James Hahn
CS1675
Homework #2

Results

After five simulations with differed configurations, below are the visual results, in addition to the settings in that simulation:

K = 5 iterations = 10 restarts = 5	
K = 5 iterations = 100 restarts = 5	
K = 2 iterations = 10 restarts = 5	
K = 2 iterations = 10 restarts = 15	
K = 4 iterations = 30 restarts = 15	

Findings

In terms of findings, there are two main categories to evaluate: running time, image segmentation quality. These can all be observed through the tuning of all three parameters: K (number of clusters), iterations (iterations for k-means), and restarts (number of times k-means was restarted).

Running time

The only parameter determined to have a significant impact on time was the *iterations* parameter. *Restarts* had an impact as well, but simply changing iterations from 10 to 100, for example in the first simulation, the running time increased from 10 minutes to 1 hour 30 minutes. As such, we want to keep the number of iterations to a minimum while still providing the k-means algorithm enough time for the cluster centers to converge. As shown in the first two examples of the **results** section, there is not a significant change in quality between 10 and 100 iterations. As such, 10 iterations was chosen as a more optimal configuration.

Image segmentation quality

Segmentation quality is an important factor to analyze in any segmentation algorithm. In the findings, k = 2 was found to produce a binarized image, while k = 4 and k = 5 produced more precise representations of the actual underlying features (i.e. edges) of the original image. There is not a significant dip in quality between k = 4 and k = 5, but k = 4 has a slightly smaller running time due to requiring a fewer number of calculations for cluster centers. As such, k = 5 has more precise segmentation quality, but k = 4 is more optimal due to the tradeoffs between running time while having strikingly similar segmentation quality.

Impact of *restarts*

The *restarts* parameter was found to have little to no effect on the segmentation quality of k-means, while only taking longer to run the simulation. This is possibly due to the chosen values to alternate between (5 and 15). The globally optimal sum of squared distances error for the k-means simulations could have easily been within the first five restarts of the simulation, so the quality only differed slightly. The *restarts* parameter might be useful while comparing one or two restarts to one hundred starts (i.e. there is a significant gap in restarts), but is useless otherwise. An impact also might not have been noticed due to the low resolution of the image (100x100), so low-level features would not have mattered and only high-level features are detected by the algorithm.

Conclusions

In terms of parameters, *K* and *iterations* largely affect k-means the most. *K* affects the segmentation quality significantly. It can make the difference between foreground/background segmentation and building-by-building segmentation. *Iterations* affects the running time significantly, but did not impact the segmentation quality a ton. Part of this may be explained by the low resolution of the images. Finally, *restarts* did not show much of an impact at all in the simulations, other than an increased running time. As such, a configuration of K = 4, iterations = 30, and restarts = 15 produced optimal results in terms of both segmentation quality (being granular but not too granular), while not taking a significant amount of time to run.