Face Detection and Recognition With Deep Convolutional Neural Network: A Survey

Rui Zhang (u5963436)
Research School of Computer Science
The Australia National University
Canberra, ACT 2601
Email: u5963436@anu.edu.au

***Abstract*— Face detection and recognition are widely used to analyze people in the images and videos and current techniques are much more accurate but their complexity is also higher so that they are hard to run in real time. This article firstly summaries the early approaches for face detection and recognition and then introduce some remarkable work using complex CNN architectures and achieving good performance.**

I. INTRODUCTION

Face detection, alignment, verification and identification have been studied actively for several decades and lots of research outcomes have already been applied on smartphones, cameras and social media platforms such as Twitter to improve user experience. This series of tasks are challenging as various background, lighting conditions, image definition, occlusion and non-frontal views in the images will increase noise in features of human faces and make it harder to detect faces and recognize faces.

Before summarizing related techniques, it is of necessity to introduce several terms: (1) Face detection: identifying regions containing human faces in digital images; (2) Face alignment: determining the shape and location of face components such as eyes and nose based on the information of faces such as shape and size; (3) Face identification: classifying a face into a large number of identity class, which can be regarded as multi-class classification; (4) Face verification refers to determining whether a pair of faces belongs to the same identity class or not, i.e., binary classification. Face identification and verification are generally called face recognition.

A. Early and Recent Working on Face Detection

Early face detection methods can be separated into three classes: appearance-based, template-based and feature-based methods [1]. Appearance-based approaches select plenty of small patches as face candidates and feed these patches into discriminative models to determine true faces[2], such as Viola-Jones algorithm[3]. Template-based approaches are not usually used as face detectors due to their low detecting speed while having an advantage of handling pose and expression variability[2] such as Active Appearance Models [4]. This kind of models usually are applied to locating key feature points for face recognition and tracking [2]. Feature-based approaches determine locations of faces according to verify whether key features of an image, such as locations of eyes, nose and mouth candidates, are in a plausible geometrical arrangement [1], such as [5]. These methods are effective to detect up-right faces but they fail to deal with extreme lighting, partial occluded faces or side faces.

The weakness of early work motivate lots of current approaches to solve these problems, which can be roughly classified into three categories [6]: (1) cascaded-based methods, such as [7], [8]. These methods are designed based on cascading classifiers proposed in the Viola-Jones algorithm[3] and achieve better performance in addressing occlusion and multi-view face detection. However, they have higher complexity as the variants of Viola-Jones algorithm and lose the advantage of real-time calculation. (2) DPM-based methods, such as [9],

[7]. The core idea behind these methods is from [10], i.e., to view a face as a face component bag and train detectors for face components and a latent SVM classifier [10] to determine whether a group of features corresponds to a face. DPM-based methods show robustness to occlusion but suffer intensive computation because for each candidate location of the image, a latent SVM is solved and to achieve state-of-the-art performance, multiple DPM detectors are combined. (3) Neural-Network-based methods. Early Neural Networks are used as a common classifier like SVM and Naive Bayesian Classifier [11], [12] to detect faces based on features. Recent Neural Network models are designed to extract face features automatically and obtain state-of-the-art performance[13], [14], [15]. Although the strong computational ability of modern computers accelerates the computations in the Neural Network, extremely high complexity of Neural-Network architectures require not only high-performance equipment but also extensive training data to adjust their parameters. Thus, the state-of-the-art Neural-Network models are hard to implement in the cameras or mobile phones.

B. Early and Recent Working on Face Recognition

In the field of face recognition, frontal faces are usually given and there is no need to deal with side faces. Early approaches to face recognition rely on extracting features of face images by such as Principal Component Analysis, called eigenfaces [16] and jointly modeling shape and appearance variations while discounting pose variations using Active Appearance Models [4]. The manually designed features requires good lighting conditions and well-aligned faces. As a result, side faces and occlusions cannot be solved effectively by these methods. The recent methods are mainly to design deep Convolutional Neural Networks (CNN) architectures to extract face features[17], [18], [19], [20]. The performance of deep CNN architectures are much better than early methods and even than human-level performance[19], [20]. However, it is still the complexity that blocks their application in wide fields.

In the following two sections, I am going to introduce the latest Deep-Neural-Network-based techniques for face detection and recognition. Their contributions, pros and cons will be summarized. After the introduction, a brief conclusion will be given and papers on advanced methods not covered in this articles will provided.

II. Advanced Face Detection

Face detection is a branch of object detection and recent work on face detection is mainly based on several remarkable foundation work in the field of the object detection.

A. Selective Search

Selective search[21] is a novel method to generate candidate regions for object detection, before which the exhaustive search is widely used, i.e., iteratively scanning images by a window and varying the size of the window. The later method is time-consuming and tends to bring about numerous trivial regions. Selective search starts with dividing an image into 1k to 2k small regions, after which it follows several grouping rules to get the union of two neighbouring regions so that a larger region replaces previous two regions [21]. This runs until all regions become a single region and returns all regions occurring during the whole process as candidates. This method has an advantage of efficiency because of which is was used in [13] to generate candidates. However, this method is much harder to implement than the exhaustive search.

B. R-CNN

The first try of CNN in the field of object detection is in[13], where a complex but effective method called Regional Convolutional Neural Networks (R-CNN) is proposed. R-CNN generates candidate regions by the selective search and extracts features of candidate regions with CNN. To category candidate regions, RCNN exploits a Support Vector Machine (SVM) classifier to classify the candidate regions based on their features. Furthermore, a linear regression model is used to refine the positions of detected candidate regions (bounding box regression) and further extended in [14], [15], [6], [22]. This method significantly improves the object detection accuracy but its detection speed is pretty low (13s/image on a GPU and 53s/image on a CPU) and the disk storage is also needed to save intermediate face features[13].

C. Fast RCNN

To address the shortcomings of R-CNN, [14] proposes Fast RCNN, which not only gets higher detection quality but also needs no disk space for feature caching and less time for face detection (9 times faster than R-CNN). Fast RCNN modifies the four-stage training in R-CNN to a two-stage training by combining CNN architecture, the SVM classifier and the bounding-box regression into a larger and more complex Neural Network architecture with two output layers to return the classification and bounding box regression results respectively. One contribution of Fast RCNN is applying truncated SVD to accelerate Neural Network[14]. The novel method achieves the better detection accuracy and higher speed in the object detection, but the selective search is time-consuming which makes it impossible to run the Fast RCNN in real time.

D. Faster RCNN

Faster RCNN is proposed in [15] to accelerate generation of candidate regions, which is realized by combining all stages of the object detection in a multi-layer CNN architecture. Faster RCNN firstly extracts features of the whole image in the first several layers, at the top of which a Regional Proposal Network shares the convolutional layer and further extract regional information based on previously-extracted features. After that, the Fast RCNN architecture is applied for classification and bounding box regression. This method realizes object detection in 0.2s per image and achieve the state-of-the-art detection accuracy of 73.2% on the dataset of VOC2007[23]. However, Faster RCNN combines two different tasks (classification and regression) into a network which reduces the detection accuracy and speed, and the later work YOLO's [24], [25] only use the CNN architecture for regression and achieve better performance in the object detection.

E. Multi-view Face Detector

The multi-view face detector [6] is inspired by the ideas behind R-CNN [13], which uses fine-tuned AlexNet[26] to extract regional features and trains the other deep architecture similar to AlexNet to classify candidate regions. Compared with the CNN architectures in the field of the object detection, this ideas behind the multi-view face detector are simpler and the structures are deeper. To improve the detection accuracy, [6] use a very large training set to train their models, as a result of which they got the best detection accuracy when the multi-view face detector was proposed. However, the two-stage process makes this method kind of complex and lacks the efficiency of one-stage Faster RCNN.

F. MTCNN

Some of recent work focuses on combining face detection and alignment and one of them is the Multi-task Cascaded CNN detector (MTCNN) [22] which outperforms the state-of-the-art methods on FDDB[27] and WIDER FACE [28] benchmarks for face detection and AFLW [29] benchmark for face alignment. MTCNN consists of three stages each of which exploits a CNN architecture. From the first to the third stage, the three independent CNN architectures are designed to generate candidate face regions, reject non-face regions and locate facial key points respectively. Three important techniques, thresholding, non-maximum suppression and bounding box regression are used on results returned by CNN architectures to reduce the trivial candidates and refine the positions of bounding boxes. As a result, this method is fairly complex and plenty of hyper-parameters are brought by thresholding and non-maximum suppression.

III. Advanced Face Recognition

Face recognition does not need so many tricks as those in face detection, e.g. techniques to generate candidate regions and then bounding boxes regression. Instead, given frontal faces, face recognition techniques determine matching pairs among them so its speed is usually much faster than face detection techniques. However, the recognition accuracy usually is affected by face detection and alignment. For example, precise positions and outlines of faces and aligned facial landmarks can improve the recognition accuracy. In the following content, I am going to show several recent work which accurately recognize faces and shows robustness to the different views of faces and occlusion.

A. DeepFace

DeepFace [17] is a milestone in the field of recognizing faces with deep CNN architectures and extended by numerous work [19], [20], [30]. The core idea behind DeepFace is to train a deep CNN architecture to extract face features and use a Siamese Network architecture for classification. To increase the recognition accuracy, it exploits explicit 3D modeling of faces to align faces before using CNN to extract features of faces and it also use the largest facial dataset at that time for training, which contains 4 million examples spanning 4000 unique identities. In the experiments, it achieves 97.35% face verification accuracy on the Labeled Faces in the Wild (LFW) dataset[31] closed to the human performance of 97.5% [32]. However, explicit 3D models of faces for face alignment is complex and time-consuming which is replaced by weak face alignment (aligning several facial landmarks) in the future work.

B. DeepID

At the same of DeepFace, DeepID was proposed in [18] and got 97.45% face verification accuracy on LFW. Similar to DeepFace, it extracts face features with a CNN architecture (whose elements in the last hidden layer is extracted as features). Different from DeepFace, DeepID only uses five facial points to align faces and splits the aligned face into 100 patches, each of which is feed into the CNN architecture to extract features. Then, the features are concatenated to a larger vector and compressed into a 150-dimensional features by Principal Component Analysis. Finally, features of two faces are input into Joint Bayesian Classifier to determine whether they are identical. Both of DeepFace and DeepID only consider the task of face verification or identification, [19] demonstrate that the multi-task learning (i.e., combining the loss function of face verification and identification to train the model) can improve the performance. Besides, the deep structure makes this method hard to run in real time and implement in some non-high-performance devices.

C. DeepID2 & DeepID3

As a DeepID variant, DeepID2[19] demonstrates that the multi-task learning can improve extracted features of faces. The loss function of DeepID2 is a weighted combination of the loss functions of the face verification and identification. It achieves 99.15% face verification accuracy on LFW and reduces 67% error rate of the best deep learning result at that time, 97.45% [18]. For the binary classification, the deep structure usually helps to improve the classification accuracy. DeepID3[20] tries a very deep CNN architecture with the joint identification and verification loss function for face recognition and gets better performance on LFW, 99.53% face verification accuracy and 96.0% face identification accuracy respectively. Compared with DeepID, both architectures are more accurate but deeper and more complex.

IV. Discussion and Conclusion

This brief review summarizes the early methods for face detection and recognition and introduces recent milestones in both fields, i.e., R-CNN and DeepFace. Ideas behind these remarkable work make a great difference to recent work and inspire lots of outstanding methods. By comparing early and current work, we can find early approaches do not have high detection or recognition accuracy but their speed is pretty high while the current methods exploit complex CNN architectures and brilliantly improve face detection and recognition accuracies but the their complexity stops them from being applied on smartphones and cameras. Notably, both of the multi-task learning and deep structures play an important role in improving the performance of both of face detection and recognition techniques. With the limitation of time and energy, I can hardly cover all sorts of remarkable work on face detection and recognition with deep CNN architectures so I just list some more strong work here for interested readers: [33], [30], [24], [34], [25], [35].

References

[1] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[2] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[3] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[5] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *International Journal of Computer Vision*, vol. 74, no. 2, pp. 167–181, 2007.

[6] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 643–650.

[7] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*. Springer, 2014, pp. 720–735.

[8] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, 2007.

[9] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[10] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[11] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1197–1215, 2007.

[12] S. Roux, F. Mamalet, and C. Garcia, "Embedded convolutional face finder," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 285–288.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[18] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[19] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[20] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[21] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc2007) results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[25] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] V. Jain and E. G. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," *UMass Amherst Technical Report*, 2010.

[28] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.

[29] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.

[30] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[31] 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 Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[32] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 365–372.

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[34] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 525–542.

[35] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *The Conference on Computer Vision and Pattern Recognition*, 2017.