Person Re-Identification, Face Recognition and Tracking/Clustering

A Project Report

Presented to

Professor Mashhour Solh

Department of Computer Science

San Jose State University

In Partial Fulfillment

Of the Requirements for the Class

Spring-2022: CS 256

By

Atharva Sharma, Abhijn Chadalawada, Devinesh Singh, Nishant Yadav

May 2022

I. ABSTRACT

The availability of high-resolution, cost-effective cameras has created billions of photographs that can be processed by computers to reveal and make sense of the visual media; an important question that comes up in many fields of study is to answer who is depicted in the image. Person Re-Identification, Face Recognition, and Tracking/Clustering are subfields in computer vision that attempt to answer that question. Particularly, face recognition aims to extract specific, distinctive details about a person's face, such as the shape of the face and location of eyes and mouth; extracted features are converted into mathematical representations and used to train a model to classify images. Finally, the model can then be used to re-identify and cluster new images of the same individuals that the model was trained with.

II. INTRODUCTION

Person Re-Identification (Re-ID) – a process that associates and groups images and/or videos of individuals from disjoint angles and cameras – has gained significant popularity among the computer vision community; particularly, Re-ID has gained significant attention for its contributions to security-related applications. Early Re-ID techniques pivoted on the appearance of individuals, such as their clothes, to cluster images; however, these systems would not be able to Re-ID a person if the individual changes their appearance or pose between frames. Instead, we must develop a robust approach to an individual's appearance.

III. LITERATURE REVIEW

Recent works employed a data driven, deep network to learn its representation directly from the pixels of the face [15, 17]; a large dataset of labeled faces is used to attain appropriate invariances to pose, illumination, and other variational conditions.

[8, 11] is based on the *Inception* model of Szegedy *et al*. which was used as the winning approach for ImageNet 2014 [16]. These networks use mixed layers that run several different convolutional and pooling layers in parallel and concatenate their responses.

[15, 17, 23] employ a complex system of multiple stages, that combines the output of a deep convolutional network with PCA for dimensionality reduction and an SVM for classification.

Zhenyao *et al*. [23] employ a deep network to "warp" faces into a canonical frontal view and then use CNN to classify each face as belonging to a known identity. Taigman *et al*. [17] propose a multi-stage approach that aligns faces to a general 3D shape model. A multi-class network is trained to perform the face recognition task on over four thousand identities. Sun *et al*. [14, 15] proposed to use an ensemble of 25 of these networks, each operating on a different face patch.

IV. METHODOLOGY

The overall flow of the process is shown by the flow diagram below :

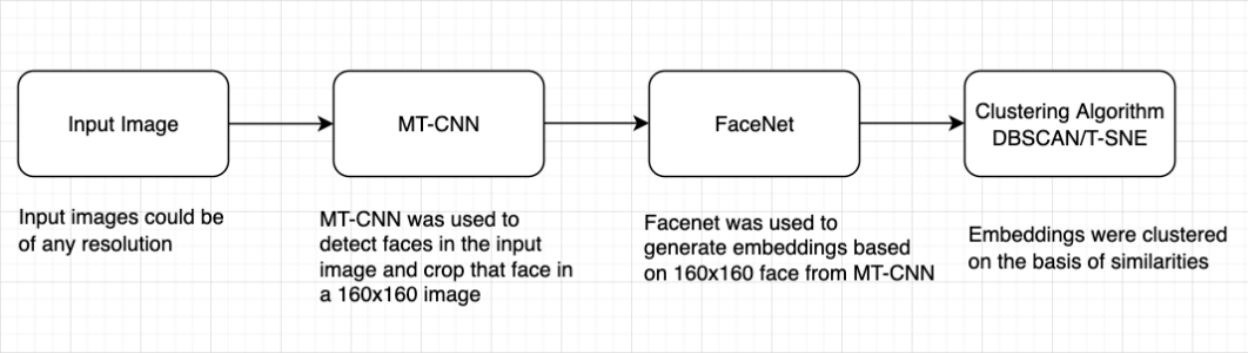


Fig. Flow diagram of clustering process.

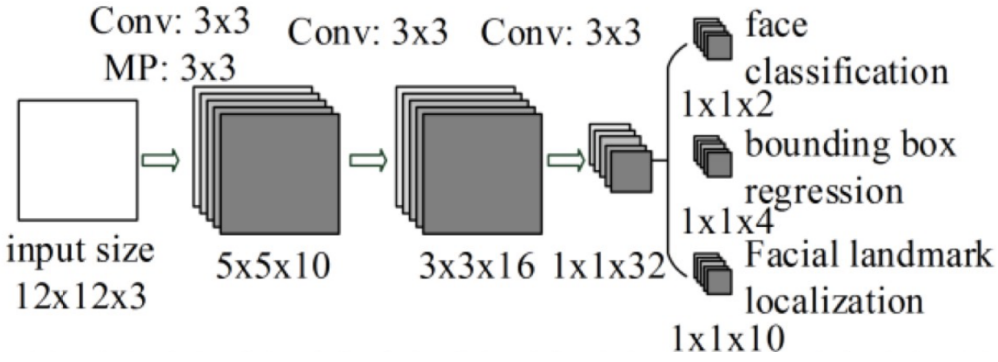
A. FACIAL FEATURE IMPLEMENTATION DETAILS

An alternative to appearance-based Re-ID is the explicit use of facial features to identify and cluster images and videos of individuals. FaceNet, a facial recognition system developed by Google, is used as the backbone of the system; particularly, FaceNet is used to extract features from images of an individual's face, which can then be used to classify images of known identities. FaceNet is trained on the MS-Celeb-1M dataset, a dataset curated by Microsoft, which contains images of 10 million images of faces.

Multi-task Cascaded Convolutional Networks (MTCNN) is a three-staged framework that accounts for both face detection and face alignment; particularly, the network involves three stages of convolutions networks, where each stage refines the models' proposed bounding boxes of landmark locations of facial features, such as eyes, nose, and mouth. [2]. Images are resized into different scales to build an image pyramid, which is used as input into the following three-stages network:

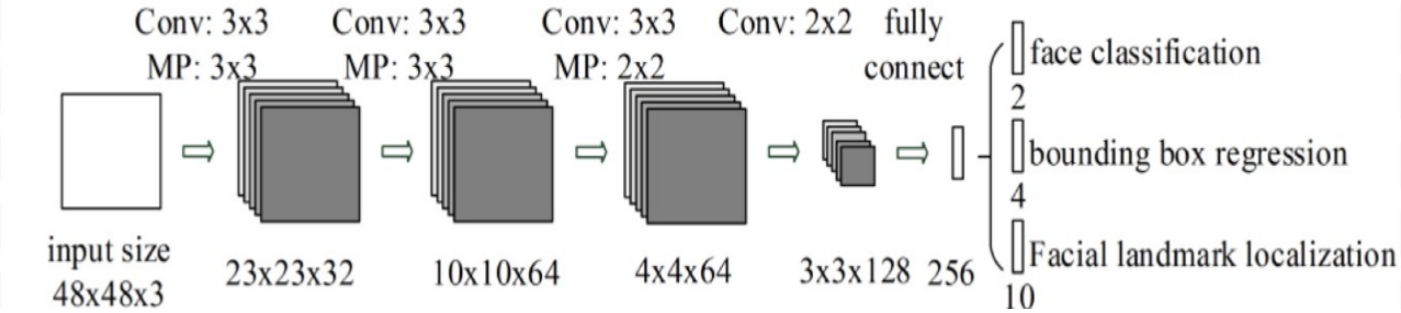
A. Stage 1: Proposal Network (P-Net)

P-Net is a fully convolutional network (FCN) that procures candidate windows and bounding box regression vectors; bounding box regression helps predict the locationalization of boxes to determine the relative location of landmark features. Refinements are required to downsize the number of candidate bounding box candidates.



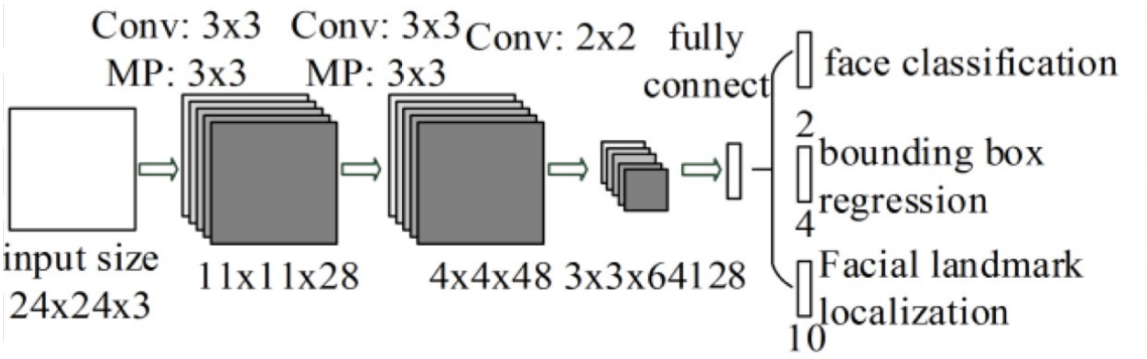
B. Stage 2: Refine Network (R-Net)

R-Net is a series of convolutional neural networks (CNNs), that are used to further reduce the number of candidate bounding boxes found in the P-Net. Output of the R-Net is as follows: 1) image of a face, 2) 4 element vector to represent the bounding box for the face, 3) 10 element vector for facial landmark localization.



C. Stage 3: Output Network (O-Net)

O-Net further refines the output of the R-Net and aims to describe the five facial landmark positions for eyes, nose, and mouth.



B. ***DATA COLLECTION AND PRE-PROCESSING***

10 samples of 5 athletes were downloaded from Google, where each image focused on the athlete without obstructions and outliers.

Fig 1. Initial data set without outliers.

After one round of testing on the initial dataset, four more athletes were added to the dataset. This time, to test the limitations of MT-CNN, the images included a second, unknown athlete. Fig. 2 shows an example of an image with a second athlete, to add noise to the dataset.

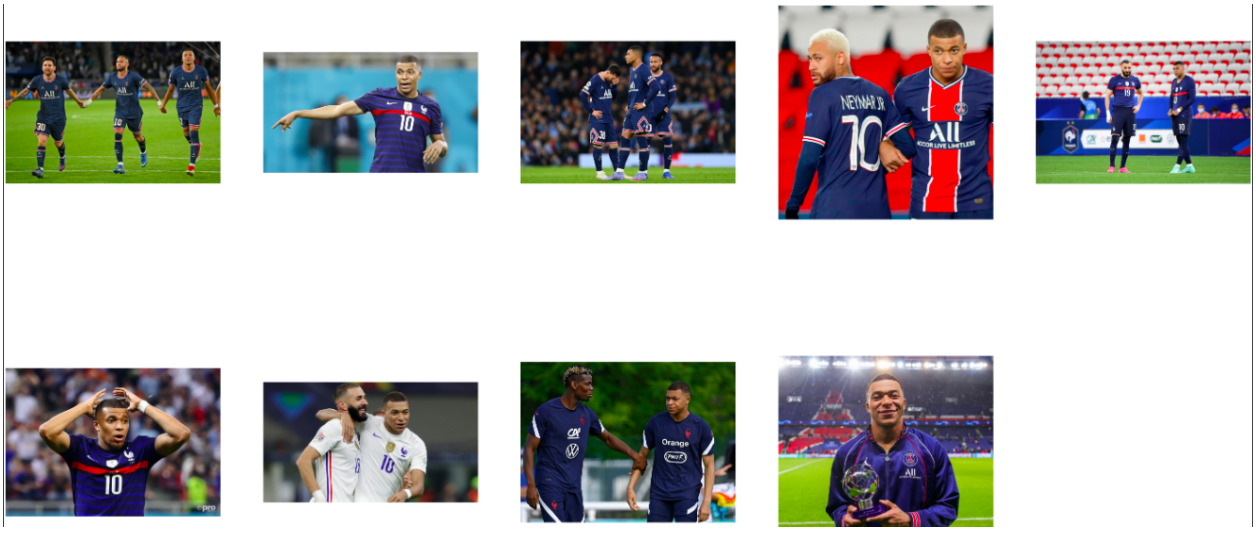


Fig 2. Sample data with added outliers.

Apart from the outliers, a miscellaneous folder of images was also added in the dataset to see clustering algorithms robustness against noise in the data. Ultimately, the training dataset was of 1M images and the testing set had 93 images.

C. DATA CLUSTERING ALGORITHMS AND TECHNIQUES

2 clustering algorithms and techniques were compared to determine which algorithm performed best for the purpose of clustering faces.

A. t-distributed stochastic neighbor embedding (t-SNE)

t-SNE is a two-staged unsupervised, non-linear statistical method used to visualize high-dimensional data; unlike principal component analysis (PCA), t-SNE preserves local similarities, whereas PCA aims to preserve large pairwise distances to maximize variance.

Stage 1 of t-SNE measures the similarity between points in a high-dimensional space using Gaussian distribution; points are normalized to give a set of similarities.

Stage 2 of t-SNE uses the student t-distribution with one degree of freedom to get a second set of probabilities; the final representation of the point in lower dimension is computed from the difference of the Gaussian and t-distribution and gradient descent is applied to minimize the KL loss.

B. Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is a density-based clustering algorithm used to group points that are closely packed; outliers, points that lie in low-density regions, are marked to indicate the point should be ignored.

IV. ***RESULTS AND DISCUSSION***

A. TESTING WITHOUT OUTLIERS

For the initial dataset which did not have any outliers, MT-CNN generated correct cropped faces. These cropped faces only contained one athlete's face from one image.

Fig. MTCNN output for images containing only 1 athlete

With our initial dataset of 5 categories, 10 images per category and no outliers. We got well defined clusters with T-SNE clustering. These clusters were highly separated and images with the same faces were clustered close together.

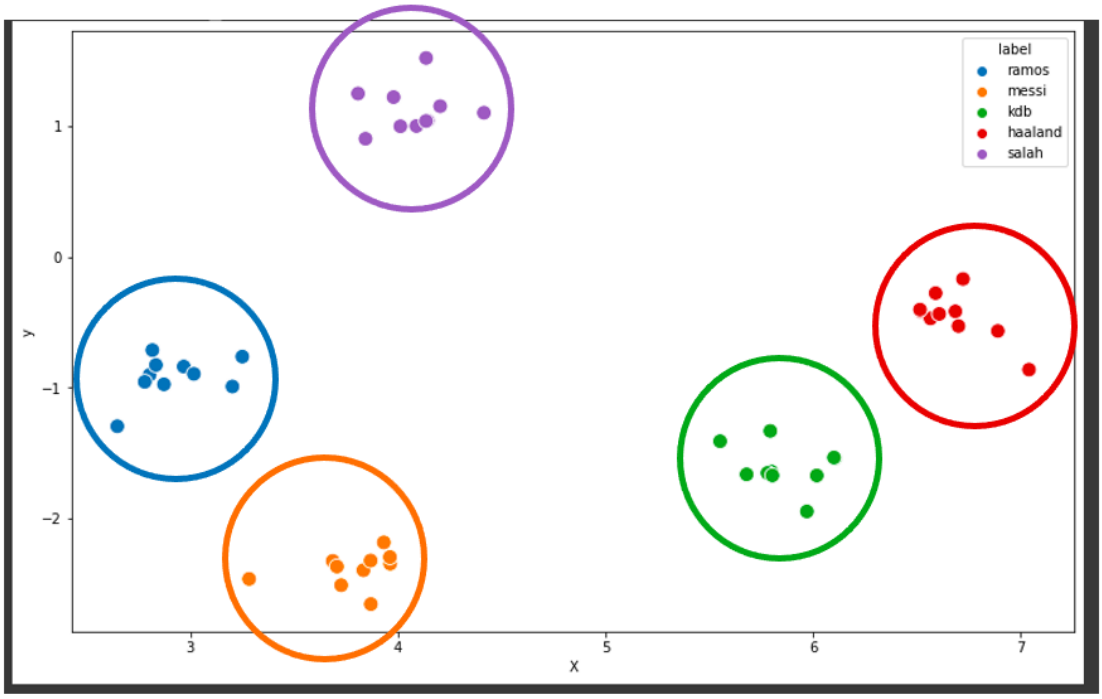


Fig : Clustering generated by T-SNE clustering on initial dataset

Another clustering algorithm that was tried was DBSCAN, it did not produce well defined results on the embeddings of the faces.

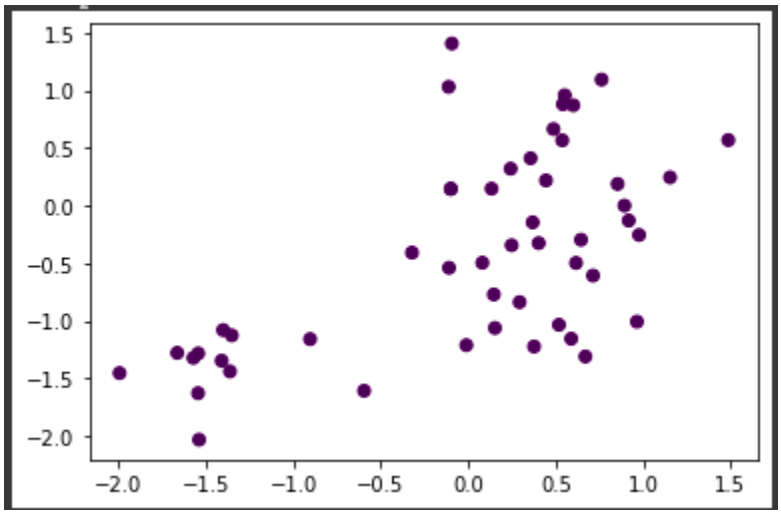


Fig : Clustering generated by DBSCAN on initial dataset

B. ***TESTING WITH OUTLIERS***

With outliers added to the dataset, one image could contain multiple faces. This confused MT-CNN and it could not generate the croppings of both the faces. It cropped the first face it could see and generated the 160x160 image.

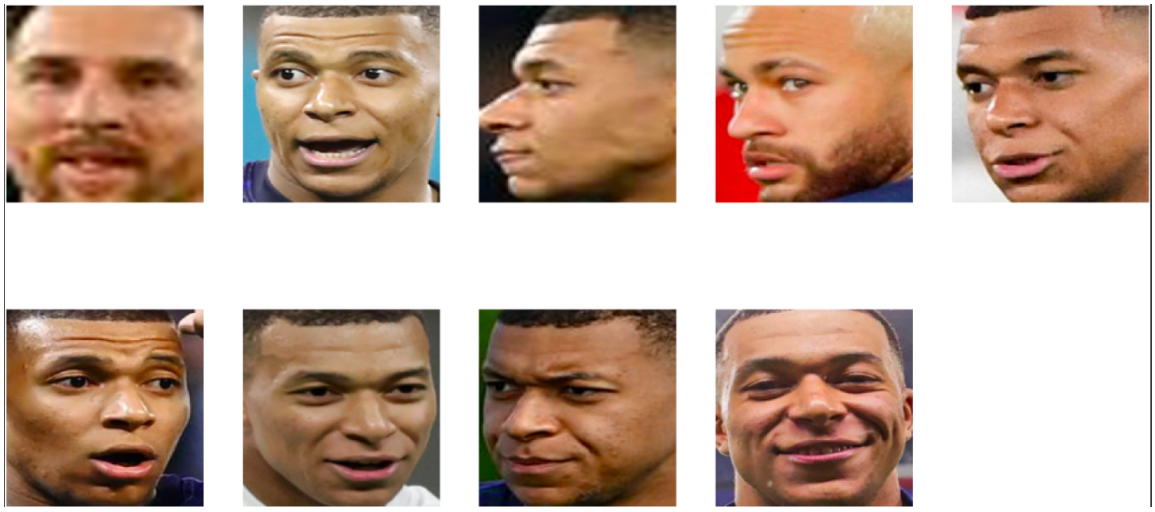


Fig : Cropping generated by MT-CNN on outlier dataset

As it can be seen from the image above, MT-CNN could only crop the first face it detected in the image. This caused non-needed faces to be considered as input for facenet. This generated incorrect embeddings and made the accuracy of the clustering algorithm to drop down.

As it can be seen in the figure below, the reason why we are seeing outliers in the clusters as well is because of the MT-CNN cropping out wrong faces. FaceNet is still performing as expected.

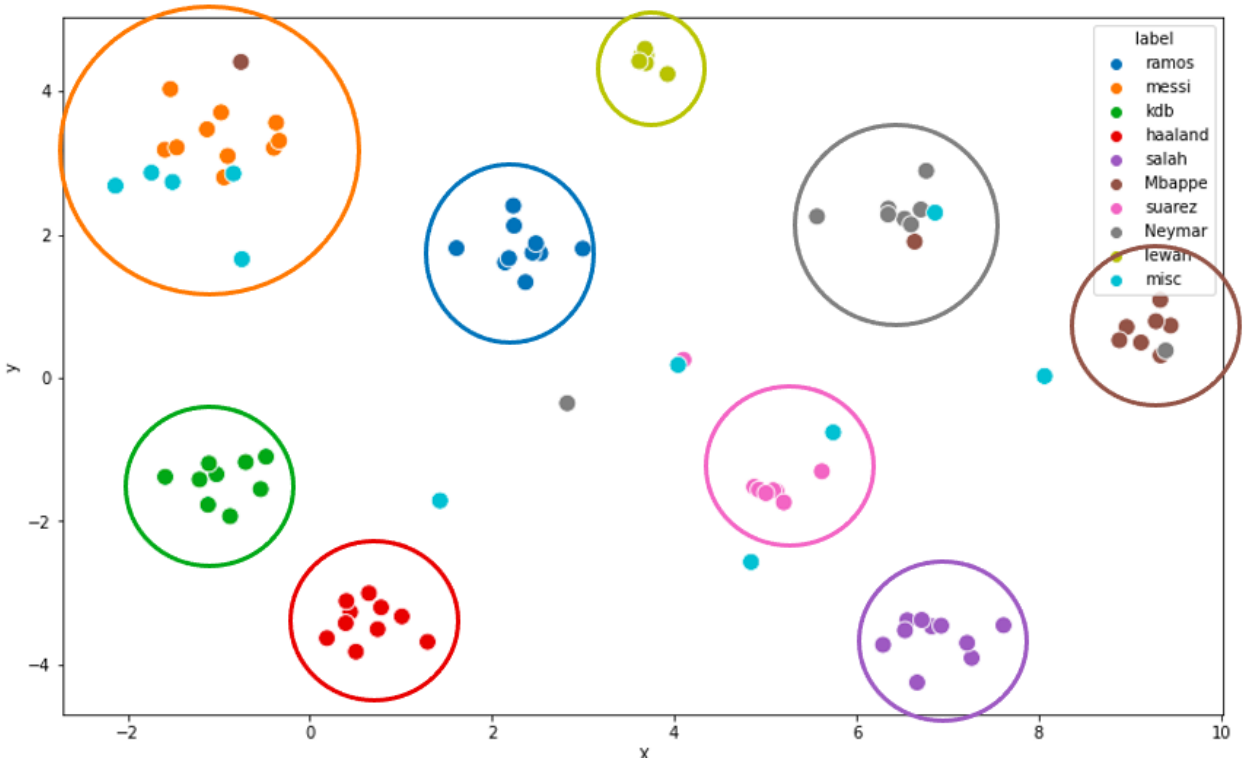


Fig. Clustering generated by T-SNE on the dataset with outliers.

Considering one specific cluster displayed in the image below : All the orange dots represent the correct categorisation. The blue and the brown dots are the images of the same athlete. But the images represented by differently colored dots are cropped by MT-CNN from a picture of

another athlete. Where the generated face was somewhere in the background and not the focus of the image.

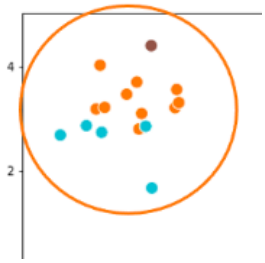


Fig. Sample cluster with outliers.

REFERENCES

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Cur- riculum learning. In *Proc. of ICML*, New York, NY, USA, 2009. 2

[2] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. ECCV*, 2012. 2

[3] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *Proc. ECCV*, 2014. 7

[4] J.Dean,G.Corrado,R.Monga,K.Chen,M.Devin,M.Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng. Large scale distributed deep networks. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *NIPS*, pages 1232–1240. 2012. 10

[5] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011. 4

[6] I. J. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *In ICML*, 2013. 4

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Re- port 07-49, University of Massachusetts, Amherst, October 2007.

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, Dec. 1989. 2, 4

[9] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 2, 4, 6

[10] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with gaussian face. *CoRR*, abs/1404.3840, 2014. 1

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 1986. 24

[12] M. Schultz and T Joachims Learning a distance metric from relative comparisons. In S. Thrun, L. Saul, and B. Schölkopf editors, *NIPS*, pages 41–48. MIT Press, 2004. 2

[13] T.Sim,S.Baker,andM.Bsat.TheCMUpose,illumination, and expression (PIE) database. In *In Proc. FG*, 2002. 2

[14] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *CoRR*, abs/1406.4773, 2014. 1, 2, 3

[15] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *CoRR*, abs/1412.1265, 2014. 1, 2, 5, 8

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 2, 3, 4, 5, 6

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verifica- tion. In *IEEE Conf. on CVPR*, 2014. 1, 2, 5, 7, 8, 9

[18] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. *CoRR*, abs/1404.4661, 2014. 2

[19] K.Q.Weinberger,J.Blitzer,andL.K.Saul.Distancemetric learning for large margin nearest neighbor classification. In *NIPS*. MIT Press, 2006. 2, 3

[20] D. R. Wilson and T. R. Martinez. The general inefficiency of batch training for gradient descent learning. *Neural Net- works*, 16(10):1429–1451, 2003. 4

[21] L. Wolf, T. Hassner, and I. Maoz. Face recognition in un- constrained videos with matched background similarity. In *IEEE Conf. on CVPR*, 2011. 5

[22] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. 2, 3, 4, 6

[23] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical- view faces in the wild with deep neural networks. *CoRR*, abs/1404.3543, 2014. 2