Review of Machine Learning Methods for Human Face Recognition in Images

Ramya Rao Basava
Department of Computer Science
University of British Columbia

Sang-Wha Sien
Department of Computer Science
University of British Columbia

Abstract

Facial recognition from images is a very interesting area of research and has useful applications in real-world scenarios. This paper presents a literature review of some of the most recognized works for facial recognition using machine learning techniques, starting from 1900s. These methods are summarized, analyzed, and discussed in detail. We give general research trends over the years and break-through contributions. In addition, we discuss some of the limitations in the current works and suggest future research directions and areas with scope for improvement.

1 Introduction

Object recognition is a term for computer technologies that can recognize certain people, animals, or other objects in images. One of the most popular ways to carry out this task robustly is through the use of machine learning algorithms. In essence, object recognition in images is primarily connected to computer vision, and its overarching goal is to produce algorithms that match or go beyond the effectiveness of how humans process and classify image data. In particular, we explore the field of facial recognition where the object to be recognized in the image is a human face. This is a widely researched area due to its importance in a large number of applications. Some of them include facial recognition software in smartphones (for example tagging faces in photos), real time video surveillance, bio-metrics, and security.

Over the years, facial recognition methods have gone through periods of trends that have made substantial advances with increasingly more effective algorithms. In earlier years, techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) were very popular due to their ease of implementation and speed for real-time recognition. Many researchers have proposed variations and extensions of these methods mainly to resolve several problems in image conditions that can affect accuracy. These conditions typically involve different aspects like lighting, viewpoint, facial expression, and occlusion. The proposed methods include local feature based methods, filtering based methods, and many more which have contributed to advancing knowledge in this area. Of late, deep learning has made the most impact with many variations of it winning competitions with significant increases in accuracy. These methods include convolutional neural networks (CNNs) and deep neural networks (DNNs).

In this paper, we provide a literature review to summarize some of the most popular machine learning algorithms in facial recognition, starting from early 1990s. Our contribution provides a glimpse into how these works on face recognition are related to each other. It also gives a general trend on how the different methods have gained popularity over the years due to advancements in research. We conclude with estimates of possible directions for future work. The remainder of the paper is organized as follows. Section 2 gives a literature review of our selections of the most popular facial recognition methods, covering algorithmic details and chronological progression. In Section

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

3, we provide a summary of the general trend over the years, our observations, and possible future directions for facial recognition research.

2 Literature Review

In the early 1990s, one of the most impactful facial recognition methods was 'eigenfaces' proposed by Turk and Pentland [20]. In this method, relevant information for a face image is gathered from principal components (eigenvectors) of the covariance matrix of a distribution of facial images (the training set). The first few eigenvectors corresponding to the largest eigenvalues are chosen to form a mutually orthogonal set of lower dimensional basis. These basis eigenvectors are termed as 'eigenfaces', and they capture the differences in variation in facial images. It is also commonly referred to as PCA in facial recognition. The training face images can be linearly transformed onto this lower dimensional basis. When a new face is sent as input, it classifies whether or not there is a face in the image by projecting the input image onto the eigenfaces and checking if it is close enough to the 'face space'. Once the image is classified as a face, it checks whether it is a known face or not. The method can also be tweaked to incorporate new unknown faces. This method gives excellent performance in general but fails when the lighting conditions or facial expressions are changed in the images, as these variations are not present in the training data.

One of the popular follow-up works to eigenfaces was 'fisherfaces' proposed by Hespanha et al. [8] and addresses the issue that is prevalent in PCA. It is also popularly known as LDA for face recognition. It is observed that in the eigenfaces method, scatter is maximized not only due to the between-class scatter but also due to within-class scatter. Between-class scatter is useful for classification while within-class scatter is deemed unwanted information. Hence unwanted effects like change in lighting conditions are captured as significant eigenfaces. In Fisher's method, the projection is based on Fisher's linear discriminant (FLD), where the optimum basis is obtained by maximizing the ratio of the between-class scatter and within-class scatter. This leads to greater between-class scatter and linearly separable classes in the projected space, as opposed to eigenfaces where the classes in the projected space are smeared. It is shown that this method gives considerable improvement in results for changes in lighting or facial expressions and computational time with respect to eigenfaces.

Both PCA and LDA are linear projection methods and perform well on linear manifolds. They can effectively see the global Euclidean structure but fail to recognize face images that lie on a non-linear manifold. Taking this into consideration, researchers had started to look into the manifold structure for face detection by the mid 2000s. One of the first works on manifolds that preserve the local structure of the image space was 'Laplacianfaces' proposed by He et al. [7]. In this method, locality preserving projections (LPPs) are used to obtain a mapping of each face image in the image space into a low-dimensional face subspace. In LPP, the face's manifold structure is modeled by a nearest-neighbor graph to preserve its local structure. The set of images in this subspace is called Laplacianfaces which was found to have more discriminating power than the subspace images produced by PCA and LDA. Researchers showed that Laplacianfaces can identify a person with different expressions, poses, and lighting conditions. Laplacianfaces yielded a significant drop in error rates in most cases when compared with PCA and LDA. Also, it was observed that LDA was more sensitive to different training data sets.

To further deal with feature selection as a way to reduce the dimensionality of the problem and also to deal with occlusion in images, Wright et al. [23] proposed a sparse representation based on L1-minimization. In this method, the test examples are represented as a linear combination of the training examples. The training matrix consists of n training examples with k object classes. Ideally if the non-zero elements of the weights are associated with a single class, the test example can be assigned to that class. The resulting set of equations is solved as a L1-minimization problem which gives a sparse solution for careful feature selection. It is also extended to be robust to corruption, occlusion, and disguise in images by adding an additional error vector which deals with the occlusion in the input image. The study is confined to human frontal face recognition and does not deal well with variation in pose or viewpoint. It is robust only to small variations due to changes in pose or displacement.

Other notable works on face detection during the phase of 2000-2010s include robust real time face detection method by [21], which is popularly known as the Viola-Jones object detection framework,

and multi-class support vector machines (SVMs) as proposed in [5]. SVMs were very effective methods in pattern recognition. They presented a better learning algorithm compared to algorithms like eigenfaces.

Approximately towards the end of 2010, there were substantial methods in facial recognition that addressed issues like lighting, viewpoint, facial expressions, and different manifolds. Also, additional local enhancement methods were researched to improve facial detection performance. These methods included using filters, getting the local textural properties, and other popular methods in the field of pattern recognition. For example, Gabor feature based enhanced LDA was proposed in [12] which yields features from face images that display scale, locality, and orientation selectivity. A local binary pattern texture feature method was proposed in [1], which enhanced errors due to pose and illumination changes. Although there was considerable work going on in improving facial recognition methods, there was no comprehensive technique which would be invariant to all the real-world changes. The methods were very specific to the issue they were addressing, one aspect at a time. Due to this, they were sometimes unstable and would fail significantly if they encountered images in real-world that were not used in the training phase. The coming decade (to date) was dominated by methods based on neural networks and deep learning due to their unprecedented performance in pattern recognition and suitability in computer vision.

By the 2010s, facial recognition researchers started to find significant results with neural networks and deep learning techniques. One such method, proposed initially by Huang et al. [9] is extreme learning machine (ELM). ELM is a feedforward neural network for classification and regression with a single layer of hidden nodes. Because the parameters of hidden nodes need not be iteratively tuned, ELMs are able to produce good generalization performance and learn thousands of times faster than networks trained using backpropagation. They can also outperform SVMs, as SVMs provide suboptimal solutions in both classification and regression applications. Additional work on ELMs by Zong and Huang [24] analyzes ELM's efficacy on classifying facial images. Noting ELM's similarity as a binary classifier to SVM, the researchers devised an experiment comparing ELM and SVM under two multiclass conditions: one against all (OAA) and one against one (OAO). Using a dimensionality-reducing technique called discriminative locality alignment (DLM), the researchers found that SVM and ELM showed comparable results in accuracy and training time. Yet they found that ELM was not sensitive to parameters, unlike SVM. Therefore, the number of hidden nodes directly affects performance, with higher number of nodes favoring ELM over SVM in terms of training time.

In 2012, Krizhevsky et al. [11] trained the deepest convolutional neural network (CNN) to date to classify images from ImageNet large scale visual recognition challenge (ILSVRC) in 2010 and 2012. ImageNet is a large-scale hierarchical image database detailed in [4]. The rationale for using a deep CNN is that images in realistic settings show considerable variability so in order to recognize them, it is necessary to use much larger training sets such as the ImageNet dataset. But to accommodate a large dataset, a model with large learning capacity like CNNs is needed. CNN's learning capacity can be controlled by varying its depth and breadth. Unlike standard feedforward neural networks, CNNs have few connections and parameters which means they are easier to train. The architecture of the proposed model consists of five convolutional and three fully-connected layers optimized to train on two GPUs. They modeled the neuron's output with ReLU to speed up performance and used a local response normalization after applying ReLU to further aid generalization. They then used a 1000-way softmax for final classification. To ensure the deep model did not overfit, they leveraged data augmentation, dropout, and overlapping max pooling. The detailed architecture is illustrated in Fig. 1. This model beat the best performance that was achieved in ILSVRC-2010 and won top place in ILSVRC-2012. They concluded that the depth of the CNN was important, as the performance was found to suffer if a layer was removed. This paper was a major contribution at that time, as it changed the direction of computer vision research. It conveyed that the performance of CNNs for computer vision problems was unmatched especially when training with large datasets. Subsequent works on this topic started to gain momentum soon after. For example, one study extended CNNs to deeper depths for improved image classification [16], and another used an optimal local sparse structure for CNNs [18] (Inception model, used for GoogleNet - a 22 layer deep network submitted in ILSVRC-2014). Research was also extended to application of CNNs to facial image recognition.

In 2013, Sermanet et al. [15] furthered Krizhevsky's work by presenting an integrated CNN framework to simultaneously classify, localize, and detect facial images on the ImageNet database. This framework modeled Krizhevsky's CNN with several modifications including no overlapping pooling

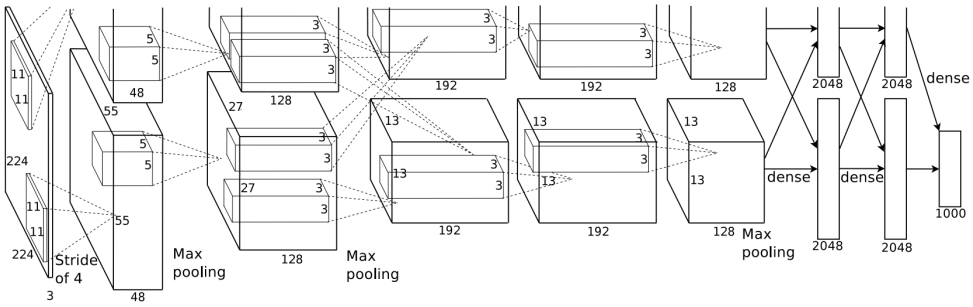


Figure 1: Deep CNN architecture by Krizhevsky et al. [11] shows the detailed model spread over two GPUs. It consists of five convolutional and three fully-connected layers and a 1000-way softmax classification at the end.

and no normalization. Furthermore, a multi-scale classification scheme, which they call a sliding window approach, more densely explores each image and at different scales. This results in more views to vote on, for a better performance and increased robustness. It consists of feature extraction and classifier layers which are repeatedly used for an image at each scale and its horizontally-flipped version. The feature extraction part of the method was developed into a program called 'Overfeat'. At scale of 6, the classification of the ILSVRC-2012 validation error rates bested Krizhevsky's 1-CNN top-1 and top-5 errors as well as ranked 5th in the 2013 classification results. For localization, classifier layers are replaced with a regression network which is then trained with L2 loss at multiple scales to predict bounding boxes. Merging the bounding boxes reinforces true positives while devaluing false positives. Using 4 scales, the model won the ImageNet 2013 competition with a top-5 error rate of 29.9%. Detection training uses the same method as localization but is fine-tuned to deal with negative objects. This, too, won first place with a mean average precision of 24.3%. This work was the first to describe how CNNs can be used for localization and detection for ImageNet data.

In 2014, two main papers independently described procedures that closed the human level performance gap in facial recognition [17; 19]. Until that point of time, no other method came close to the 97.53% accuracy rate achieved by human performance. Sun et al. [17] used a deep CNN to achieve 97.45% accuracy, while Taigman et al. [19] used a DNN to achieve 97.35% accuracy rate. Other methods like 'PCANet' [3], based on learning-based filters in CNNs, enjoyed limited popularity which was overtaken mainly by the above two methods described in detail below.

Sun et al. [17] devised a procedure they called 'DeepID' (Deep hidden IDentity features), a deep CNN where facial images are classified to form progressively higher level features with deeper layers. What this also means is that the number of neurons per layer gets more reduced as the lower level features start forming higher level features. The output from the final DeepID layer is used to predict thousands of identity classes simultaneously. This is opposed to other studies that use binary classifiers like SVM. The small number of nodes in the DeepID layer forces the model to learn shared features of the faces of different classes which helps with classification. The architecture consists of four convolutional layers with max pooling and ReLU nonlinearity that hierarchically form features. A final softmax layer outputs the identity classes. A face verification neural network model based on joint Bayesian (JB) technique was also developed. These two models were trained on two datasets, CelebFaces and the CelebFaces+ [13], and tested on the labeled faces in the wild (LFW) [10] dataset. They showed that when using the JB + DeepID verification method, combining more image features and adding more identity classes significantly increased accuracy. The DeepID model with transfer learning JB [2] achieved a test accuracy of 97.45% on the LFW dataset when trained on 100 patches and 10,477 classes on the CelebFaces+ dataset.

The second best reported result on the LFW dataset with 97.35% was the 'DeepFace' method, developed by Taigman et al. [19]. This differed from the DeepID approach in that images go through more preprocessing: a 3D alignment step that learns from raw pixel RBG values to produce a very compact yet sparse descriptor. This is opposed to other systems that use tens of thousands of image descriptors with features that are often combined to improve performance. Preprocessing was deemed

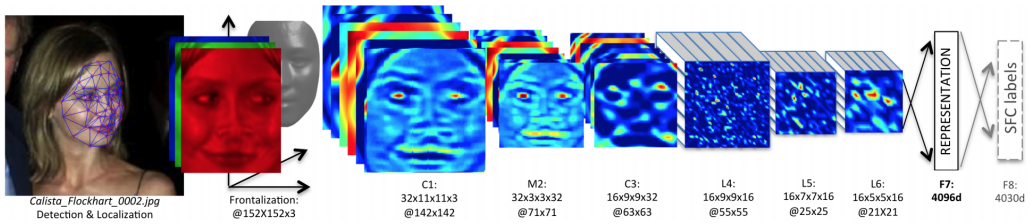


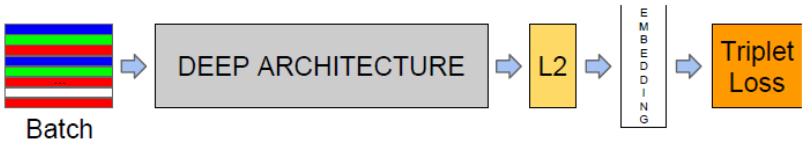
Figure 2: DeepFace CNN architecture taken from Taigman et al. [19]: a 3D-aligned facial image is filtered through a convolution-pooling-convolution section, then fed to three locally-connected layers and 2 fully-connected layers before softmax classification on the SFC dataset.

to be needed because an unconstrained scenario is considered a difficult problem due to variances in pose and expression. This 3D representation is fed to three layers in the model to extract low-level features (simple edges, texture): a convolutional layer, a max pool layer, and another convolutional layer. The next three layers are locally connected so that the model can learn a different set of filters at every location. The next two layers are fully connected to capture correlations between more distant features (for example, positions of eyes or mouth). The output is then fed to a K-way softmax for classification. ReLU and dropout are used to produce highly sparse and nonlinear features in the model, and normalization is used to reduce sensitivity to changes such as illumination. The overall architecture is illustrated in Fig. 2. The model was evaluated by learning from the social face classification (SFC) dataset from Facebook and tested on the LFW database and the YouTube faces (YTF) dataset [22]. An ensemble DNN achieved a 97.35% accuracy, closely approaching human performance of 97.53%. Testing on YTF was to show that the algorithm can generalize to a different dataset. As this dataset is of lower image quality, accuracy was shown to be somewhat lower at 91.4%. However, this reduced the previous best error by 50%.

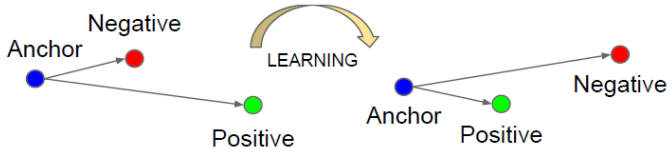
The problem with DNNs and deep CNNs at the time, however, was that they were difficult to train. Very deep networks in general had started to produce great accuracy results, but they became more difficult to train as models increased in depth. To resolve this, He et al. [6] proposed a new residual learning framework (ResNets) to ease the difficulty when training substantially deeper networks. The specific problem they hoped to resolve was the degradation problem with deeper models: accuracy gets saturated and then degrades rapidly. The contradiction with this was that the degradation was not due to overfitting, as deeper models were shown to produce higher training errors. Their proposed algorithm leveraged the fact that training the weights of multiple nonlinear layers to approximate identity mappings can drive down the cost of training, even when adding more layers. They did this by skipping certain layers in the model, and these residual mapping shortcut connections were shown to be easier to learn and optimize. Experimental results showed that in plain networks, those that do not have shortcuts, the degradation problem was observed because validation error increased as the number of layers increased; however in the residual networks, the validation error decreased with depth. In fact, even with very deep layers (1000+), ResNets showed no optimization difficulty and were able to achieve low training errors. However, overfitting still remained a concern. An ensemble of six ResNets of different depths resulted in a 3.57% top-5 error to win 1st place in ILSVRC-2015.

Later in 2015, the best known method in terms of performance was published by Google. This was proposed by Schroff et al. [14] and was called 'FaceNet'. It achieved a record accuracy of 99.63% on the LFW dataset and 95.12% on the YTF dataset. The network consists of a batch input layer and a deep CNN followed by L2 normalization, which results in the face embedding layer. This is followed by the triplet loss during training. The full network is illustrated in Fig. 3a. A Euclidean distance based embedding space shows the similarity between face images: faces of the same person have smaller distances and faces of distinct people have larger distances. The triplet loss tries to enforce a margin between each pair of faces from one person to all other faces, as shown in Fig. 3b. Proper triplet selection is crucial for fast convergence. There is no more a 3D alignment step as in the case for the DeepFace method. The method only required minimal alignment to give excellent results.

It is to be noted that this paper provides a review of a limited number of papers which contributed greatly to the advancement of facial recognition in the field of computer vision. There is an exhaustive



(a)



(b)

Figure 3: FaceNet [14]: Figure 3a shows the CNN framework for the FaceNet method. Figure 3b shows how the triplet loss minimizes the anchor (particular example) and a positive example, both of which have similar identity. The triplet loss also maximizes the distance between the anchor and a negative example, having a different identity.

list of research papers published in this field, that have also added value and helped the advancement of facial recognition methods over the years.

3 Discussion

Our literature review of facial recognition methods shows the general trend in algorithms since the 1990s and gives an analysis on the strengths and weaknesses of the methods. Over time, these methods have shown different capabilities and advances that addressed many facial recognition tasks to a high level of accuracy. Some have even surpassed human performance. The advent of CNNs and DNNs made learning huge datasets possible due to their layered hierarchical architectures which could be molded as required. These deep learning methods turned out to be extremely useful tools in facial recognition and image classification in general, giving robust results almost every time.

Some of the next possible steps in this research area would be to reduce the computational cost of training large-scale real-world data. Even though neural networks give outstanding results, it is a known fact that the time required to train these networks is extremely high. This is because computing time increases as the training data increases and also as the depth of the network increases. Datasets will keep increasing in size in the coming years, and it is an unstoppable process. A promising direction is to leverage parallel computing or parallel GPU's to speed up the computation. Not much or very few discussions about this were found in the papers, though Krizhevsky et al. [11] used it in their study.

Although current methods show considerable effectiveness, many issues still remain that will no doubt create problems for facial recognition research. Some of these are listed below:

- Effects in the image like illumination, facial expressions, scaling, viewpoint, pose, occlusions, and background have not been addressed comprehensively together.
- Different methods use different strategies in pre-processing, processing and post-processing. This leads to a very specific sequence of steps catered to that method and can hurt generalizability. For example, the different layers in CNNs or DNNs can be modified and many local differences exist between methods.
- Most methods are based on 2D face characteristics from images for facial recognition which might not capture the full 3D features effectively.

Keeping the above points in mind and also thinking about future directions, some areas for further research and opportunities to explore are listed below:

- The high accuracies being obtained by the present methods are limited to the dataset being used during training and validation. There is a need to consider more complex, real-world data that show more variance in race, color, age, and gender.
- Many of the studies above have yielded satisfactory performance under controlled scenarios. More research is needed with real-world unconstrained images that show varied and unusual but possible scenarios. That is, the methods should be more robust and fail-proof to these special test cases.
- It is a known fact that if the training data somehow does not have a test case encountered in real-world, the method fails. This failure is very likely in real-world scenarios and might cause very bad results and losses. Ways to deal with these cases have to be researched.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (12):2037–2041, Dec 2006. ISSN 1939-3539.

[2] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun. A practical transfer learning algorithm for face verification. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[3] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. PCANet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 24(12):5017–5032, Dec 2015. ISSN 1941-0042.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] Guodong Guo, S. Z. Li, and Kapluk Chan. Face recognition by support vector machines. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 196–201, March 2000.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:328–340, 2005.

[8] J. P. Hespanha, D. J. Kriegman, and P. N. Belhumeur. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(07):711–720, Jul 1997. ISSN 1939-3539.

[9] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1):489 – 501, 2006. ISSN 0925-2312. Neural Networks.

[10] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, Oct. 2008.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[12] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. *Trans. Img. Proc.*, 11(4):467–476, Apr. 2002. ISSN 1057-7149.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[15] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[17] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[20] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1): 71–86, 1991.

[21] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57, 2004. ISSN 1573-1405.

[22] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, June 2011.

[23] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2): 210–227, Feb 2009. ISSN 1939-3539.

[24] W. Zong and G.-B. Huang. Face recognition based on extreme learning machine. *Neurocomputing*, 74(16):2541 – 2551, 2011. ISSN 0925-2312.