Deep Face Recognition

Omkar M. Parkhi
omkar@robots.ox.ac.uk

Andrea Vedaldi
vedaldi@robots.ox.ac.uk

Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

The goal of this paper is face recognition – from either a single photograph or from a set of faces tracked in a video. Recent progress in this area has been due to two factors: (i) end to end learning for the task using convolutional neural networks (CNNs), and (ii) the availability of very large scale training datasets. We make two contributions: first, we show how a very large scale dataset (2.6M images spanning more than 2.6K identities) can be constructed by semi-automatic annotations with humans in the loop, investigating the trade-off between annotation purity and cost; second, we introduce a very deep convolutional neural network and a corresponding training procedure that achieve face recognition accuracy comparable to the current state of the art on public benchmarks such as "Labelled Faces In the Wild" and "YouTube Faces Dataset", while at the same time using a fraction of the data used by competitors.

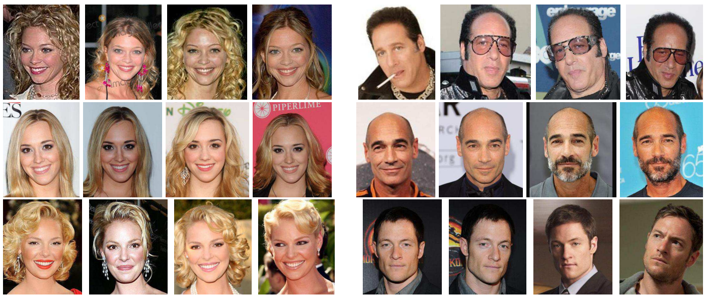

Figure 1: Example images from our dataset.

The availability of large scale datasets such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [4], MIT places [10] and Microsoft COCOA [2] have been instrumental in the giant CNN based leapforward we have witnessed in the community. Such datasets however, were missing from the face recognition domain. Most recent advancement in the field have been from the internet giants like Facebook and Google [5, 8, 9]. For example, the most recent face recognition method by Google [5] was trained using 200 million images and eight million unique identities. The size of this dataset is almost *three orders of magnitude* larger than any publicly available face dataset (see Table 1). The first part of this paper proposes a procedure to create a reasonably large face dataset whilst requiring only a limited amount of person-power for annotation. One of the key ideas was to use weaker classifiers to rank the data presented to the annotators. This procedure has been developed for faces, but is evidently suitable for other object classes as well as fine grained tasks. We employ this procedure to build a dataset with over two million faces, and will make this freely available to the research community.

Dataset	People	Images
LFW	5,749	13,233
WDRef [1]	2,995	99,773
CelebFaces [7]	10,177	202,599
Ours	**2,622**	**2.6M**
FaceBook [8]	4,030	4.4M
Google [5]	8M	200M

Table 1: **Dataset comparisons:** Our dataset has the largest collection of face images outside industrial datasets by Goole, Facebook, or Baidu, which are not publicly available.

The second part of this paper investigates various CNN architectures for face identification and verification, including exploring face alignment and learning the task specific embeddings, using the novel dataset for training. Many recent works on face recognition have proposed numerous variants of CNN architectures for faces, and we assess some of these modeling choices in order to filter what is important from irrelevant details . The outcome is a much simpler and yet effective network architecture without any embellishments but with appropriate training, achieving near state-of-the-art results on all popular image and video face recognition benchmarks using a single network (see Table 2 and 3). Again, this is a conclusion that may be applicable to many other tasks.

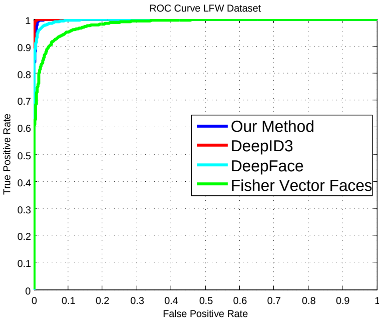

Figure 2: ROC curve for the LFW dataset unrestricted protocol setting.

No.	Method	Images	Networks	Acc.
1	Fisher Vector Faces [6]	-	-	93.10
2	DeepFace [8]	4M	3	97.35
3	Fusion [9]	500M	5	98.37
4	DeepID-2,3		200	99.47
5	FaceNet [5]	200M	1	98.87
6	FaceNet [5] + Alignment	200M	1	99.63
7	Ours	2.6M	1	98.95

Table 2: **LFW unrestricted setting.** Left: we achieve comparable results to the state of the art whilst requiring less data (than DeepFace and FaceNet) and using a simpler network architecture (than DeepID-2,3). Note, DeepID3 results are for the test set with label errors corrected – which has not been done by any other method.

No.	Method	Images	Networks	100%- EER	Acc.
1	Video Fisher Vector Faces [3]	-	-	87.7	83.8
2	DeepFace [8]	4M	1	91.4	91.4
3	DeepID-2,2+,3		200	-	93.2
4	FaceNet [5] + Alignment	200M	1	-	95.1
5	Ours	2.6M	1	92.8	91.6
6	Ours + Embedding learning	2.6M	1	97.4	97.3

Table 3: **Results on the Youtube Faces Dataset, unrestricted setting.** The value of k indicates the number of faces used to represent each video.

[1] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. ECCV*, pages 566–579, 2012. 1

[2] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 1

[3] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *Proc. CVPR*, 2014. 1

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, S. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and F.F. Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[5] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, 2015. 1

[6] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE PAMI*, 2014. 1

[7] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014. 1

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deep-Face: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014. 1

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *Proc. CVPR*, 2015. 1

[10] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. *NIPS*, 2014. 1