# Information Visualization Final Project: One-hundred plant species leaves data set Data Set

## Data description

One-hundred plant species leaves data set Data Set is from the UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set ).

This dataset consists of one-hundred varieties of leaves. For each variety, 16 examples of leaves were collected(Mallah, Cope and Orwell, 2013). The leaves were placed on a white background, and processed via semiautomated method, then transformed the color image into binary gray image.

Three different features were then created: The shape feature was created by using Centroid Contour Distance Curve (CCDC), and then scaled down to a 64 by 1 vector; The texture feature was created by Gabor Co-Occurrences (Cope, Remagnino, arman & Wilkin, 2010). A 20-feature vector generated by Gabor Co-Occurrences was then quantized to one of 64 predetermined vectors. A histogram is then built from the number of the 1024 vectors assigned to each of the 64 quantization vectors, giving a 64-feature vector for describing each leaf. (Mallah, Cope and Orwell, 2013); The margin feature used the method described by Beghin, Cope, Remagnino & Barman, 2010. The results of this method were then used to produce a 64-feature vector describing the margin, using the same quantization and histogram method as for the textures.

Three features are stored separately in three files, each file has 64 columns and 1600 rows (total 100 species, 16 rows for each species). The leaves images are also provided in the data folder(total 1600 images).

List of variables: The data was created by the algorithm mentioned above, features are all 1 by 64 feature vectors.

## Visualizations

Figure 1 is the leaves images form the dataset, 49 species shown in the figure were selected randomly. Original images are in different size; each image plotted above was transformed to a size of 170\*250 and then align the center of leaves to the center of the image.

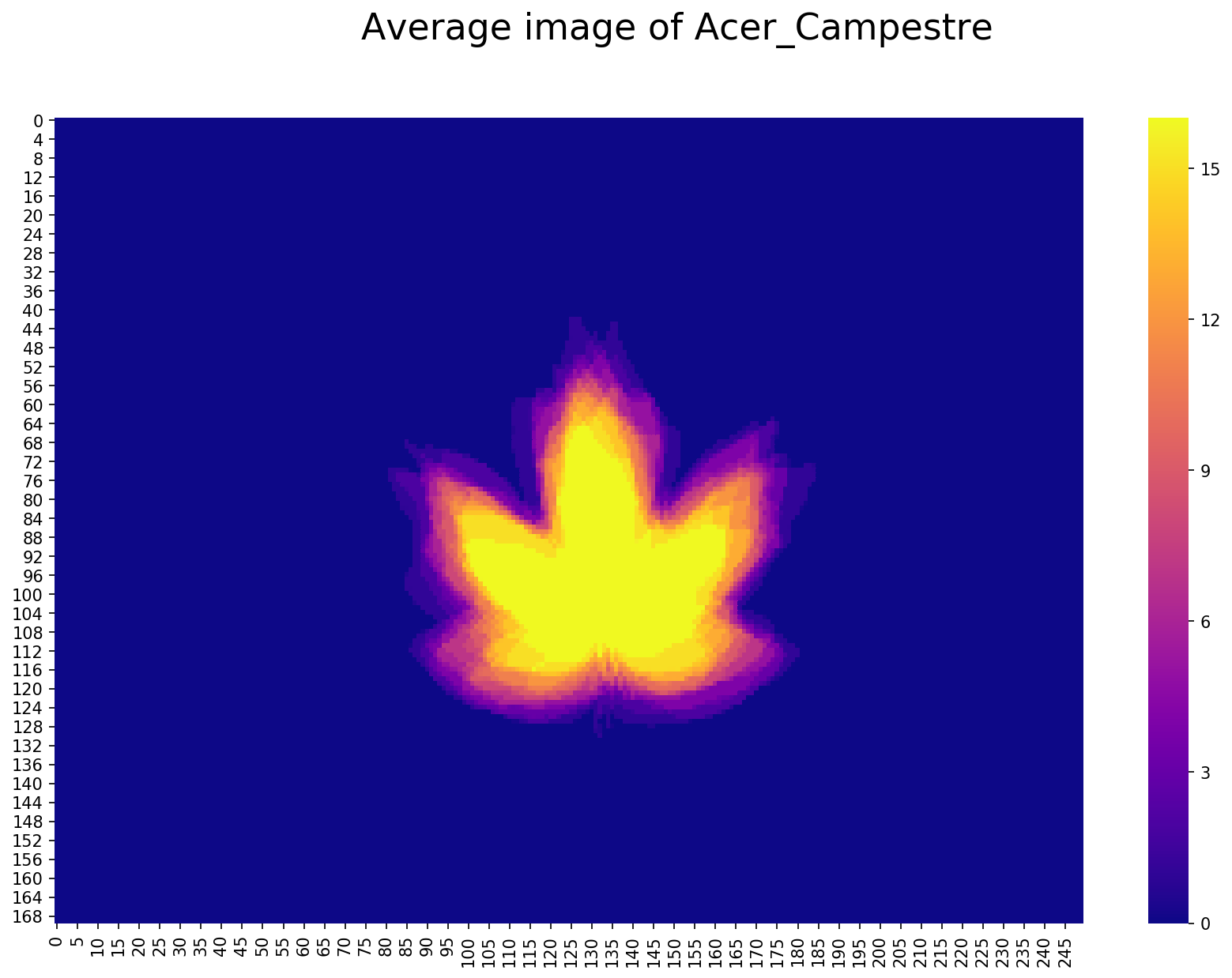
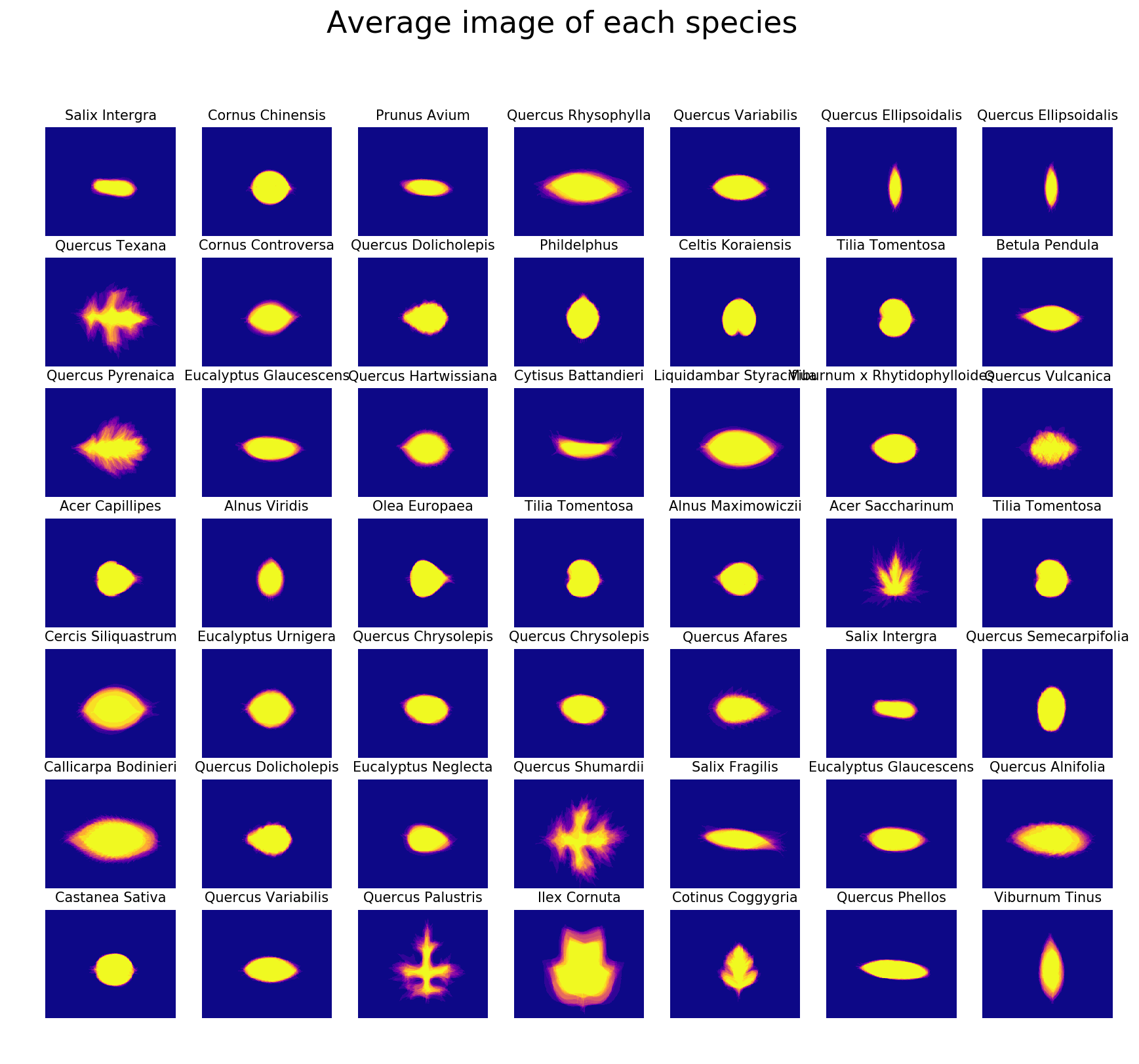


Figure 2 is the average image of 16 acer campestre leaves. The center of the leaves shows in yellows, which means almost all 16 leaves overlap in this area; the edge shows in pink, which means the shape at the edge of the leaves varies more.

Same as Figure2, Figure3 are average images of 49 randomly selected species.

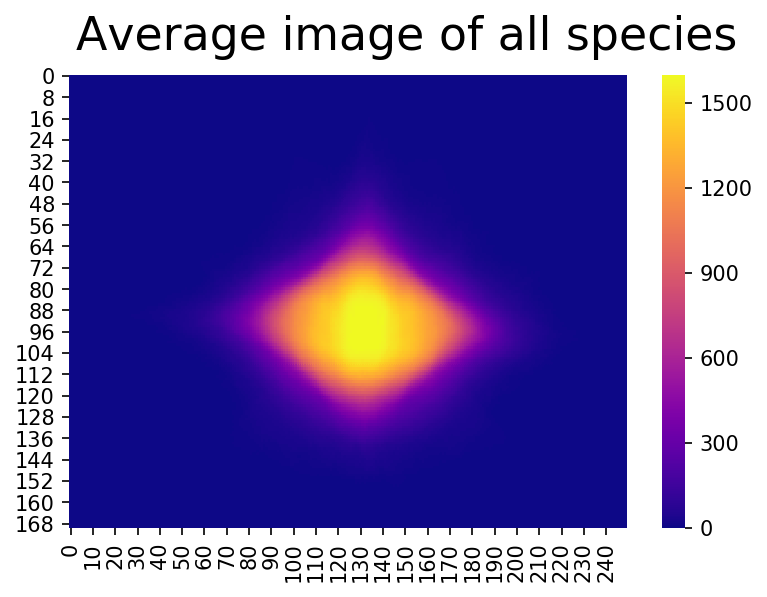


Figure4 is the average image of all 1600 images.

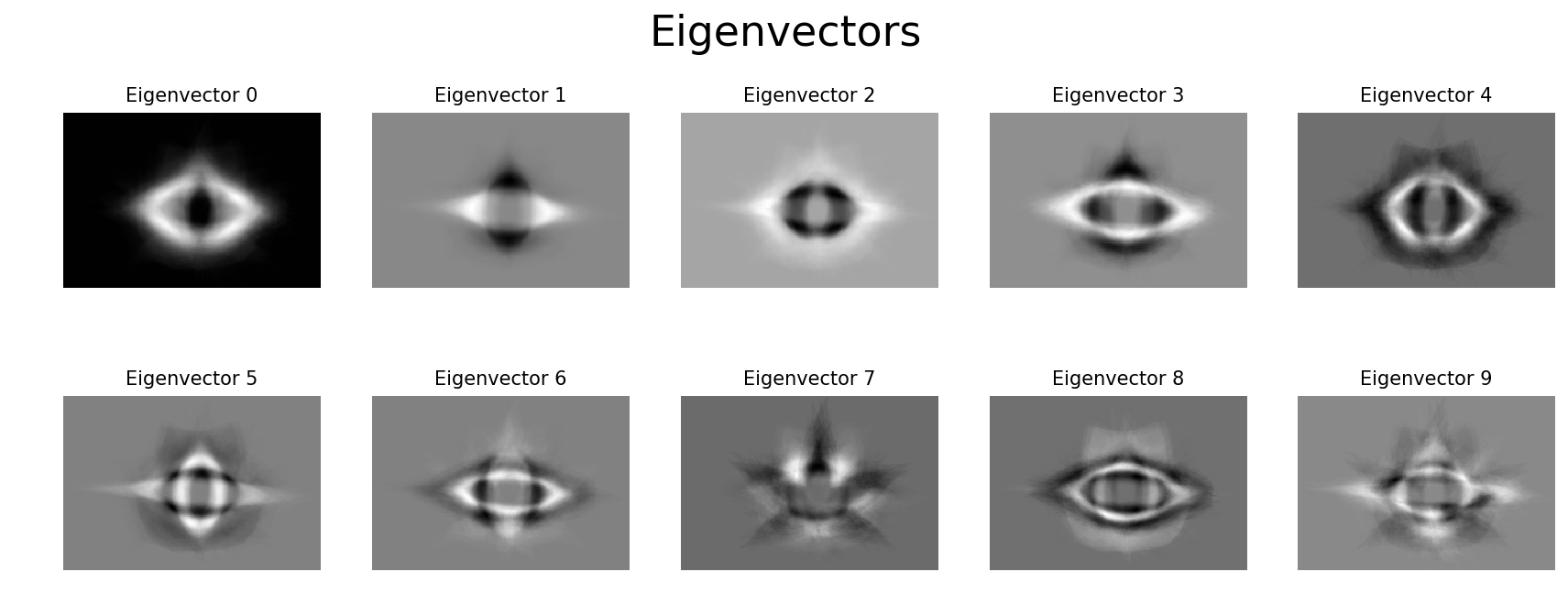


Figure 5 shows the 10 eigenvectors calculated by using PCA.

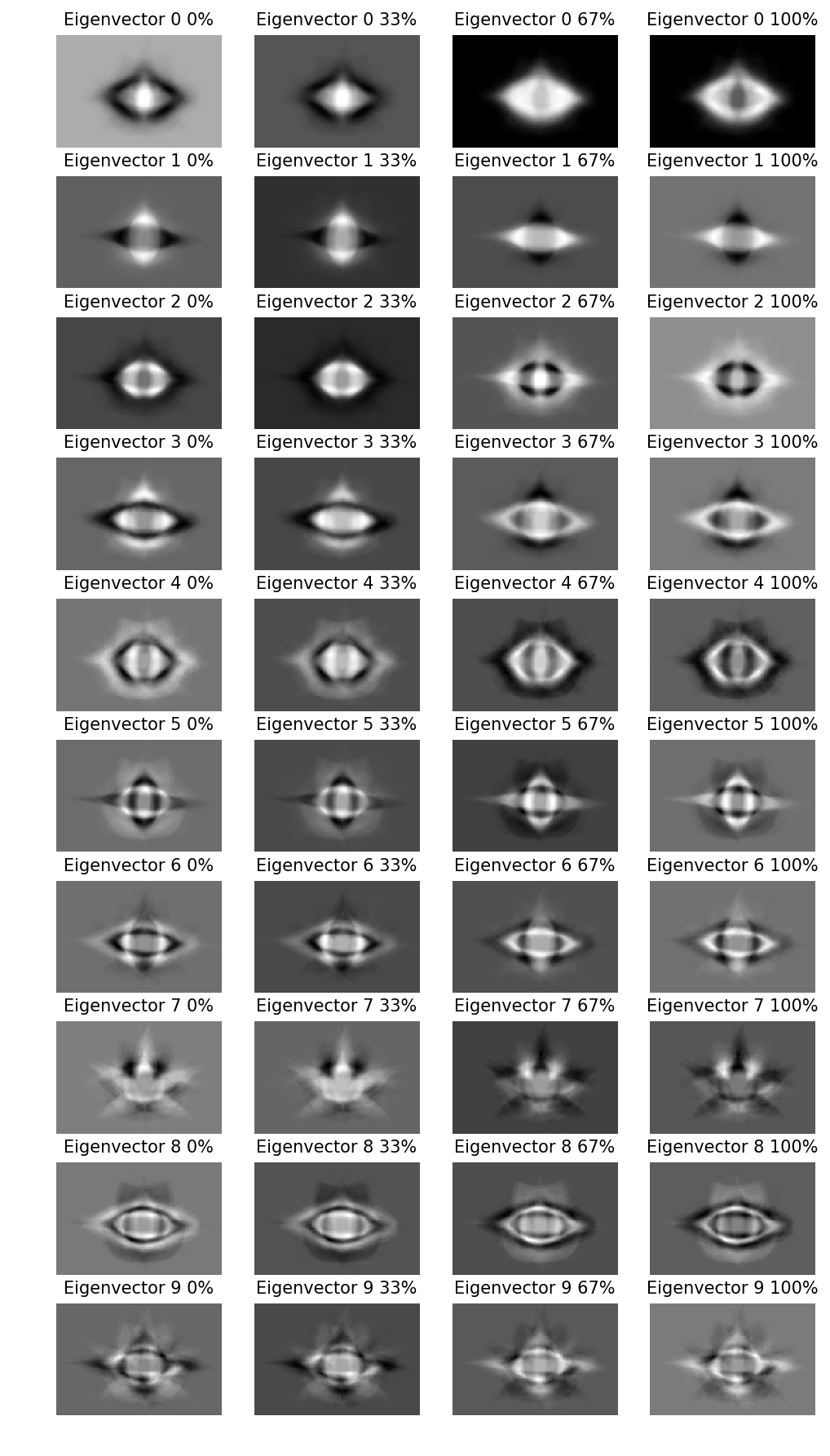


Figure 6 shows the leaves reconstructed by different principal component. There are 10 rows in the figure, each row represents one of 10 principal components of all images. Four images in each row shows the image reconstructed by adjusting the weight of this principal components. For example, the first row is the first principal component: PC0, which captures the direction with maximum variation. The first image in the first row is reconstructed by add 0% of PC0 to mean image, the second to forth images are reconstructed by adding 33%, 67% and 100% of PC0 to the mean image.

By comparing the difference between each image in the same row, we can see what feature a principal component capture. The background changes a lot in the four images in the first row, which means the first principle component captures the background of the image; Similarly, the second principle component captures whether an image is vertical or horizontal; the third principle component captures shape of the edge of a leaf. Etc.

## Observations

## Appendix

## Reference

1. Mallah, C., Cope, J., & Orwell, J. (2013). Plant leaf classification using probabilistic integration of shape, texture and margin features. *Signal Processing, Pattern Recognition and Applications*, *5*(1).