Clustering Millions of Faces By Identity

Nisim Hurst

Tuesday 14 May 2019

Clustering Millions of Faces by Identity

The article was written by (Otto, Wang, and Jain 2018). 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 an approximate 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 (Simonyan and Zisserman 2014). The rank-order clustering algorithm is based on (Zhu, Wen, and Sun 2011). Their k-d tree implementation for calculating just the 200-top nearest neighbors is based on (Muja and Lowe 2014).

Evidence and Results

Evidence is presented first over a small dataset and the over an augmented version of the datasets with millions 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.	(Huang et al. 2008)
YTF	3425 videos of 1595 subjects.	100x100, variable enclosing area	Color, different Poses and Backgrounds.	(Wolf, Hassner, and Maoz 2011)
Webfaces	123,654,141 distractor images.	N/A	N/A	(Otto, Wang, and Jain 2018)
CASIA-webface	494,414 images of 10,575 subjects.	120x165	Color, different Poses and Backgrounds.	(Yi et al. 2014)

Results

First, the authors present Pairwise F-measure evaluated in the LFW dataset without any distractor images. The algorithm obtained the highest F-Measure and lowest run-time. The algorithm is proficient at handling class unbalance.

Second, the authors show performance on the augmented LFW dataset and the decay rate of each algorithm under these conditions. The proposed algorithm shows the highest resiliency compared to the decays ensued in the other algorithm.

Also, given that having a high number of similar frames on each video can affect grouping identities between videos, the authors present the results of the algorithm using a sample of 3 frames per video in contrast to the results obtained over all the frames.

Finally, the authors presents

Contribution

Firstly, the authors improved the Rank-Order clustering algorithm proposed by (Zhu, Wen, and Sun 2011). 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 (Silpa-Anan and Hartley 2008). Just one iteration is used. This approximate version had better performance compared to the exact rank-order and was faster than all the methods tested.

Secondly, the authors improved the internal quality metric of Modularization quality (MQ) (Mancoridis et al. 1998) 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

The method uses a representation that needs to be distributed in chunks across servers, each one process about a million image instances. However, the authors don't provide an efficient algorithm for merging the results nor prove that the algorithm is unaffected in single-thread environments.

Also, the method is dependent of a k that depends on the number of instances, but the authors don't specify how k should be modified. They tested with different

Future Work

Otto et al. mentions that the dimensional vector representation could be improved through a better deep model architect that perform better on profile/side faces.

It would be beneficial to enforce pairwise constraints like must-link and cannot-link. Also the authors

References

Huang, Gary B. et al. (Oct. 2008). "Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments". In: *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie. Marseille, France. URL: https://hal.inria.fr/inria-00321923.

Mancoridis, S. et al. (June 1998). "Using automatic clustering to produce high-level system organizations of source code". In: *Proceedings. 6th International Workshop on Program Comprehension. IWPC'98 (Cat. No.98TB100242)*, pp. 45–52. DOI: 10.1109/WPC.1998.693283.

Muja, M. and D. G. Lowe (Nov. 2014). "Scalable Nearest Neighbor Algorithms for High Dimensional Data". In: *IEEE Transactions on Pattern Analysis and*

Machine Intelligence 36.11, pp. 2227–2240. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2321376.

Otto, C., D. Wang, and A. K. Jain (Feb. 2018). "Clustering Millions of Faces by Identity". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.2, pp. 289–303. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2017.2679100.

Silpa-Anan, C. and R. Hartley (June 2008). "Optimised KD-trees for fast image descriptor matching". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. DOI: 10.1109/CVPR.2008.4587638.

Simonyan, Karen and Andrew Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: pp. 1–14. ISSN: 09505849. DOI: 10.1016/j.infsof.2008.09.005. arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

Wolf, L., T. Hassner, and I. Maoz (June 2011). "Face recognition in unconstrained videos with matched background similarity". In: *CVPR 2011*, pp. 529–534. DOI: 10.1109/CVPR.2011.5995566.

Yi, Dong et al. (2014). "Learning face representation from scratch". In: *arXiv preprint arXiv:1411.7923*.

Zhu, C., F. Wen, and J. Sun (June 2011). "A rank-order distance based clustering algorithm for face tagging". In: *CVPR 2011*, pp. 481–488. DOI: 10.1109/CVPR.2011.5995680.