September 30, 2019

Experiment 1

1. Input: 2194 images—5 Categories

- Basketball: 225 images
- Fighter: 574 images
- Golf: 558 images
- Soccer: 735 images
- Tennis: 102

Source: https://github.com/SoumitraAgarwal?tab=repositories

2. Added categories to the file names
e.g. Alexander_Volkov → Alexander_Volkov_fighter

3. Merged all images from the 5 categories

4. Created a directory of each image (necessary for Facenet)

5. Face Detection—align images with MTCNN from Facenet
Output:
Total number of images: 2194
Number of successfully aligned images: 510

6. Feature Extraction—Facenet
Output:
Number of images: 2180
Number of batches: 5

- embeddings: 2180 x 512 array
- labels: 2180 x 1 array – indices of the images
- label_string: 2180 x 1 array—file names of the images

Run time: 99.389898777

*I am not sure why 14 images are lost from Facenet during this step

7. Create ground truth dictionary

- The dictionary assigns the indices of the images to the category. The file names can be retrieved from the indices.
- The keys are the categories and the values are arrays of image indicies

e.g. {'tennis': [41, 54, ...], 'basketball': [0, 2,...], 'golf': [...], 'fighter': [...], 'soccer': […]}

8. Create cluster (from algorithm) dictionary

The output of the labels from the clustering algorithms is an 1D array. The cluster dictionary uses the label number as the key and the indices of the image as the values.

e.g. clustering algorithm output: [3, 2, 3, ….]
→ cluster dictionary: {3: [0, 2, …], 2: [1,…]}

9. Create pairs of labels for evaluation using F-Measure

e.g. Ground truth: {‘tennis’: [1, 2, 3], ‘golf’: [4, 5]} → Label pairs: {(1, 2), (1, 3), (2, 3), (4, 5)}

e.g. Cluster labels: {0: [2, 3], 1: [1, 4, 5]} → Cluster pairs: {(2, 3), (1, 4), (1, 5), (4, 5)}

10. Compute F-Measure

Suppose:

L: {(1, 2), (1, 3), (2, 3), (4, 5)} is the ground truth labels

C: {(2, 3), (1, 4), (1, 5), (4, 5)} is the labels from the clustering algorithms

Then:

- True Positive = TP = | L intersect C| = |{ (2, 3), (4, 5)}| = 2
- False Positive = FP = | C – L | = |{ (1, 4), (1, 5)}| = 2
- False Negative = FN = | L – C | = |{ (1, 2), (1, 3)}| = 2
- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F-Measure = 2*((Precision*Recall)/(Precision + Recall))

11. Evaulation from Experiment 1 with different clustering algorithms:

- **2180 images**

Clustering method	Number of clusters	F-Measure	Precision	Recall	False Positives	Runtime (seconds)
K-Means	5	**0.49**	0.52	**0.46**	258,594	1.67
Hierarchical Agglomerative	5	0.44	0.46	0.43	306,756	1.31
DBSCAN min_dist = 1 min_samples = 3	2	0.41	0.26	0.99	1,727,307	6.78
Mean Shift Bandwidth = 1	14	0.41	0.26	0.98	1,712,011	115.03
Spectral	5	**0.45**	0.51	0.4	236,447	**1.09**
EM (Gaussian Mixture Model)	5	**0.50**	**0.57**	0.45	**210,382**	4.31
Birch Threshold = 0.48	5	0.41	0.42	0.41	342,778	1.08
Affinity Propagation	162	0.04	0.64	0.02	5,523	4.05

Hierarchical Agglomerative Clustering Error Examples

False Positives:

Left: Huga Ayala (Soccer)

Right: Ray Borg (Fighter)

Left: Jose Fonte (Soccer)

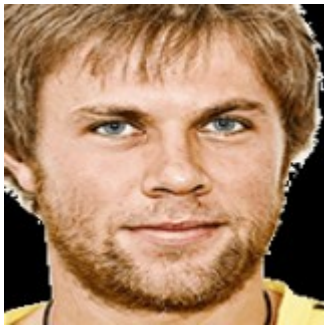
Right: Tommy Hass (Tennis)

False Negatives:

Left: Andy Murray (Tennis)

Right: Joao Sousa (Tennis)

Left: Hyeon Chung (Tennis)



Right: Radu Albot (Tennis)

12. Analysis

- I supposed the best way to evaluate the algorithm is to set the number of clusters to 5 since there are 5 categories.
- Best performing algorithms are EM (Gaussian Mixture Model), K-Means, and Spectral clustering.
- The non-parametric methods (DBSCAN, Mean Shift, Affinity Propagation) have hyperparameters that are hard to tune and the results are either not accurate, have a slow runtime, or both.
- The examples above of the false positives are reasonable as the faces are similar even though they are from different categories.
- The false negatives are also reasonable as ethnicity and appearances like hair color, facial hair may influence this.
    - e.g. Even though Hyeon Chung and Radu Albot are both tennis players they have a completely different look from ethnicity, hair, eye color, etc.

13. Next Steps

- Webscrape the categories (careers) of the faces (labeled with names) from the Labeled Faces in the Wild (LFW) dataset (12,233 images of 5,749 distinct people)
    - Or look for another dataset with labeled names to webscrape
- Experiment 2: Run the experiment again with 4,936 images of 2,013 people
- Experiment 3: Run the experiment again with the whole dataset
- Only consider the parametric methods: K-Means, Hierarchical Agglomerative, Spectral, EM, and Birch Clustering
    - Keeping the number of clusters to 5