of clusters, while convergence time increases nearly linearly. However, there is a small drop in convergence time happening at k = 3 that caught my attention and needs extra investigation.
- The two initialization methods did not show notable difference at small values of k. However, as K increases, it can be seen from figure 4 that the random initialization achieved overall better MSE and convergence time as the convergence plateaued at K = 20 while it kept increasing using Max Distance.

1) Segmenting Second Image:

Shown in figure 5 is the original image chosen for the second experiment. Figure 6 shows the effect of increasing the number if clusters (k) on the final reconstruction of the image.

Figure 5: Original version of second image used in experimentation.

Figure 6: Effect of varying the number of clusters K on the final reconstructed Image.

Table 2 shows a summary of the results acquired from this experiment. Figures 7,8 shows a visual plotting of the same data for easier interpretation.

Table 2: Summary of the results of experiment 2

Number of clusters (k)	Run	Initialization Method	MSE	Converge at Iteration
2	0	Random	97.6965	29
	1	Max Distance	97.6965	29
	2	Random	97.6965	29
3	0	Random	87.6293	18
	1	Max Distance	87.6293	18
	2	Random	87.6293	18
10	0	Random	40.6338	67
	1	Max Distance	41.1285	142
	2	Random	40.8886	60
20	0	Random	24.5843	99
	1	Max Distance	23.9973	113
	2	Random	24.6281	177
40	0	Random	14.4366	143
	1	Max Distance	14.5988	154
	2	Random	14.4243	143

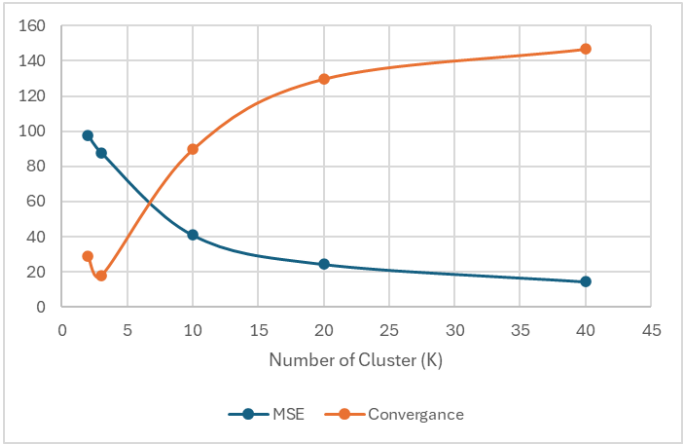

Figure 7: Summary of results (averaged over runs of each K)

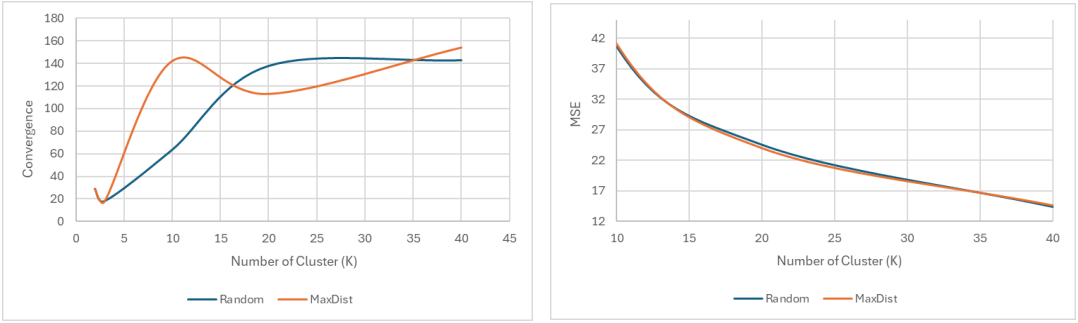

Figure 8: Comparison between different initialization for each K

Notes on the second experiment:

- MSE decreases nearly logarithmically with increasing the number of clusters while the convergence time increases.
- The two initialization methods don't show a notable difference at small values of K. However, as K increases, the random initialization achieved better convergence time (except around K=20) and it plateaued while the random initialization kept increasing. The MSE is nearly the same for both methods at large values of K.
- There is also a small drop in convergence time happening at K = 3 just like in experiment 1.

III. Conclusion

The following can be concluded:

- As K increases, MSE decreases nearly logarithmically, and convergence time increases nearly linearly. This shows the trade-off between the granularity of the segmentation and the computational cost.
- Different initialization methods do not show notable effect on both the MSE and convergence time at low number of K, but start to show effect with increasing K.
- The results show that using different initialization methods has an unidentical effect on the MSE and convergence time across different images. This indicates that there is no best initialization method in general, but it's totally dependent on the image itself and its distributions of colors.
- However, unlike the Max Distance initialization, there is a consistency with the convergence time results using the random initialization method as the curves follow nearly the same shape.
- In general, it can be said that Random initialization of clusters archives somewhat more promising results in both MSE and convergence time, yet it is still dependent on the image.
- For both images, there is a noticeable drop in convergence iterations at K=3, indicating that this might be a natural clustering level for images or that three clusters better match the intrinsic grouping of pixels within an image than two.
- A smaller MSE results in an image that is closer to the original image. However, I cannot say whether a smaller MSE corresponds to a more pleasing visual reconstruction because it's subjective. One may find a simple abstract image with few colors more artistic than a reconstruction that is closer to the original image yet still shows some artifacts.