Face Recognition by Coarse-to-Fine Landmark Regression with Application to ATM Surveillance

Ya Li[1], Lingbo Liu[2], Liang Lin[2], and Qing Wang[2(✉)]

[1] Guangzhou University, Guangzhou 510006, China
liya@gzhu.edu.cn
[2] Sun Yat-sen University, Guangzhou 510006, China
liulingb@mail2.sysu.edu.cn, linliang@ieee.org, wangq79@mail.sysu.edu.cn

Abstract. While ATM provides us convenient banking services, it has great security risks. The authentication of only password requiring is not safe enough. With the rapid development of face recognition technology based on deep convolutional neural network (CNN), undoubtedly, applying it into ATM authentication will improve security further. In this paper, we explore a new authentication mode combine face recognition and basic password for ATM. We think that it would prevent the economic crime on ATM fundamentally. However, computational and storage costs of CNN based methods are still high. To this end, we propose a new face recognition method by landmark regression. Our pipeline integrates a landmark localization network with a light face recognition network. For landmark localization, we employ a fully convolutional neural network to produce facial landmark response maps directly from raw images in a coarse-to-fine manner. For face recognition, we train a light CNN to obtain a compact representation, where the rectified linear unit (ReLU) is replaced by max-feature-map (MFM). Our approach shows good performance on several datasets. And it is practicable due to its high speed, good accuracy, and low storage space requirement.

Keywords: Face recognition · Landmark localization
ATM surveillance · Deep CNN · Face verification

1 Introduction

ATM provides us many convenient services about banking, because most ATMs are open 24 h and their locations are spread all over a city. But at the same time, ATMs are one of the most vulnerable sites without any manual security. The criminal behaviours such as physical attack, damage on ATM, using a duplicated or stolen bank card to withdrawal or transfer money often occur. Many researches have been done to improve the security of ATM. Some works focus on suspicious wearing detection. Ray et al. [12] detected mask wearing by Viola-Jones algorithm, while Wen et al. [23] detected safety helmets wearing

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 62–73, 2017.
https://doi.org/10.1007/978-981-10-7302-1_6

using modified Hough Transform. Some works focused on abnormal behaviour detection. Tang et al. [22] detected peeping behavior based on Omni-Directional Vision sensor and computer vision technology. Arsic et al. [1] applied Bayes Markov chains to resolve heavy occlusions and detect robberies.

Most researchers paid their attention to the surrounding environment monitoring of ATM kiosks, and ignored the potential safety issue coming from ATM authentication mode itself. Currently, the authentication of only password requiring is not safe enough. With the rapid development of face recognition technology based on deep convolutional neural network (CNN), undoubtedly, applying it into ATM authentication will improve security further. Other biometric information like fingerprint and iris are often used for security, but in ATM system they are not very appropriate. Because we can grant privileges to other persons by upload their facial images if using face recognition technology, while the authorization is difficult to implement if using fingerprint or iris. In daily life, it would be inconvenient if only cardholders themselves are permitted to handle business, and authorization is a good resolution to guarantee both security and flexibility.

In this paper, we explore a new authentication mode combine face recognition and basic password for ATM. We think that it would prevent the economic crime on ATM fundamentally. However, real-time response, high accuracy, and low storage space requirement make this application very challenging. On one hand, there are two separated steps including face detection and alignment before recognition generally in face recognition pipeline and result in low efficiency. We simplify the general processing and present a novel landmark localization method to produce facial landmark from raw images without face detection preprocessing. On the other hand, computational and storage costs of CNN based methods are still high. To this end, we propose a new face recognition method by landmark regression. Our pipeline integrates a coarse-to-fine landmark localization network and a light face recognition network. In network architecture, we use fully convolutional network (FCN) in landmark localization and network-in-network with small convolutional kernel size in face recognition for saving storage space. Besides, considering the speed of process we obtain landmark locations in a coarse-to-fine manner. And the nine layers CNN of face recognition ensure the accuracy further.

For landmark localization, there are no fully connected layers in our network and only convolutional layers are utilized. In particular, by taking the whole image as input, the coarse locations of facial landmarks are roughly detected in global context. Then they are further refined by local regions, which taking the cropped patches containing coarse landmarks as input and produce a fine and accurate prediction. For face recognition, we train a light (CNN) where the rectified linear unit (ReLU) is replaced by max-feature-map (MFM) like paper [24]. The CNN with MFM can obtain a compact representation and shows good performance in terms of both computational cost and storage space [24].

The main contributions of this work are summarized as follows. (1) We propose a new face recognition method by coarse-to-fine landmark regression, which

produces facial landmark response maps directly from raw images without relying on the result of face detection or other preprocessing done in advance. (2) Our approach is practicable due to its good accuracy and high efficiency.

The reminder of this paper is organized as follows. In Sect. 2, we briefly review some related works on landmark localization and face recognition. Then in Sect. 3, we introduce our system in detail. Finally, we present our experimental results in Sect. 4 and conclude this paper in Sect. 5.

2 Related Work

As well known, deep learning methods, especially the deep CNNs and deep auto-encoders, have made dramatic progress on many computer vision tasks including face landmark localization and face recognition. Therefore, in this section, we mainly review the works using deep learning.

2.1 Face Landmark Localization

Luo et al. [11] proposed a hierarchical face parser by combined deep belief network (DBN) with deep auto encoder (DAE). Sun et al. [16] proposed three-level cascaded CNNs to detect facial landmarks gradually, in which the first CNN produced the initial predictions and the following two CNNs refine the results. Specifically, the adjustment from third level is more subtle than the second level. Zhang et al. [30] used correlated tasks, e.g. facial expression recognition and head pose estimation to optimize facial landmark detection by deep CNN. In paper [28] Zhang Jie et al. introduced a new stacked DAEs pipeline to progressively refine the facial landmark locations by taking gradually higher resolution image version as input. Zhang Cha et al. [27] built a deep CNN to learn the face/non-face decision, the face pose estimation, and the facial landmark localization simultaneously. Lai et al. [9] proposed an end-to-end CNN architecture to learn highly discriminative shape-indexed-feature for face alignment. Recently, a deep cascaded multi-task CNN framework [29] was proposed to exploit the inherent correlation between detection and alignment. Existing methods mentioned above either require a facial bounding box or need to extract proposals by inefficient sliding window, so their performance are not good enough for applying in unconstrained settings.

2.2 Face Recognition

Currently, deep learning method has become the most common method to be used for face recognition. Earlier work proposed in paper [17] used CNN for face verification, which was composed of hybrid CNNs lower part for feature learning and restricted Boltzmann machine (RBM) top layer for classification. Lower part CNNs took twelve different face regions as input and RBM merged the twelve group outputs to give the final prediction. In 2014, DeepFace [21] and

DeepID [18] appeared almost at the same time. DeepFace employed 3D alignment for preprocessing and achieved 97.35% accuracy on LFW dataset [6]. While DeepID aligned face image based on landmarks and achieved 97.47% on LFW. Later, the improved model of DeepID named DeepID2 [15] and DeepID2+ [19] improved the verification accuracy on LFW to 99.15% and 99.47% respectively. For DeepID, the outputs of different CNNs trained on different local patches were assembled as deep feature and Joint Bayesian was applied for face verification. For DeepID2, verification loss and classification loss were further combined to increase accuracy. Compared with DeepID2, DeepID2+ was improved by increasing the dimension of patch feature and adding supervision to early convolutional layers. Recently, a triplet-based CNN model named FaceNet [13] achieved the best result 99.63% on LFW. Although these CNN based methods have achieved a practicable accuracy on LFW dataset, however, their computational costs are still high due to their deep or multi-network architectures. Hence, it is inevitable to design a light model with a smaller number of parameters for lower computation and storage cost.

3 Our Approach

3.1 Pipeline

Our model consists of two main cascaded parts: face landmark localization and face recognition, both of them are realized by using CNN. We train the whole model by stochastic gradient descent (SGD) and use the Caffe library [7]. The detailed pipeline is illustrated in Fig. 1. To begin with, the image is acquired at the moment of the card insertion. And the following operations are preprocessing, landmark localization, face alignment and face recognition. If current image is matched with card-holder or one of the authorized user's images, the process is continued. Otherwise, while asking him whether take photo again we can implicitly recognize the current user's identification to judge whether the user is a suspect, if yes, send alarm signal to monitoring center by background program. Specifically, we perform standard histogram equalization preprocessing to adjust the image contrast. And face alignment is done by horizontal rotation based on landmarks. In the next two subsections, we introduce face landmark localization and face recognition respectively.

3.2 Coarse-to-Fine Landmark Localization

Landmark locations are obtained in a coarse-to-fine way by FCNs, where convolution is used like a filter. Suppose $L_i^k = (x_i^k, y_i^k)$ is the location of i-th landmark of type k in image I, where (x_i^k, y_i^k) represents the landmark coordinates. The filtering is performed on patches. We denote $F_{\mathbf{W}^k}(P)$ as the filtering function of type k on patch P with parameters $\mathbf{W}^k$. Suppose the patch size is $w \times h$, and the filtering function is in sliding window manner with stride δ. The response map $F_{\mathbf{W}^k} * I$ whose value at location (x, y) can be computed by:

$$(F_{\mathbf{W}^k} * I)(x, y) = F_{\mathbf{W}^k}(I(x\delta : x\delta + w, y\delta : y\delta + h)), \tag{1}$$

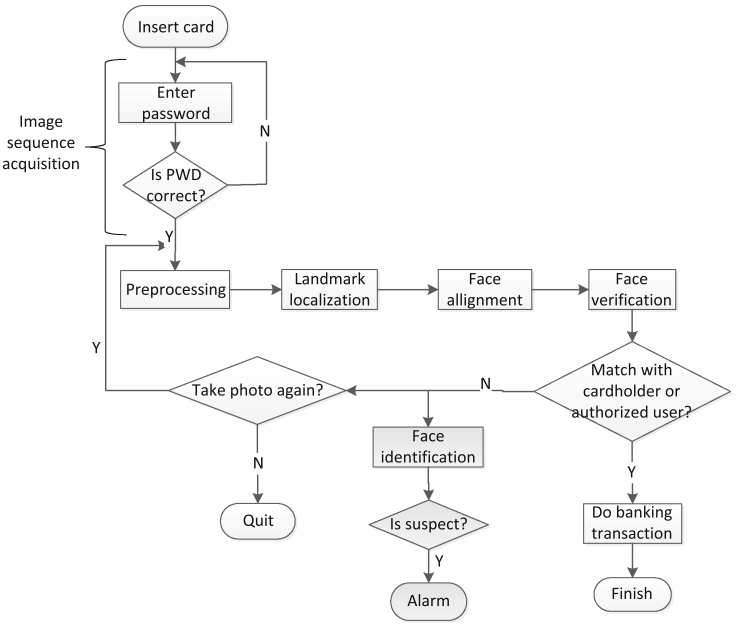

Fig. 1. The proposed pipeline of face recognition. The image is acquired at the moment of the card insertion. If current image is matched with card-holder or one of the authorized user's images, the process is continued. Otherwise, while asking him whether take photo again we can implicitly recognize the current user's identification to judge whether the user is a suspect, if yes, send alarm signal to monitoring center by background program.

where $I(x\delta : x\delta + w, y\delta : y\delta + h)$ is the image patch. At first, we take the whole image as input and obtain the response map. We wish that the response map should distinguish whether the patches containing landmarks or not. Next we only input the patches containing coarse landmarks for refining by taking the same filtering process. Thus, we can construct a multi-level networks pyramid by decreasing the path size for filtering gradually.

Our destination is learning the filtering functions which have the property: patches containing the target landmarks should have strong response, otherwise should have weak response. Suppose the threshold is θ and the landmark location is the patch center, which can be computed by:

$$Det(I) = \{(x\delta + w/2, y\delta + h/2) | (F_{\mathbf{W}^k} * I)(x, y) > \theta\}. \tag{2}$$

Let $H^k(I; \mathbf{W}_p)$ and $H_0^k(I)$ denote the predicted and ground truth response map of image I for landmark type k, where $\mathbf{W}_p$ denote the parameters of the pth level network pyramid. The value of $H^k(I; \mathbf{W}_p)$ at position (x, y) can be computed

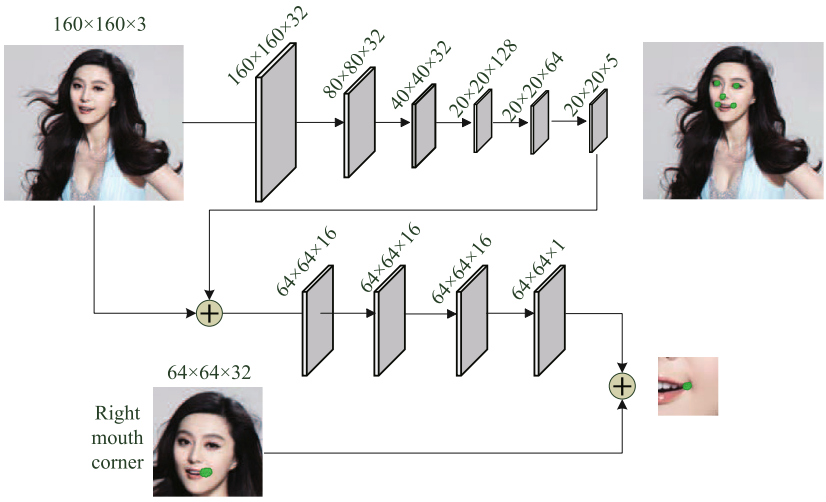

Fig. 2. The network architecture of our coarse-to-fine landmark localization. The first level network is shared by all of the landmarks. The second level sub-networks are unique to each landmark. We only illustrate "right-mouth-corner" sub-network for simplification.

by Eq.(1). The loss function used to train the network is:

$$\mathcal{L}_p(I; \mathbf{W}_p) = \sum_{k=1}^{K} ||H^k(I; \mathbf{W}_p) - H_0^k(I)||^2. \tag{3}$$

We evaluate our approach by stacking two level networks on five landmarks localization, including two eyes' centres, left mouth corners, right mouth corners and nose tip. The first level network is shared by all of the landmarks, and the second level sub-networks are unique to each landmark. The network architecture is showed in Fig. 2, where only "right-mouth-corner" sub-network is illustrated for simplification. In summary, the kernel sizes of the first level convolutional layers are 5×5, 5×5, 5×5, 9×9, 1×1, 1×1; and for the second level they are 5×5, 7×7, 9×9, 1×1 respectively.

4 Face Recognition

As well known, CNN-based face recognition systems need large amounts of memory and computational power. Although they perform well on GPU-based machines, it is often a challenge to run them on target low-power devices like ATMs due to overtaxing the compute capabilities and the limited storage. Hence, it needs us to seek a light model for the deployment on ATMs. Much research work is done to speed-up and compact CNNs. Here we utilize an off-the-shelf

named light B model proposed in paper [24] for face recognition, which is cascaded with our landmark localization network. In particular, a rough face alignment is made before they are input for recognition according to the five landmarks location. We rotate two eye landmarks horizon-tally to overcome the pose variations in roll range and fix the distance between the midpoint of eyes and the midpoint of mouth for facial image normalization.

The light B model we used has two main characteristics: (1) defining Max-Feature-Map (MFM) activation function for compact representation and (2) employing net-work in network (NIN) [10] concept between convolution layers for improvements on speed and storage.

MFM is defined to obtain competitive feature maps, it outputs element-wise maximum of two convolutional feature maps. Given an input convolution layer $C \in \mathbb{R}^{h \times w \times 2n}$, where $2n$ is the channel number of convolution layer and $h \times w$ denotes the map size. The MFM can be formulated as:

$$f_{ij}^{k} = \max_{1 \leq k \leq n} (C_{ij}^{k}, C_{ij}^{k+n}), \tag{4}$$

where $1 \leq i \leq h$ and $1 \leq j \leq w$.

The light B contains 5 convolution layers, 4 NIN network layers, MFM activation function, 4 max-pooling layers and 2 fully connected layers. The small convolution kernel size with NIN can reduce the number of parameters for the model; it makes the model 20 times smaller than VGG [14] while 9 times faster on CPU time. Please refer to [24] for more academic details.

5 Experiments

In this section, we evaluate our landmark localization approach first and then show the performance of face recognition based on landmark locations using light model.

5.1 Landmark Localization

For model training, we collect 7,317 face images and 1,671 natural images from Internet. Among them 6,317 face images and 1,218 natural images for training, the rest of 1,000 face images and 453 natural images for validation. Each face is annotated with five landmarks. We employ AFLW [8] and AFW [31] for evaluation. For AFLW, we selected 3,000 faces randomly from AFLW for testing like TCDCN [30].

We report our result on mean error, which is measured by the distances between estimated landmarks and the ground truths, normalized with respect to inter-ocular distance. It is defined as:

$$err = \frac{\sqrt{(x - x')^2 + (y - y')^2}}{l}, \tag{5}$$

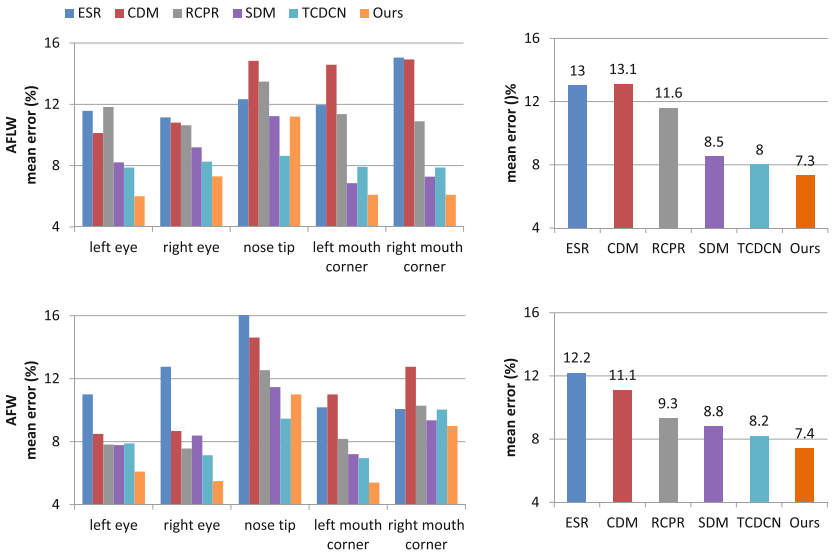

Fig. 3. Comparison with other methods on AFLW and AFW. The comparison on five landmarks is illustrated in the left sub-image, and the average mean errors of these methods are summarized in the right sub-image.

where (x, y) and (x', y') are the ground truth and predicted locations, respectively, and l is the inter-ocular distance.

We compare our method against: (1) Explicit Shape Regression (ESR) [3]; (2) A Cascaded Deformable Model (CDM) [26]; (3) Robust Cascaded Pose Regression (RCPR) [2]; (4) Supervised Descent Method (SDM) [25]; (5) Task-Constrained Deep Convolutional Network (TCDCN) [30]. The results are reported in Fig. 3. On the AFLW dataset, our method achieves 7.3% for average mean error, 8.75% improvement over TCDCN. On AFW, our average mean error is 7.4% over five parts, 9.76% improvement over TCDCN.

5.2 Face Recognition Based on Landmark Locations

We evaluate the cascaded model which integrates our landmark localization with the light B for face verification task on LFW and CACD-VS [4] datasets. LFW contains 13,233 images of 5,749 identities. For face verification, face images are divided in 10 folds and each folder contains 600 face pairs of different identities. CACD-VS dataset contains 2,000 positive pairs and 2,000 negative pairs which are collection of celebrity images on Internet.

Face verification is one-to-one matching, which identifies whether two facial images are from the same person. We compare our method against WebFace [20], VGG [14], FaceNet [13] and Light B [24] on LFW and against HFA [5], CARC [4], VGG and Light B on CACD-VS. The results are reported in Table 1.

Table 1. The accuracies of different methods on LFW and CACD-VS datasets.

LFW		CACD-VS	
Method	Accuracy	Method	Accuracy
WebFace	96.13%	HFA	84.4%
VGG	97.27%	CARC	87.6%
FaceNet	**99.63%**	VGG	96%
Light B	98.13%	Light B	97.95%
Ours	98.5%	Ours	**98.2%**

Our cascaded model shows very competitive results compared to other state-of-the-art methods on both datasets.

We also show some verification results of our approach on LFW dataset in Fig. 4. The first line illustrates some errors, where the first three pairs are errors on mismatched pairs, that is, the pairs of mismatched images that were incorrectly reported as matched pairs. And the last three pairs are errors on matched pairs, that is, the pairs of matched images that were incorrectly reported as mismatched pairs. And the second line demonstrates the robustness of our approach, where the hard pairs are verified correctly despite of variations on pose, expression, occlusion, and illumination. It is worth mentioning that our model is practicable to apply in ATM system due to the good performance on speed and storage space. The entire model is about 33 MB and it requires about 140 ms for face verification on low power PC with Xeon(R) E5-2620 v3 @ 2.40GHz CPU.

Fig. 4. Illustration of verification results of our method on LFW dataset. The first line illustrates some errors, where the first three pairs are errors on mismatched pairs, the last three pairs are errors on matched pairs. And the second line demonstrates the robustness of our approach, where the hard pairs are verified correctly despite of variations on pose, expression, occlusion, and illumination.

6 Conclusion

In this paper, we explore a new authentication mode combine face recognition and basic password for ATM. We propose a cascaded deep model which integrates a landmark localization network with a light face recognition network.

Our model shows good performance on several datasets. It is practicable due to its high speed, good accuracy, and low storage space requirement. The entire model is about 33 MB and it requires about 140 ms for face verification on low a power PC with Xeon(R) E5-2620 v3 @ 2.40 GHz CPU. The face recognition network is the bottleneck on speed and storage. The time of landmark localization on one image is only 7 ms, but for face verification is 133 ms. The landmark localization network needs less than 1 M storage space while recognition network needs 32 MB. Therefore, we will continue to optimize the recognition network further. In future, we also generalize our approach in more challenging object recognition, such as human action understanding.

Acknowledgments. This research is supported by the Research Project of Guangzhou Municipal Universities (No. 1201620302), National Undergraduate Scientific and Technological Innovation Project (No. 201711078017), the Science and Technology Planning Project of Guangdong Province (No. 2015B010128009, 2013B010406005). The authors would like to thank the reviewers for their comments and suggestions.

References

1. Arsic, D., Lyutskanov, A., Kaiser, M., Schuller, B., Rigoll, G.: Applying Bayes Markov chains for the detection of ATM related scenarios. In: Applications of Computer Vision, pp. 1–8 (2010)
2. Burgosartizzu, X.P., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: International Conference on Computer Vision. pp. 1513–1520. IEEE (2013)
3. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vision **107**(2), 177–190 (2014)
4. Chen, B.C., Chen, C.S., Hsu, W.: Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimedia **17**(6), 804–815 (2015)
5. Gong, D., Li, Z., Lin, D., Liu, J., Tang, X.: Hidden factor analysis for age invariant face recognition. In: International Conference on Computer Vision, pp. 2872–2879. IEEE (2013)
6. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report (2008)
7. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
8. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: International Conference on Computer Vision Workshops, pp. 2144–2151. IEEE (2011)
9. Lai, H., Xiao, S., Cui, Z., Pan, Y., Xu, C., Yan, S.: Deep cascaded regression for face alignment. arXiv preprint arXiv:1510.09083 (2015)
10. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1502.03167 (2013)

11. Luo, P.: Hierarchical face parsing via deep learning. In: Computer Vision and Pattern Recognition, pp. 2480–2487. IEEE (2012)
12. Ray, S., Das, S., Sen, A.: An intelligent vision system for monitoring security and surveillance of ATM. In: IEEE India Conference, pp. 1–5. IEEE (2015)
13. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Computer Vision and Pattern Recognition, pp. 815–823. IEEE (2015)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1412.5903 (2014)
15. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988–1996 (2014)
16. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition, pp. 3476–3483. IEEE (2013)
17. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: International Conference on Computer Vision, pp. 1489–1496. IEEE (2013)
18. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Computer Vision and Pattern Recognition, pp. 1891–1898. IEEE (2014)
19. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Computer Vision and Pattern Recognition, pp. 2892–2900. IEEE (2015)
20. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Web-scale training for face identification. In: Computer Vision and Pattern Recognition, pp. 2746–2754. IEEE (2015)
21. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Computer Vision and Pattern Recognition, pp. 1701–1708. IEEE (2014)
22. Tang, Y., He, Z., Chen, Y., Wu, J.: ATM intelligent surveillance based on omnidirectional vision. In: Computer Science and Information Engineering, pp. 660–664. IEEE (2009)
23. Wen, C.Y., Chiu, S.H., Liaw, J.J., Lu, C.P.: The safety helmet detection for ATM's surveillance system via the modified hough transform. In: IEEE International Carnahan Conference on Security Technology, pp. 364–369. IEEE (2003)
24. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. arXiv preprint arXiv:1511.02683 (2015)
25. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition, pp. 532–539. IEEE (2013)
26. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: International Conference on Computer Vision, pp. 1944–1951. IEEE (2014)
27. Zhang, C., Zhang, Z.: Improving multiview face detection with multi-task deep convolutional neural networks. In: Winter Conference on Applications of Computer Vision, pp. 1036–1041. IEEE (2014)
28. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_1

29. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
30. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
31. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)