Clustering Millions of Faces By Identity

Nisim Hurst

Friday 10 May 2019

Clustering Millions of Faces by Identity

The article was written by (**otto2018**). It was was cited 44 times according to Google Scholar. The task performed was face clustering. They used the Pairwise F-measure metric over clusters with distractor images. They also developed their own metric for measuring internal cluster quality using just the k-top nearest neighbors.

Hypothesis

Deep features clustered using only the top-k nearest neighbors in rank-order clustering will produce a more scalable and a more accurate face clustering algorithm. This algorithm will be able to overcome the presence of millions distractor images and class imbalance.

The network architecture to produce a 320D feature vector was VGG16 proposed by (**Simonyan2014**). The rank-order clustering algorithm is based on (**zhu2011**). Their k-d tree implementation for calculating just the 200-top nearest neighbors is based on (**muja2014**).

Evidence and Results

Evidence is presented first over a small dataset and the over an augmented version of the datasets with million of distractor images.

Dataset

The feature extractor was trained with the CASIA-webface. LFW, YTF were used for cluster evaluation, the former over static images and the latter over videos. Webfaces was used to augment the LFW. Here is a brief description of each:

Table 1: Main characteristics of the four datasets that were used to test the improved CW.

	# Instances	Resolution	Scenery	Author
LFW	13233 images of 5749. Only 1680 subjects have two or more photos.	??, variable head angle	Color, different Poses and Backgrounds.	(**huang2008**)
YTF	3425 videos of 1595 subjects.	100x100, variable enclosing area	Color, different Poses and Backgrounds.	(**wolf2011**)
Webfaces	123,654,141 distractor images.	N/A	N/A	(**otto2018**)
CASIA-webface	494,414 images of 10,575 subjects.	120x165	Color, different Poses and Backgrounds.	(**yi2014**)

Results

First, the authors present Pairwise F-measure

Contribution

Firstly, the authors improved the Rank-Order clustering algorithm proposed by (**zhu2011**). The original Rank-Order has the disadvantage that it requires $O(n^2)$. The authors propose to use the FLANN library implementation of the randomized k-d tree algorithm to compute the list of top-k nearest neighbors. Just one iteration is used.

Secondly, the authors improved the internal quality metric of Modularization quality (MQ) (**mancoridis1998**) by just counting shared neighbors in the top-k nearest neighbors list. Cluster's external quality was obviated.

Thirdly, the authors provide an augmented dataset as a matter of baseline to assess the accuracy of the algorithm under the effect of distractor images that are out of the face clusters.

Weaknesses

Future Work