Parkinson Disease Dynamic/Spiral Drawings

Overview

Context

More than 10 million people worldwide are living with Parkinson's disease. Improving machine learning model which identifies Parkinson's disease will lead to helping patients with early dialogs and reduction of treatment cost. Implemented an unsupervised classification on the dynamic spiral drawings of the Parkinson's test dataset, and classify normal people from Parkinson's disease ones.

My Approach

Initially, I have thought of implementing Convolutional Neural Networks for image pattern matching. However, since the data available is relatively small and that neural networks need large volumes of the data to be trained on, I implemented distance based clustering techniques to classify the given drawings. Converted images to their pixel data and implemented t-SNE dimensionality reduction technique to bring down to $[25\ X\ 2]$ dimensions. This helped me visualize the image data in 2 dimensions and better understand the data spread.

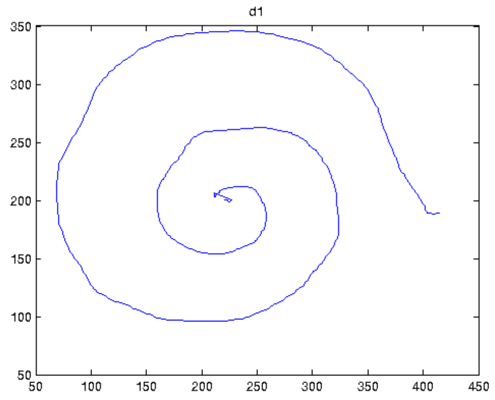
Scaled the data and explored multiple clustering techniques (Kmeans, Hierarchical, DBSCAN and Spectral) and tuned the models for better classification. Tested including the Static Spiral drawings and provided my observations.

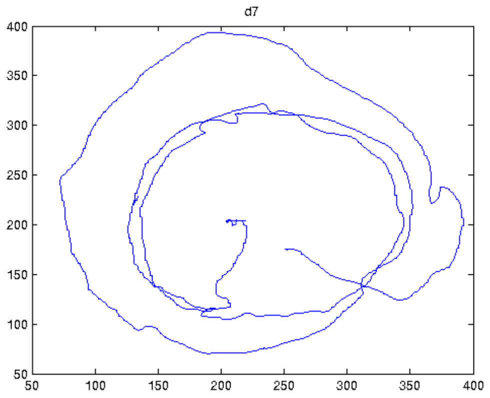
Steps Involved

1. Loaded images and converted to data frame
2. Data Exploration
3. Feature Generation
4. Dimensionality Reduction using t-SNE technique
5. Clustered on Dynamic Spiral Drawings
6. Evaluated including Static Spiral Drawings

Data Exploration and Feature Generation on Dynamic Spiral Drawings

Although all the images are of same size $561\ X\ 420$, observed that few drawings have different scales as shown below. This inconsistency could have significant effect on the cluster formation, as most of the clustering techniques are distance based.





To avoid this inconsistency, I have masked all the pixels to NULL values excluding blue pixels. This way obtained data frame with values (1) for blue pixels and rest as 0in each image.

Dimensionality Reduction

The final data frame obtained is of size $25\ X\ 235620$, with each row representing an image and the columns representing pixels within an image. Implemented t-SNE dimensionality reduction technique and obtained data frame of dimensions $25\ X\ 2$. My reasoning's on why dimensionality reduction are given below;

Why Dimensionality Reduction?

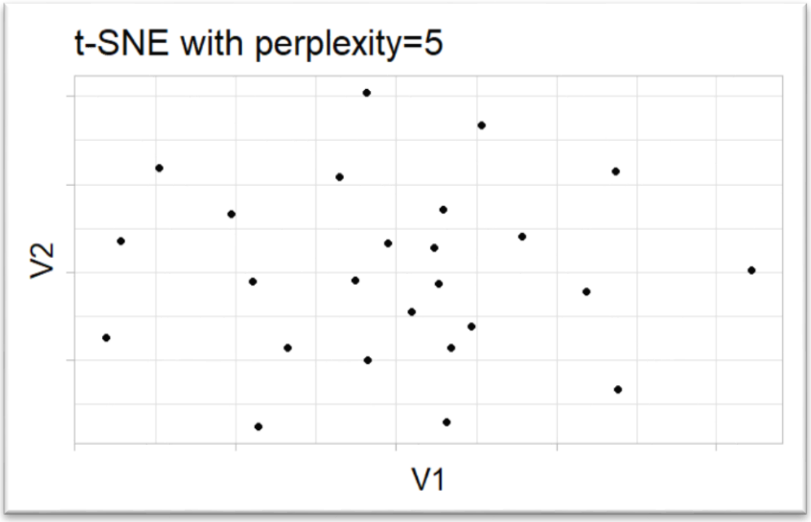
1. The image data we are dealing here is quite disperse with values only for blue pixels in an image of size $561\ X\ 420.$ i.e. variance is spread among large number of features.
2. Dimensionality reduction enables us to summarize and visualize the data in fewer dimensions, with minimal loss of information.

Why t-SNE over PCA?

1. PCA captures the variance in the fewer data points however, it will lose the data's structural information. In our case, where we are dealing with spiral structure of an image I believe it is more important to retain the structural information along with the variance.
2. To achieve this, t-SNE is implemented which preserves the local structure in the data by capturing in fewer dimensions.

t-Distributed Stochastic Neighbor Embedding and its hyper parameters

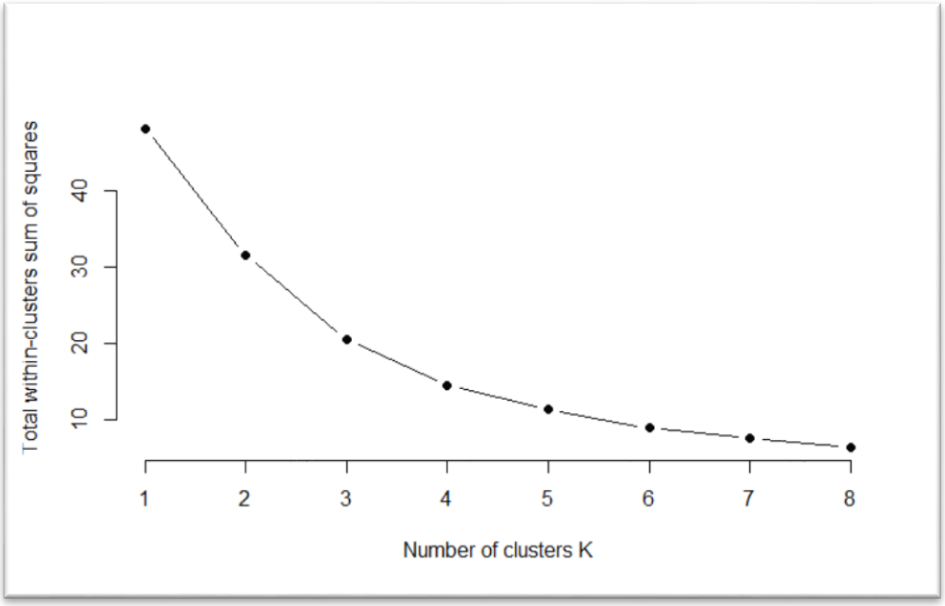
1. Dimensions – 2. For the ease of visualizing the spread in the data and further viewing clusters.
2. Perplexity – In general, it is viewed as a knob that sets the number of effective nearest neighbors and that it should be much smaller than the number of data points. I have experimented with perplexity values 2 till 8 and obtained best results at 5.



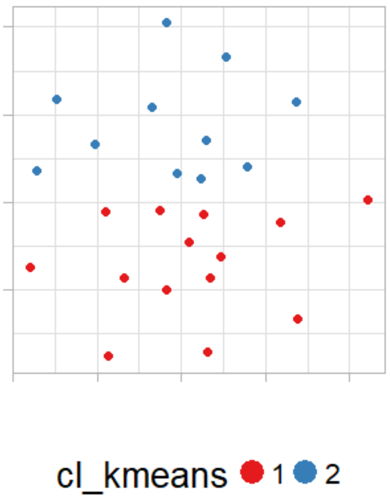
Unsupervised Clustering

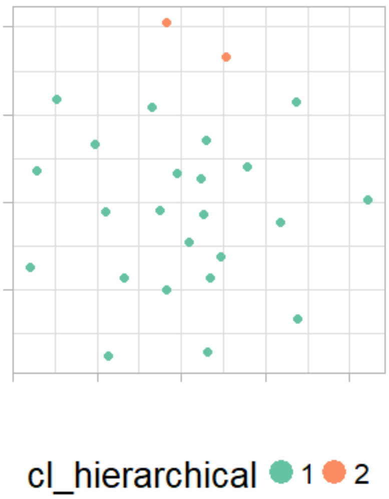
Elbow Method

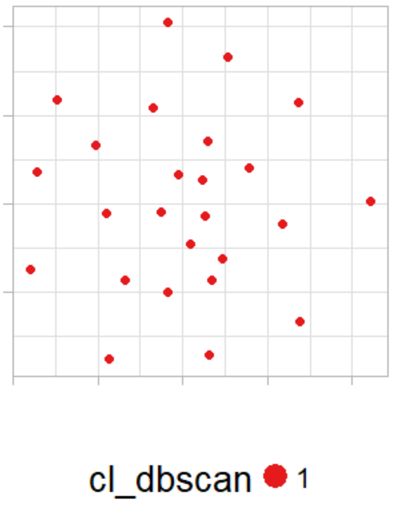
Initially implemented elbow method to identify the ideal number of clusters to be 4, as shown below, based on the data. However, it made more sense to cluster into just 2 groups as healthy people and Parkinson's disease ones.

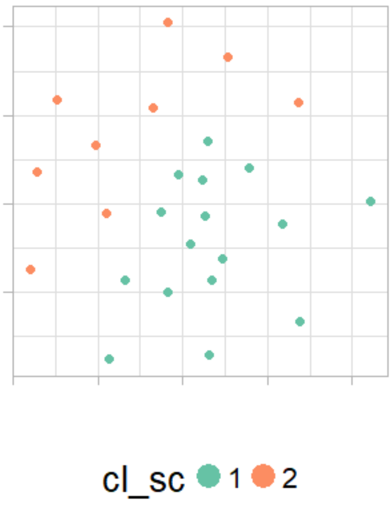


Implemented KMeans, Hierarchical, DBSACN and Spectral clustering techniques with number of clusters as 2. Shown below are the cluster plots;



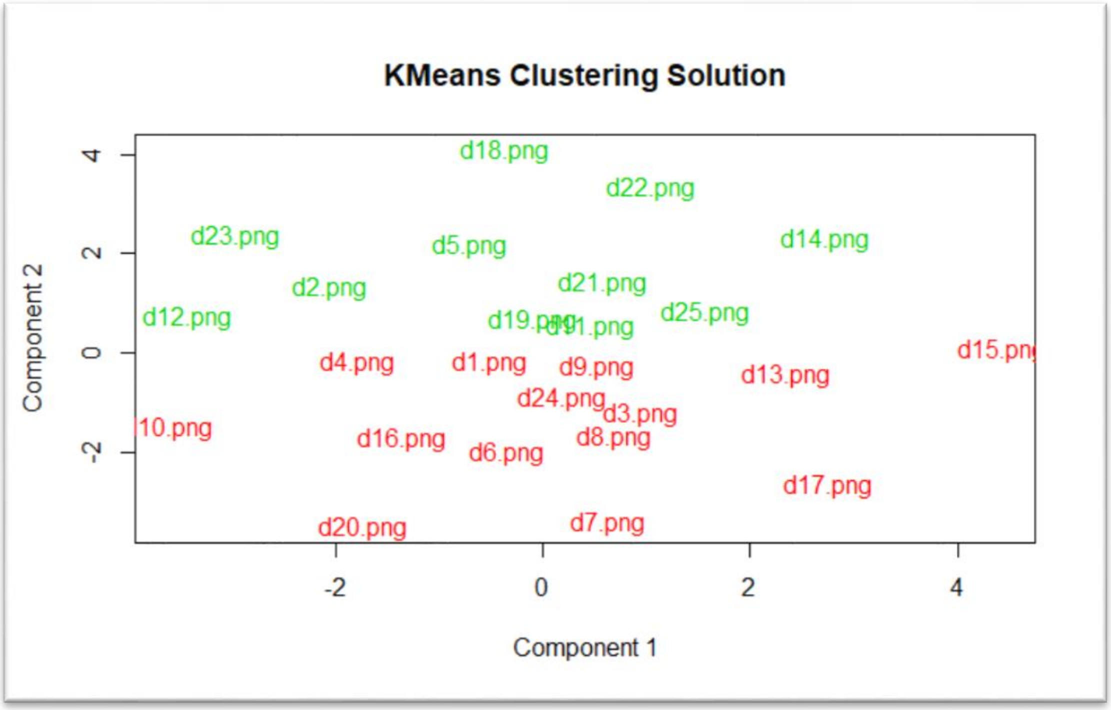




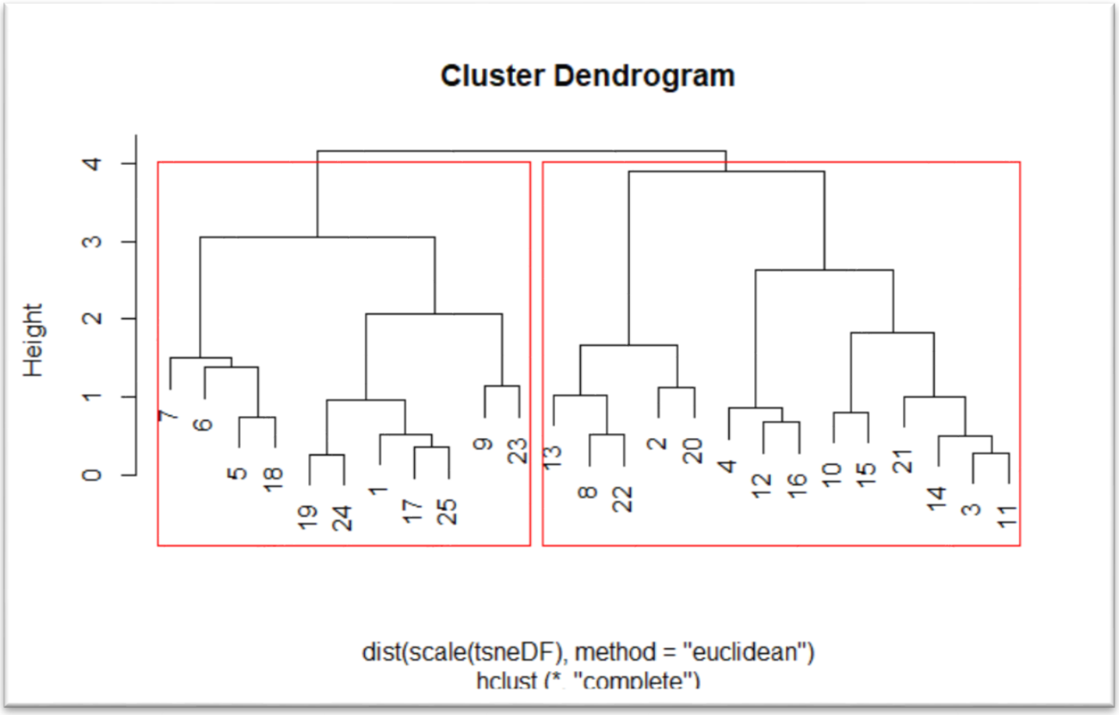


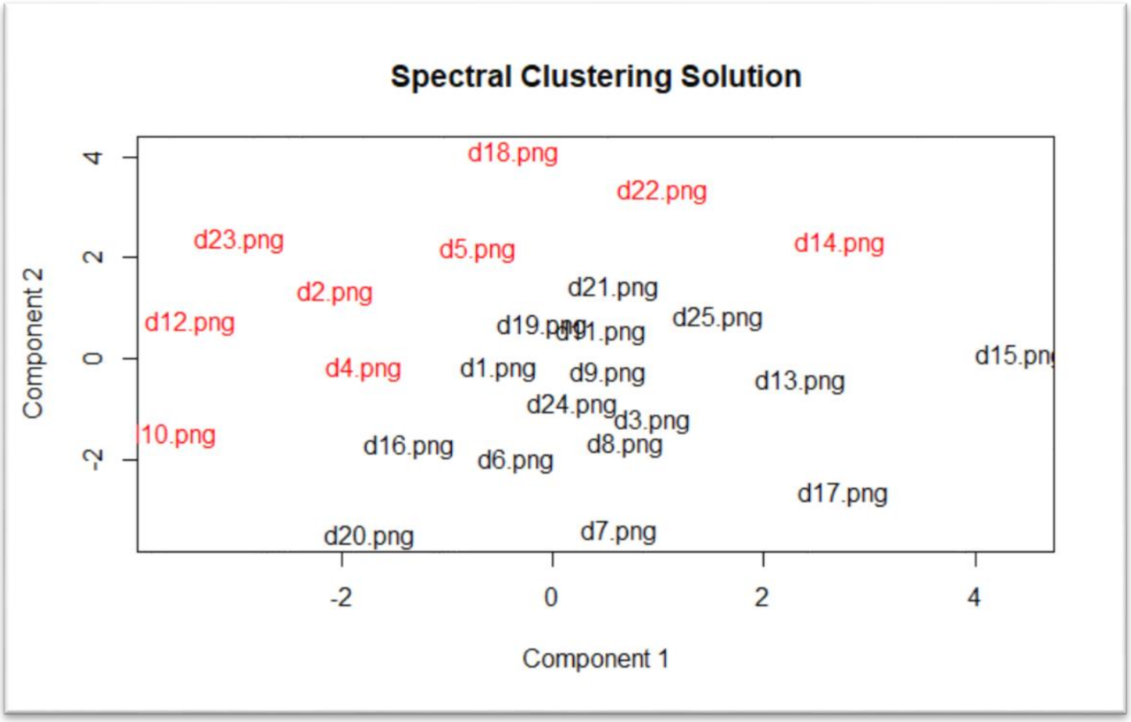
Above plots give a higher level overview of clusters on the dynamic spiral drawings. We can see that there is a better segmentation with Kmeans, Hierarchical and Spectral Clustering, while DBSCAN method clustered all into one group.

Tuned these clustering methods such that the distance between clusters is maximum and within the clusters is minimum. Here are the few more plots to understand the data spread in two clusters with each method.



Hierarchical clustering is tuned with different distance methods and linkages (‘single’, ‘average’, ‘complete’, ‘ward’). Complete linkage with Euclidean distance outperformed other linkages.





Evaluating Clusters

Manually classified the images as drawn by Parkinson's and non-Parkinson's. Created a confusion matrix and calculated accuracy as a measure of metric to compare the clustering results.

KMeans:
Cluster Sizes – 15,10
Accuracy – 60%

	0	1
np	9	5
p	5	6

Hierarchical:
Cluster Sizes – 11, 14
Accuracy – 56%

	0	1
np	7	7
p	4	7

Spectral Clustering:
Cluster Sizes – 16, 9
Accuracy – 40%

	0	1
np	4	10
p	5	6

We see 60% accuracy with K-means clustering. However, this involves manual intervention and may not be accurate way to evaluate clusters. The best way is to test on dynamic including the static spiral drawings.

Testing including Static Spiral Drawings

Results obtained on testing the above fine-tuned cluster methodologies on the dynamic including static spiral drawings are as shown below. Accuracy/error rate are used as a metric to compare the clustering solutions.

KMeans:
Cluster Sizes – 28,22
Accuracy – 62%

	0	1
d	17	8
s	11	14

Hierarchical:
Cluster Sizes – 36, 14
Accuracy – 46%

	0	1
d	17	8
s	19	6

Spectral Clustering:
Cluster Sizes – 6, 46
Accuracy – 48%

	0	1
d	3	23
S	3	22

We can clearly see that k-means outperforms other clustering methods with 68% accuracy and a good split size (27,23) in the clusters. Spectral and Hierarchical clusters tend to group most of the images into one cluster and this may be the case with dynamic drawings too.

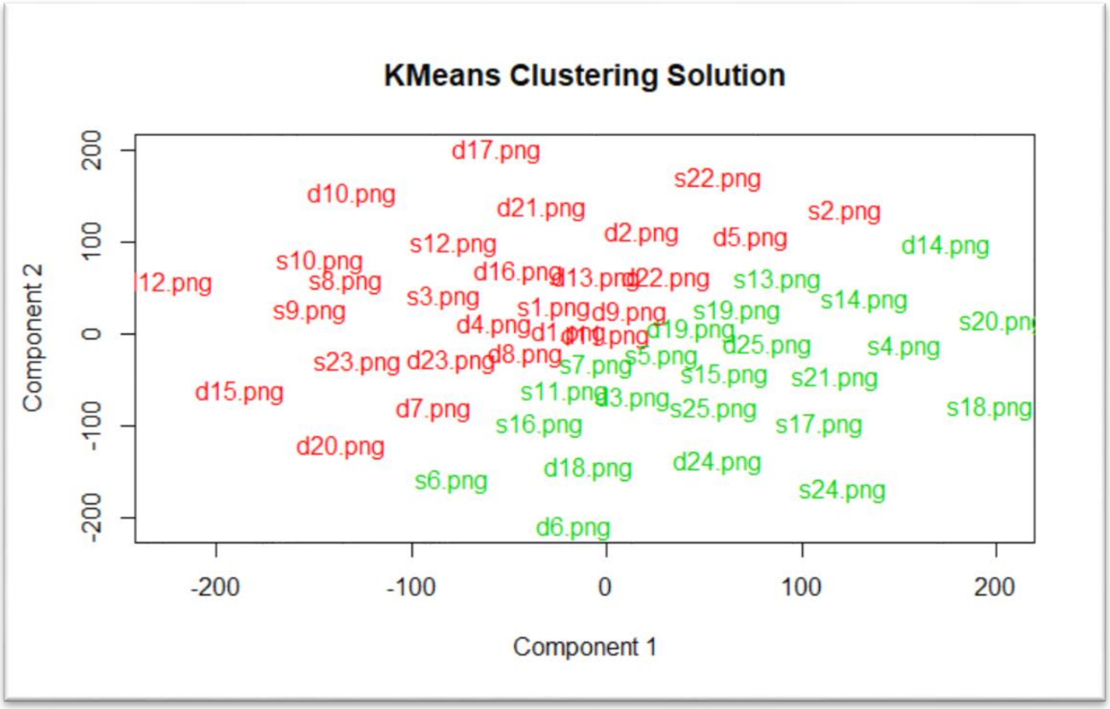
Observations & Further Enhancements

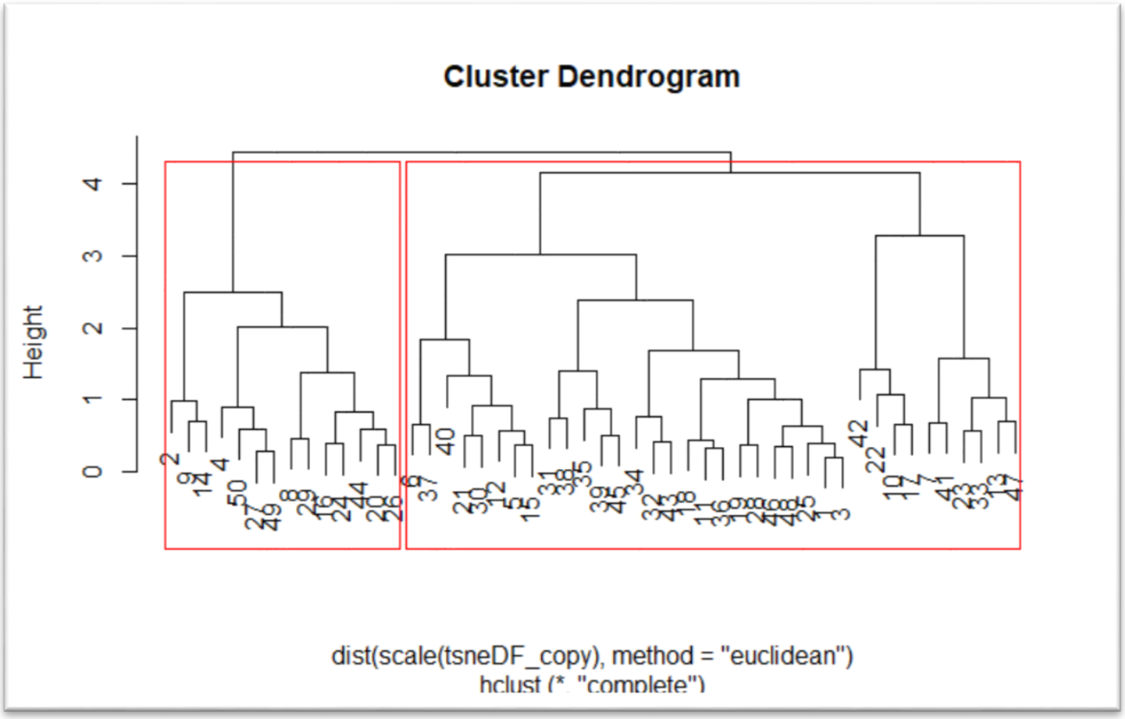
I observed that the 8 dynamic drawings - *d3, d6, d7, d18, d19, d20, d24, d25* that were the only ones classified wrongly as Static were similar in shape but have few spikes along the lines. This made me conclude that the algorithm performs better to a greater extent with regards to grouping similar shapes, however needs to be improved on classifying how they are drawn.

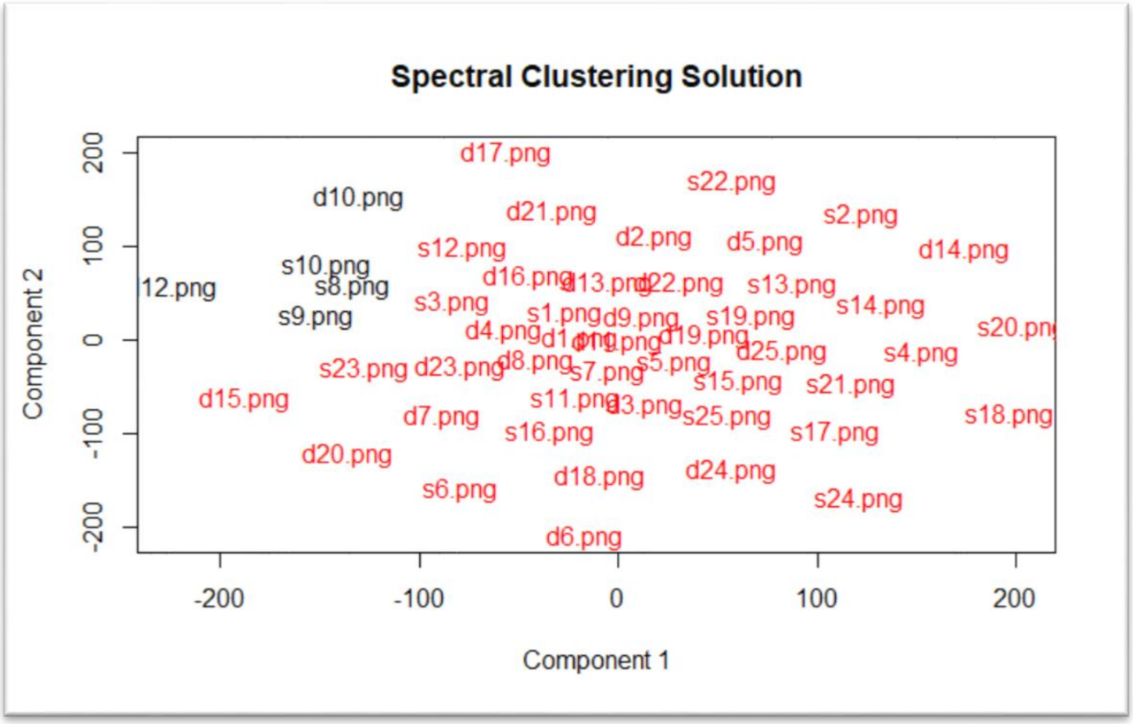
I believe this can be improved by extracting SURF features of an image which includes local feature detectors and descriptors. While t-SNE techniques help retain the structural information, these techniques can be used to compare the spikes along the lines.

Appendix

Plots pertaining to Static and Dynamic Drawings:




Cluster Dendrogram
Height
0
1
2
3
4
2
9
14
4
50
27
49
8
29
19
24
44
20
26
6
37
40
21
30
12
5
15
31
38
35
39
45
34
32
43
18
3
42
22
10
17
7
41
23
33
13
47
dist(scale(tsneDF_copy), method = "euclidean")
hclust (*, "complete")


Spectral Clustering Solution
Component 2
Component 1
200
100
0
-100
-200
-200
-100
0
100
200
d17.png
s22.png
d10.png
d21.png
s2.png
d2.png
d5.png
s12.png
d14.png
s10.png
s8.png
12.png
d16.png
d13.png
d22.png
s13.png
s3.png
s1.png
d9.png
s19.png
s14.png
s9.png
d4.png
d1.png
d11.png
d19.png
s20.png
d25.png
s4.png
s23.png
d23.png
d8.png
s5.png
s7.png
s15.png
s21.png
d15.png
s11.png
d3.png
s25.png
s18.png
d7.png
s16.png
s17.png
d20.png
d18.png
d24.png
s6.png
s24.png
d6.png