Facial Detection and Clustering

Zerun Wang

The Chinese University of Hong Kong

April 11, 2023

Overview

Dlib Facial Detection

Dlib has two trained models for facial detection: one based on Histogram of Oriented Gradients (HOG) + Linear SVM, and the other one based on CNN.
HOG + SVM:

- object shape is characterized using the local intensity gradient distribution and edge direction.
- simple, explainable
- only works with straight and front faces $\rightarrow$ not a good fit for real-time video / event photos

CNN:

- More robust face detection than HOC
- Can utilize GPU accelerator CUDA in Dlib

FaceNet Detection

FaceNet also provides efficient facial detection tools based on Multi-Task Cascaded Convolutional Neural Network (MTCNN).

- MTCNN takes less time to perform detection compared to the CNN model in dlib
- Dlib is more robust to faces with different angles, thus it produces more face chips from the dataset

Comparison

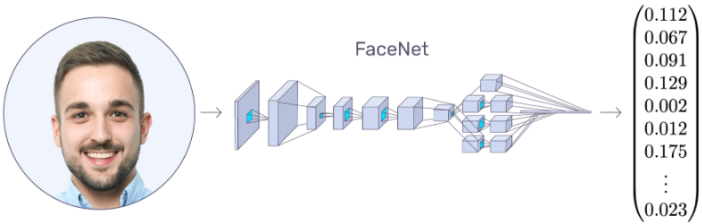
(a) Dlib

(b) FaceNet

Figure: Face detection

Embeddings (Dlib and FaceNet)

- Mapping high-dimensional data into low-dimensional representations (embeddings)
- The network is trained such that the squared L2 distances in the embedding space directly correspond to face similarity.



Chinese Whisper Clustering (CWC)

- Random: running the process on the same network several times can lead to different results
- flat clustering: no hierarchical relations between clusters

The predetermined threshold for the number of the iterations is needed because it is possible that process does not converge.

On the other hand in a network with approximately 10000 nodes the clusters does not change significantly after 40-50 iterations even if there is no convergence

In our study, a threshold 0.5 is handy, but it only gives 21 clusters.

Agglomerative Clustering (AC)

- Hierarchical Clustering: iteration (1) identify the two clusters that are closest together (2) merge the two most similar clusters; until all the clusters are merged together (Bottom up)
- Distance between two clusters measured by Euclidean distance (most popular)
- Linkage function: group objects into hierarchical cluster tree. Includes Ward (minimizes the total within-cluster variance), complete linkage (maximum value of all pairwise distances between clusters), single linkage (min), average linkage...

We initialized cluster numbers to be 70.

Evaluation

As the dataset is not labeled, it is hard to compute the true accuracy of our clustering. However, there are methods that can bypass true label by comparing information within and across clusters.

- Silhouette score : compare the average distance of a data within its cluster and the min distance of it to the other cluster. (cannot apply concentric data, e.g., for DBscan)
- Calinski Harabasz (CH) Score: the ratio of average variance across clusters and average variance within each clusters.

Evaluation

Silhouette score

- between -1 to 1. A desired result should be closer to 1 compared to others;
- practically , larger cluster numbers lead to higher score
- Agglomerative Clustering with 69 clusters gives highest score in hyperparameter tuning

CH score:

- do not have a upper limit. A higher score is preferred;
- practically , less cluster numbers lead to higher score
- CWC method with threshold 0.5 (21 clusters) is perfered compared to AC with clusters 69.

Facial Quality Check

Face quality assessment aims at estimating the suitability of a face image for recognition.

Unsupervised Solution: SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness by Philipp Terhörst et al.

Methods: using the robustness of an image representation as a quality clue

We applied this method to some of our face chips:

(a) Score: 0.7021

(b) Score: 0.4166

Facial Quality Check

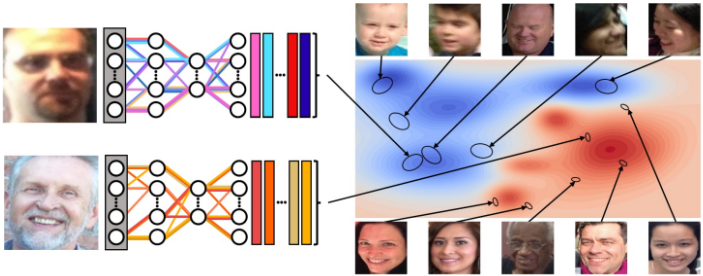


Figure: An image that produces small variations in the stochastic embeddings (bottom left in figure), demonstrates high robustness (red areas on the right) and thus, high image quality. Vise versa.