WEAKLY-SUPERVISED DEEP SELF-LEARNING FOR FACE RECOGNITION

Binghui Chen, Weihong Deng

Beijing University of Posts and Telecommunication, No 10, Xitucheng Road,
Haidian District, Beijing, PR China
{chenbinghui, whdeng}@bupt.edu.cn

ABSTRACT

For recent years, state-of-the-art deep learning systems for face recognition task completely use supervised training. Their performances depend critically on the amount of manually-labeled examples and the correctness of label data. In real life, however, it is very costly and time-consuming to collect and label such database. Therefore, we intend to build a feasible self-learning system, handling the face images which are unlabeled. In this paper, we first build a challenging unlabeled database and propose an efficient Self-Learning DCNN structure (SL-DCNN) to handle weakly-supervised training for face recognition using complicated and unlabeled training data. Our main contribution is that we introduce a novel modification signal as an ingenious supervision to distinguish the misclassifications, and to correctly reduce intra-class variations and enlarge inter-class distances in combination with identification-verification. Then, we investigate the method of feature merging and whether rich identity improves feature learning under unlabeled data. Finally, 97.47% face verification accuracy on LFW [1] is impressively achieved by our method, which is much higher than state-of-the-art methods under noise.

Index Terms— weakly-supervised, DCNN, modification, face recognition, self-learning

1. INTRODUCTION

Deep Convolutional Neural Networks have taken the computer vision field by storm, significantly improving the state-of-the-art performances in face recognition. Deep learning methods have shown more impressive performances than traditional shallow models which represent faces with over-complete low-level features, such as [2] [3]. Deep models [4] [5] [6] [7] [8] [9] have been proved to be effective to extract high-level visual features. Throughout the history of LFW benchmark, deep learning models [8] [10] [11] [12] [13] have achieved surprising improvements in face verification, wildly considered as state-of-the-art. And [8] [10] [11] [12] are trained with two supervised signals (i.e. identification, verification signals) using deep convolutional neural networks. The only difference among them is that [12] is retrained by embedding loss in second round. In [13], the significant result can also be achieved by naive CNN structure followed by PCA feature reduction. It is obvious that the deep models are powerful in face verification whether the identification signal is used in combination with verification signal at the same stage or not.

However, most deep learning results of state-of-the-art models mentioned above are based on their training datasets. Most are trained with big accurate database collected from Internet.The premise of regarding these database as training examples is that the class labels are precise. Unfortunately, this assumption often does not hold in real life. And there is little work on training with unlabeled database for face recognition using deep convolutional neural networks.

In this work, we collect large-scale unlabeled training samples sets which are automatically downloaded from the search engine by keywords. Then to learn representative features from unlabeled data, we propose a Self-Learning DCNN structure by using a novel modification signal which can be regarded as self learning to bootstrap the identification and verification phase. Motivated by Scott et al. [14], we take advantage of the prediction, the output probability distribution over the n classes, to bootstrap the learning phase. Because in traditional training phase, useful features which are extracted in the top hidden layer before the n-way softmax layer is obtained by exact classification. And in consideration of the possibility of the existence of false labels in our training set, when the labels of training examples are wrong, the learning stage can be influenced on different levels. Particularly, in verification stage, a pair of images with the same weak label may belong to different classes. Thus, the closer we keep these two feature vectors extracted from the ”same” identity, the worse we do. Modification signal is to correctly handle the weak identity labels or pair labels and it will dominant the process of learning by potential right labels. And the distance between a pair of images can be rightly tackled with.

To evaluate our work from different aspects, complementary experiments have been done. We took DeepID2 and Naive Deep models as our baseline, tested them on our noisy database for comparison and investigated whether rich identity information would improve the feature learning under the environment of unlabeled examples. We also investigate that which signal is mainly affected on by our modification signal and the schedule of multi-region merging to improve the performance of our model. Then we report the results of our method against the state-of-the-art approaches. And training with these three signals, we finally have achieved a significantly 97.47% face verification accuracy on LFW benchmark, which is nearly 9% higher than DeepID2 [10] and 4% higher than Naive Deep (two patches input) [13] trained on our database.

2. UNLABELED WEB FACE DATABASE

Throughout the varied databases collected by many institutions, they are marginally labeled. To minimize the cost in the collect stage, we are the first to build a straightforward database, referred to as Unlabeled Web Face (UWF) database. In this section, we will introduce our collected database and show its difficulty in details.

The costless database, which is automatically downloaded from Google Image Search by keywords without any extra hand-label, contains 0.45M large-scale (250*250) and high resolution RGB face images of 14,416 individuals (celebrities). That is to say, the face images only have keyword labels (weak labels). Most worthy of mention is that people in UWF and LFW are mutually exclusive. Subsequently, it is randomly separated into three parts followed by naive alignment, i.e. training set (0.43M images) and validation set (0.01M images) are from 14316 identities, test set (3K images) is from 100 identities.

Fig. 3 shows the distribution of our UWF database. Due to the reason of no hand-label, the distribution looks more uniform and have a little differences with the discovery of long-tail characteristic of web-collected data by Zhou et al [14]. And the image number of each individual is approximately 31 on average.

Without manually labelling, we could not explain exactly how much proportion of negative classification. However, we still believe that the proportion of wrong examples is considerable. Fig. 2 shows a part of examples with the same identity (randomly selected). To demonstrate more clearly, we also have labeled the test set and Fig. 1 shows the result. So we have reason to believe the challenging characteristic of our UWF database. And the bad performances of the state-of-the-art presented in Sec. 4.5 assist to prove it.

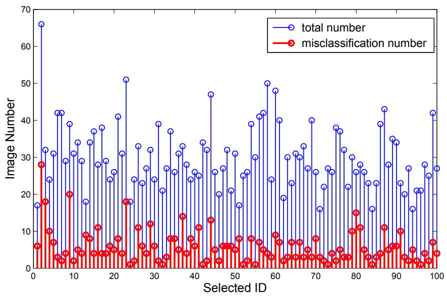

Fig. 1. The misclassification ratio of 100 individuals is 19.61% in average.

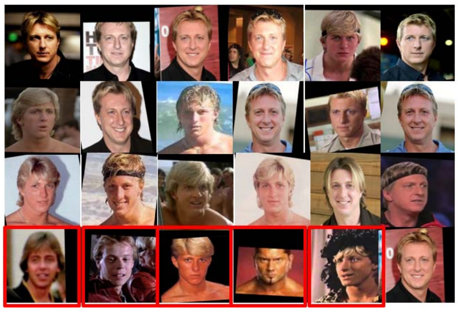

Fig. 2. Part of instance images from UWF database for one identity. Red box marked images indicate that they do not belong to this class in essence.

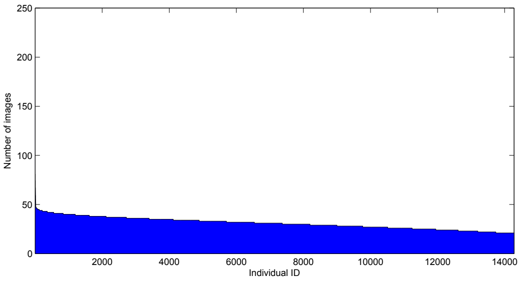

Fig. 3. The Distribution of UWF Database. All individuals are sorted by the number of instances.

3. SL-DCNN MODEL AND TRAINING DETAILS

Bootstrapped by modification signal, the deep features can be learned well. And this section describes the SL-DCNN model and its training details.

3.1. SL-DCNN architecture

In our experiment, the input to our network is a fixed-size 104×96 RGB face image ,which is cropped from the 250×250 image, with the average face image(computed from the training set) subtracted. Inspired by [15], our network contains eleven layers, i.e. nine convolutional layers and two fully-connected layers. Each of the nine convolutional layers is followed by ReLU. And we use filters with a small receptive field: 3×3 throughout the convolutional layers. The convolution stride is fixed to 1 pixel and the spatial padding for conv1-conv6 is 1 pixel, 0 pixel padding for other conv.layers. Then there is a Dropout [16] layer with 0.3 dropout-ratio following Conv9. Spatial pooling is carried out by three max-pooling layers. Max-pooling is performed over a 2×2 window, with stride 2. The last layer is a n-way(e.g. n=14316) softmax layer which is set in training phase for weakly-supervised learning. The fully-connected

Table 1. The architecture of SL-DCNN

	Conv1	Conv2	MaxP1	Conv3	Conv4	MaxP2	Conv5	Conv6	MaxP3	Conv7	Conv8	Conv9	FC	softmax
support	3	3	2	3	3	2	3	3	2	3	3	3	1	1
num filts	64	64	-	128	128	-	256	256	-	512	512	512	3000	14316
pad	1	1	0	1	1	0	1	1	0	0	0	0	0	0
Stride	1	1	2	1	1	2	1	1	2	1	1	1	1	1
ReLU	Y	Y	N	Y	Y	N	Y	Y	N	Y	Y	Y	Y	-

layer is 3000 dimension and is regarded as features layer. An illustration of the SL-DCNN structure used to extract representative features is shown in Table. 1. At the verification process, we measure the similarity between two images through a simple Cosine Distance.

3.2. Modification signal guides feature learning

The representative features are learned with weakly-supervised signals. To correctly deal with the misclassifications, modification signal plays part in identification-verification stage. And the core idea of modification is taking into account the class of maximum predicted probability by the model which are regarded as predicted label p when given an input image, and guiding feature learning in combination with weak label y.

Identification bootstrapped by modification signal is achieved by the n-way (e.g. n=14316) softmax layer. Our network is trained to minimize the following loss function, which was proposed by Flatow et al.[17], and we call it Identification-Modification loss (L_{IM}). In accordance with the softmax loss, it is denoted as

$$L_{IM} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}((1-v)y_{ij} + vp_{ij})log(P_{ij}) \tag{1}$$

where capital letter P_{ij} is the predicted probability respect to jth class, given the ith input image. y_{ij} is the weak label, where $y_{ij} = 0$ for all j except $y_{ij} = 1$ for the target class. Lowercase $p_{ij} = 1\{j = \arg\max P_{ij}, j = 1 \sim n\}$ is an indicator for whether j is the class of highest predicted probability by the model for the ith input data. And v is the validation accuracy. When the validation accuracy increases, the formula of loss function will change accordingly, i.e. at initial, $v = 0$ and only weak label guides the model learning, when the value of v increases we will leverage the prediction by our model. That is to say, we do not trust our model's prediction at initialization at all until it has the learning capacity. In identification, modification is embodied in this convex combination of predicted label and original label, $(1 - v)t_{ij} + vz_{ij}$, with continuously modifying the affects of weak labels of the current input face images.

Without wrongly handling the distance between two images, verification guided by modification signal can effectively encourage the deep features learning. For example, it will enlarge the distance between feature vectors extracted from the "same" class because they belong to different subjects in nature. We take the following loss function based on L2 norm, which we call verification-modification loss (L_{VM}),

$$L_{VM} = \begin{cases} \frac{1}{2}M & if\ t_{ab} = z_{ab} = 1 \\ \frac{1}{2}(\beta M + (1-\beta)\max(0, mar - M)) & if\ t_{ab} = 0\ \&\ z_{ab} = 1 \\ \frac{1}{2}\max(0, mar - M) & if\ z_{ab} = 0 \end{cases} \tag{2}$$

where $M = \left\|f_a - f_b\right\|_2^2$, f_a and f_b are the deep feature vectors extracted from a pair of images. M indicates their L2 distance. β and mar are hyper-parameters, and we will discuss some results about β afterwards. t is the predicted label and z is the weak label. $t_{ab} = 1\ or\ 0$ means whether these two face images are predicted to belong to the same class. While, $z_{ab} = 1\ or\ 0$ means whether they have the same weak identity. When $t_{ab} = z_{ab} = 1$, it will minimize the L2 distance between these two feature vectors. And when $z_{ab} = 0$, it directly requires the distance larger than the margin "mar" because in this case the probability of belonging to the same class is very low. However, when the results of prediction are diverse from the weak labels, that is when $t_{ab} = 0\ and\ z_{ab} = 1$, it will take into account the convex combination of inter-class and intra-class distances.

By combining Eq. 1 and Eq. 2, we formulate our system as the following optimization problem:

$$\min_{W,b} L = L_{IM} + \frac{\lambda}{m}\sum_{m} L_{VM} \tag{3}$$

where λ is hyper-parameter and m indicate the batch size.

Overall, modification signal encourages right classifications by weakening the negative effect of weak labels. Especially at verification stage, given the potential right labels, intra-class variations are significantly reduced with few mistakes.

3.3. Training details

We trained our SL-DCNN using SGD(stochastic gradient descent) with a mini-batches of 50 samples and momentum of 0.9. To regularise our network, we use popular weight decay and set it to 0.005. The dropout layer can also regularise our model with a rate of 0.3 outlined above. Then the hyper-parameter margin mar and λ are separately set to 1 and 0.01. We set our learning rate to 0.01 at initial stage and then it was decreased by factor of 10 when the validation set accuracy stopped increasing. Overall, the learning rate decreased four times throughout the learning process.

The initialisation of the network weights is very important, since traditional initialisation procedure of random sampling from a Gaussian distribution can hardly guide the model learning due to instability of training examples and deep net. So we adopted the method of progressive training. That is to say, we will train a larger model based on a well trained small model. So the training sequence of our models is that 6 layers net (lacking conv2, conv4, conv6, conv8 and conv9 layers mentioned in Table. 1), 9 layers net (lacking conv2 and conv4 layers mentioned in Table. 1) and 11 layers net. And we will also show the performances of these two relative smaller models in Sec. 4.5. As a result, without considering the convergence of deep network, we can easily train our model with the complicated training set. During training, the input was randomly sampled and mirrored with 50% probability. However, we did not perform any colour channel augmentation as described in [18].

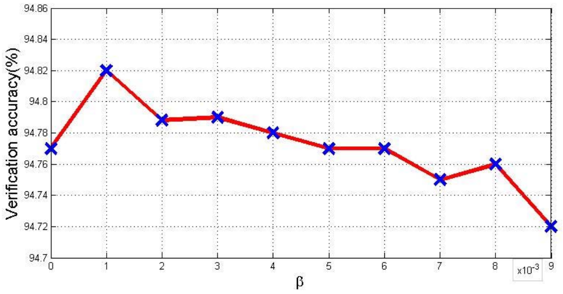

Fig. 4. Face verification accuracy by varying the mixture parameter β. We just show the corresponding accuracy with β varying from 0 to 0.009.

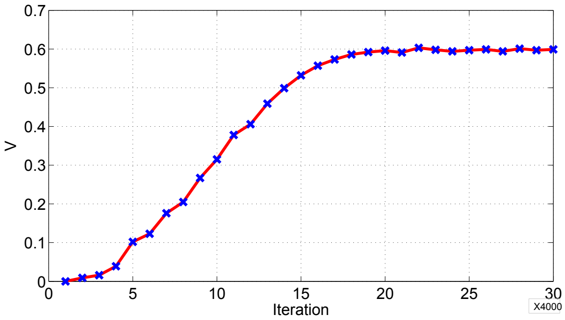

Fig. 5. The curve of parameter v (accuracy on validation).

4. EXPERIMENTS AND RESULTS

In this section, all of the experiments are based on our UWF database and tested on 6000 given face pairs in LFW. In Sec. 4.1, we will investigate the appropriate value of mixture parameter β and show the curve of parameter v described in Sec. 3.2. Then we will investigate the effects of our modification signal on identification-verification in Sec. 4.2 and demonstrate our final system in Sec. 4.3. For comparison, in Sec. 4.4 and Set. 4.5, we implement some state-of-the-art models on UWF database and show their performances.

4.1. Optimal mixture parameter

As demonstrated in Sec. 3.2, the formula of loss function will change with accuracy on validation accordingly. So the fact that wether it could reach a steady level affects our system. As shown in Fig. 5, the value of v finally reached a steady level which indicated that the prediction capability of our system is credible. Then we investigate the coefficient of intra-class and inter-class distances on feature learning, by varying β from 0 to 1. Fig. 4 shows the results of our model on the test set (removing the misclassifications) of UWF, trained with examples of 4096 identities randomly selected from UWF. At $\beta = 0.001$, the performance is the best. As β increasing, the intra-class distance gradually dominates the learning phase with negative effects and the performance turns to be worse, because these two feature vectors are gradually pulled together which should be regarded as being extracted from different classes and kept away from each other due to the predicted label. And when β comes to 1, it will directly reduce the distance between these two feature vectors without considering the potential positive classification and the accuracy dropped to 88.9% which we don't show in Fig. 4. At the other extreme of $\beta = 0$, only inter-class distance remains, the performance is also surprising. We think it is because our model's prediction is already precise with a few misclassifications. Finally we set the mixture parameter β to 0.001 throughout our experiments, with taking into account the effect of reducing intra-personal variations and increasing inter-personal variations simultaneously.

4.2. Investigating the effects of modification signal

Table 2. Comparison of different bootstrap strategies. (i,v) separately means identification and verification signal, and +m/-m means guided by modification signal or not. Training ID number is 14316.

	i(+m), v(+m) **#1**	i(+m), v(-m) **#2**	i(-m), v(+m) **#3**	i(-m), v(-m) **#4**
ACC(%)	97.33	93.39	96.88	92.45

To investigate the effects of modification signal separate on identification and verification, we do different bootstrap experiments. And the face verification accuracies of the learned features on the test set (removing the misclassifications) of UWF, measured by Cosine distance, are listed in Table. 2. Comparing **#1** with **#3** and **#2** with **#4**, the features learned with i(-m) signal are only slightly worse than those learned with i(+m) signal. This may be due to that ConvNet has inherent anti-noise ability, i.e. the feature learning can not be easily harmed by misclassifications to some extend. In contrast, comparing **#1** with **#2** and **#3** with **#4**, v(+m) signal helps a lot in feature learning. And the features from **#3** are more effective than that from **#2**. This is a strong evidence that modification signal mainly affects verification signal since in verification stage correctly tackling with the intra-class and inter-class variations by the potential class label can significantly improve the discrimination of our network and as a result, face verification accuracy can be largely improved.

4.3. Multi-region merging and results

To make our system more simple and easy to learn when merging features, we straightly take three pathes from the fixed-size 250 × 250 face images. An illustration of our method to merge features extracted from these three pathes is shown in Fig. 6. At test time, the same operation of cropping described in Fig. 6 is done to the 6000 given face pairs in LFW. For face verification, three 3000-dimension feature vectors, which are extracted from the same individuals with different scales by a total of 3 SL-DCNN models, are directly concatenated to a 9000-dimension feature vector without PCA reduction. Then we use cosine distance to measure the similarity between the given pair. And we have achieved 97.47% verification accuracy on LFW shown in Table. 3.

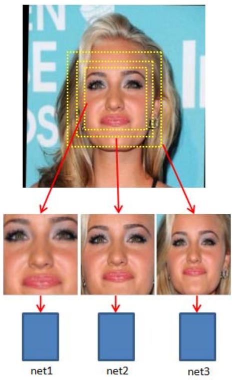

Fig. 6. The procedure of multi-region merging. They are all cropped from the same image and centered on nose tip. They vary in scale, separately defined as small(left), middle(mid) and large(right). The left and right pathes are resized to 104 × 96. Accordingly, we will get 3 SL-DCNN models trained with these pathes respectively.

We have also tried many other methods, e.g. we cropped many different scale patches (centralized at eye-center, nose tip and mouth corner) with their horizontal reflections, then feature vectors extracted from these patches are concatenated to a representative vector followed by PCA reduction. However, without impressive improvements, these methods need more time to train their models and more operations of dimensional reduction in verification phase.

Table 3. Face verification accuracy on LFW with features extracted from three pathes. And time used for features extraction is shown bellow

selected patches	accuracy(%)	time(ms)
small scale	95.88	4.6
middle scale	96.97	4.6
large scale	96.95	4.6
merging	**97.47**	13.8

4.4. Rich identity improves feature learning?

As shown in Table. 4, the performance of DeepID2 [10] is very terrible, with just 88.87% verification accuracy achieved. We think that it may be caused by the questionable process of verification, in which the distance between two feature vectors may be wrongly handled due to the weak label. And with the number of identities increasing, the performance of DeepID2 is getting worse since the more instances, the higher probability of negative learning. However, trained by our novel modification signal, the performances of our models are better and better with the ID number increasing. We have significantly improve the accuracy rate by nearly 9%. And the conclusion is that rich identity not always improves feature learning under the environment of strong noise unless misclassifications are correctly modified.

Table 4. Verification accuracy on LFW with identity number increasing.

Acc(%) ID number models	14316	8192	4096	time(ID=14316)
DeepID2 [10]	87.9	88.23	88.87	8(day)
DeepID2(-verif)	91.65	90.32	89	6(day)
DeepID2(+modif)	*94.67*	*93.04*	*91.43*	8(day)
Naive Deep [13]	93.97	92.88	92.02	10(day)
ours(6 layers)	**95.1**	93.95	92.87	2(day)
ours(9 layers)	**96.32**	95.02	93.93	2.4(day)
ours(11 layers -modif)	*93.06*	-	-	3(day)
ours(11 layers)	**96.97**	95.78	94.67	3(day)
ours(merging)	**97.47**	96.23	95.4	9(day)

4.5. Comparison with others and results analysis

From the results shown in Table. 4, without training with verification signal, the results of DeepID2 (-verif) and Naive-deep [13] are a little higher than DeepID2 and increase with the number of identity. It demonstrates that, ConvNet has inherent anti-noise ability and robustness under big noise environment. Unfortunately, if it was learned with verification signal which is not bootstrapped by modification signal may harm its performance since a pair of images with the same weak label may belong to different classes and the closer we keep these two feature vectors extracted from the same identity the worse we do, e.g. DeepID2 VS DeepID2 (-veri). Similarly, the 93.06% verification accuracy achieved by our model which is not bootstrapped by modification signal is also unsatisfying. For comparison, all of our models and DeepID2 (+modifi) which are bootstrapped by modification signal perform better than that without modification. And in another aspect, with correctly handling the misclassifications by modification signal, the verification accuracy increases with the identity number. Finally, by deepening the network and adopting effective procedure of multi-region merging, our model (merging) which is trained with three weakly-supervised signals has shown its significant 97.47% verification accuracy against with others and the ROC comparison is shown in Fig. 7.

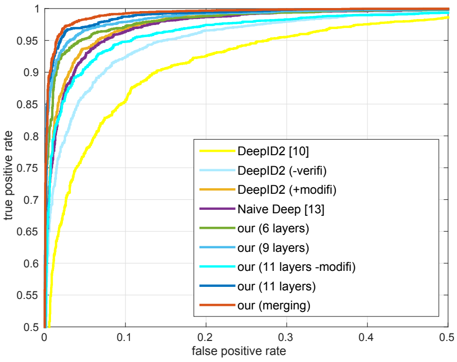

Fig. 7. ROC of face verification accuracy on LFW. Best viewed in color. The ID number is 14316.

5. CONCLUSIONS

In this paper, we have collected a challenging unlabeled database UWF. And we have proposed an effective method, in which modification signal is used to get the potential true labels, to improve the feature learning for face recognition under the environment of unlabeled examples. Matched pairs and mismatched pairs will be correctly distinguished by our model trained with three weakly-supervised signals. Using a simple and efficient procedure of multi-region merging, we finally achieved our SL-DCNN system with 97.47% face verification accuracy on LFW. And we also hope our work will inspire further study on face recognition under strong noise environment or self learning.

6. ACKNOWLEDGEMENT

This work was partially sponsored by supported by the NSFC (National Natural Science Foundation of China) under Grant No. 61375031, No. 61573068, No. 61471048, and No.61273217, the Fundamental Research Funds for the Central Universities under Grant No. 2014ZD03-01, This work was also supported by the Beijing Higher Education Young Elite Teacher Program, Beijing Nova Program, CCF-Tencent Open Research Fund, and the Program for New Century Excellent Talents in University.

7. REFERENCES

[1] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” *Workshop on Faces in ’Real-Life’ Images: Detection, Alignment, and Recognition*, 2007.

[2] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun, “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3025–3032.

[3] Xudong Cao, David Wipf, Fang Wen, Genquan Duan, and Jian Sun, “A practical transfer learning algorithm for face verification,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3208–3215.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] Gary B Huang, Honglak Lee, and Erik Learned-Miller, “Learning hierarchical representations for face verification with convolutional deep belief networks,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2518–2525.

[6] Xinyuan Cai, Chunheng Wang, Baihua Xiao, Xue Chen, and Ji Zhou, “Deep nonlinear metric learning with independent s ubspace analysis for face verification,” in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 749–752.

[7] Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Hybrid deep learning for face verification,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1489–1496.

[8] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lars Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708.

[9] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Recover canonical-view faces in the wild with deep neural networks,” *arXiv preprint arXiv:1404.3543*, 2014.

[10] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, “Deep learning face representation by joint identification-verification,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[11] Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Deeply learned face representations are sparse, selective, and robust,” *Eprint Arxiv*, 2014.

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, A Vedaldi, K Lenc, M Jaderberg, K Simonyan, A Vedaldi, A Zisserman, K Lenc, et al., “Deep face recognition,” *Proceedings of the British Machine Vision*.

[13] Erjin Zhou, Zhimin Cao, and Qi Yin, “Naive-deep face recognition: Touching the limit of lfw benchmark or not?,” *arXiv preprint arXiv:1501.04690*, 2015.

[14] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich, “Training deep neural networks on noisy labels with bootstrapping,” *arXiv preprint arXiv:1412.6596*, 2014.

[15] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Eprint Arxiv*, 2014.

[16] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0580, 2012.

[17] David Flatow and Daniel Penner, “On the robustness of convnets to training on noisy labels,” .

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.