Pose-Invariant Face Recognition Based on a Flexible Camera Calibration

Xiaohu Shao, Cheng Cheng(✉), Yanfei Liu, and Xiangdong Zhou

Chongqing Institute of Green and Intelligent Technology,
Chinese Academy of Sciences, Chongqing, China
{shaoxiaohu,chengcheng,liuyanfei,zhouxiangdong}@cigit.ac.cn

Abstract. In this paper, we present a flexible camera calibration for pose normalization to accomplish a pose-invariant face recognition. The accuracy of calibration can be easily influenced by errors of landmark detection or various shapes of different faces and expressions. By jointly using RANSAC and facial unique characters, we explore a flexible calibration method to achieve a more accurate camera calibration and pose normalization for face images. Our proposed method is able to eliminate noisy facial landmarks and retain the ones which best match the undeformable 3D face model. The experimental results show that our method improves the accuracy of pose-invariant face recognition, especially for the faces with unsatisfied landmark detection, variant shapes, and exaggerated expressions.

Keywords: Camera calibration · 3D alignment · Face recognition

1 Introduction

Face recognition plays an important role in pattern recognition and computer vision applications. In recent years, face recognition has made great progress with deep learning technique developing. Methods using deep learning and large training dataset [1–4] have almost achieved super-human accuracy on the LFW benchmark [5,6]. However, it remains a difficult problem for faces in the wild due to the variations in pose, illumination and expression. More specifically, different poses of the same face have dramatically different appearances, causing fatal problems to most of current face recognition systems.

In order to solve the aforementioned problems, many approaches have been explored, they can be categorized into feature-based methods and normalization-based methods.

The pose insensitive feature-based methods are widely used, they try to extract specific features which are invariant or insensitive to different poses. Wiskott *et al.* [7] collapse face variance of pose and expression by extracting concise face descriptions in the form of image graphs. Gross *et al.* [8] develop the theory of appearance-based face recognition from light-field, which leads directly to a pose-invariant face recognition algorithm that uses as many images of the

T. Tan et al. (Eds.): CCPR 2016, Part I, CCIS 662, pp. 191–200, 2016.
DOI: 10.1007/978-981-10-3002-4_16

face as are available. Lai *et al.* [9] use wavelet transform and multiple view images to determine the reference image representation. Restricted to capacity of these representations and limited dataset, above mentioned methods are not able to get satisfied features which is insensitive to pose of faces in the wild. DCNN based face recognition have been widely reported in recent studies, because features trained by DCNN with huge size of dataset have a strong representation for variant of object, they achieve state-of-the-art performances on recognition of different poses of faces. Taigman *et al.* [1] derive a face representation form a nine-layer deep neural network. Sun *et al.* [2] propose to learn a set of high-level feature representations which called DeepID feature through deep learning for face verification. In 2014, they proposed two very deep neural network architectures to achieve a higher face identification accuracy [3]. Liu *et al.* [4] combine a multi-patch deep CNN and deep metric learning to extract low dimensional but very discriminative feature for face recognition.

Normalize-based method tries to normalize different faces to a unified frontal face to improve the accuracy of recognition. Chai *et al.* [10] use locally linear regression (LLR) to generate the virtual frontal view from a given non-frontal face image, this method is not able to always preserve the identity information. Berg [11] takes advantage of a reference set of faces to perform an identity-preserving alignment, warping the faces in a way that reduces differences due to pose and expression. Hu *et al.* [11] reconstruct a 3D face model from a single frontal face image, and synthesize faces with different PIE to characterize face subspace. Wang [12] proposes a fully automatic, effective and efficient framework for 3D face reconstruction based on a single face image in an arbitrary view. Asthana *et al.* [13] build a 3D Face Pose Normalization system which improves the recognition accuracy of face variation up to ±45° in yaw and ±30° in pitch angles. Zhu *et al.* [14] present a pose and expression normalization method to recover the neutral frontal faces without little artifact and information loss. Hasser *et al.* [15] use an unmodified 3D reference to approximate shape of all query faces and synthesize frontal faces. These 3D-based methods estimate the normalization transformations from correspondence between 2D and 3D facial landmarks, they are often efficient but suffers from errors and variety of landmarks which are caused by landmark detection, various shapes and exaggerated expressions.

Inspired by the above approaches, we present a flexible camera calibration for 3D alignment in order to improve pose-invariant face recognition. Different with work [14], we present a flexible camera calibration based on RANSAC [16] and facial unique characters to estimate poses of faces for pose normalization of faces. Our flexible camera calibration is insensitive to outliers of landmarks caused by landmark detection or variant of shape and expressions. The experimental results show that our method improves the accuracy of pose-invariant face recognition, especially for the faces with unsatisfied landmark detection, variant shapes, and exaggerated expressions.

Our pose-invariant face recognition includes three steps: First, we estimate the pose of a face using our proposed flexible camera calibration from

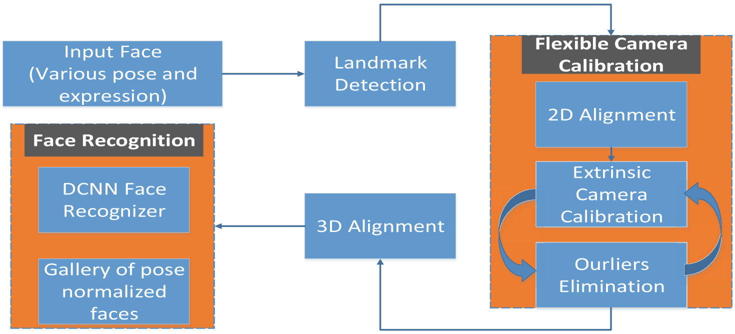

Fig. 1. Framework of the pose-invariant face recognition system

correspondence between 2D landmarks and an undeformable 3D face model. Second, we calculate the transformations of 3D alignment based on the estimated pose. Finally, we get the pose-normalized face and use them to train DCNN model for face recognition. The framework of our pose-invariant face recognition system is shown in Fig. 1.

The remainder of this paper is organized as follows: Sect. 2 introduces the details of flexible camera calibration and framework of our pose-invariant face recognition. Section 3 provides the experimental results of proposed method compared with other methods on face recognition. The conclusion and future work is provided in Sect. 4.

2 Facial Pose Normalization

Previous work of face recognition have witnessed the efficiency of the pose-normalized face and 3D face. In this section, we normalize poses of faces by proposed flexible camera calibration from correspondence between 2D landmarks and an undeformable 3D face model.

The problem of camera calibration can be described as follows: Given a mean 3D model of face $\mathbf{S} \in \Re^{3\times n}$ with total n vertices, landmarks on the 2D face $\mathbf{s} \in \Re^{2\times n}$, the goal is to estimate the intrinsic camera parameters $\mathbf{A} \in \Re^{3\times 3}$, rotation matrix $\mathbf{R} \in \Re^{3\times 3}$ and translation vector $\mathbf{t} \in \Re^{3\times 1}$. $[\mathbf{R}, \mathbf{t}]$ is also known as extrinsic camera parameters. To find the parameters that best project the 3D face model to the 2D landmarks, we solve the nonlinear least squares optimization problem:

$$\{\mathbf{A}^*, \mathbf{R}^*, \mathbf{t}^*\} = \min_{\mathbf{A},\mathbf{R},\mathbf{t}} \|\mathbf{f}(\mathbf{A}, \mathbf{R}, \mathbf{t}, \mathbf{S}) - \mathbf{s})\|_F^2, \tag{1}$$

$$\mathbf{f} = \mathbf{f}_1 \circ \mathbf{f}_2, \tag{2}$$

$$\mathbf{f}_1(\mathbf{A}, \mathbf{R}, \mathbf{T}, \mathbf{S}) = \mathbf{A}(\mathbf{RS} + \mathbf{T}), \tag{3}$$

$$\mathbf{f}_2(\mathbf{S}) = \begin{bmatrix} \mathbf{S}_1^\top \oslash \mathbf{S}_3^\top \\ \mathbf{S}_2^\top \oslash \mathbf{S}_3^\top \end{bmatrix} \tag{4}$$

where $\mathbf{T} = [\mathbf{t}, \mathbf{t}, ...] \in \Re^{3\times n}$ consists of n copies of $\mathbf{t}$, $\mathbf{f}_2$ projects 3D vertices into 2D image, $\oslash$ denotes element-wise division, $\mathbf{S}_i$ is the row vector of i.

In order to get the correspondence of 3D face model and 2D landmarks, we get a mean 3D face model obtained from USF Human ID 3D face [18] and 2D landmarks by recent methods of facial landmark detection. We select 49 vertices from 70000 vertices to reconstruct a simple 3D face model. Automatic facial landmark detection on face images has been well studied [17–22], We select the method [19] for its satisfied accuracy on faces with large poses and its efficiency. Similarly with work [15], we retain 49 facial landmarks and exclude the contour landmarks, because different poses would change the matching relationship of contour landmarks and vertices of the 3D model.

2.1 Intrinsic Parameter Unit by 2D Alignment

Estimating the intrinsic parameters $\mathbf{A}$ and extrinsic parameters $[\mathbf{R}, \mathbf{t}]$ at the same time for a single image is an ill-pose problem. Work [22] estimates $\mathbf{A}$ by using many frames as its initialization. Work [15] uses a fixed $\mathbf{A}$ for aligned LFW images. The sizes and locations of faces on LFW images are almost the same, they can be seen sharing the same intrinsic matrix. But for an arbitrary image, its unsuitable to use the supposed intrinsic parameters. An approximate $\hat{\mathbf{A}}$ can be fixed when a face image I is aligned into coordinate of standard LFW dataset by similarity transformation. The source is the 2D facial landmarks, and target shape is the reference landmarks $\bar{\mathbf{s}}$, which can be calculated from the mean shape of all shapes in LFW images. The aligned landmarks $\hat{s}$ and image $\hat{I}$ is shown in Fig. 2.

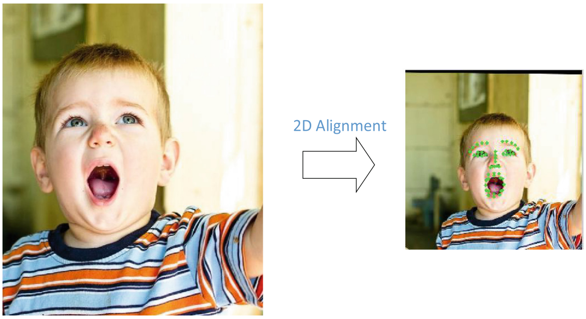

Fig. 2. An example of 2D alignment.

2.2 Flexible Extrinsic Camera Calibration

After the face is 2D aligned by similarity transformation, $[\mathbf{R}, \mathbf{t}]$ is to be estimated from the 2D facial landmarks and the 3D face model:

$$\{\mathbf{R}^*, \mathbf{t}^*\} = \min_{\mathbf{R},\mathbf{t}} \left\| \mathbf{f}(\hat{\mathbf{A}}, \mathbf{R}, \mathbf{t}, \mathbf{S}) - \hat{\mathbf{s}} \right\|_F^2 \tag{5}$$

The above problem is known as 3D pose estimation, which is usually solved by iterative method based on *Levenberg-Marquardt Algorithm* (LMA) [23]. This optimization is efficient and accurate when the vertices of 3D face are able to match the 2D landmarks very well. However, as noises often exist in landmark detection and different person with expressions have various shapes of landmarks, it is impossible to match the various 2D landmarks with the undeformable 3D model accurately. These matching errors decrease the accuracy of pose estimation, so we need to eliminate these large errors of landmarks before the iteration.

RANSAC is an iterative method to estimate parameters of a mathematical model from a set of data which contains outliers [24]. However, when the number of iteration computed is limited, the solution may not be optimal. Considering efficiency and accuracy of pose normalization, we cannot afford no limited iterations.

When we use RANSAC to eliminate the outliers of facial landmarks on a large dataset, we observe that outliers often appear as landmarks of particular parts, such as eyebrow, top and bottom of mouth. It seems that the accuracy of these landmarks location is less than other landmarks, or these landmarks are not able to match the undeformable 3D model very well caused by variant of person and expressions. The probability distribution of each landmark which is labeled as an outlier in dataset by general RANSAC is shown in Fig. 3.

In order to speed up outlier elimination of landmarks, we separate all N landmarks in two pools according with their probability distribution labels as an inlier in training dataset: inliers pool $\mathbf{\Phi} = \{\phi_1, \phi_2, ..., \phi_p\}$, outliers pool $\mathbf{\Psi} = \{\psi_1, \psi_2, ..., \psi_q\}$, where ϕ_i denotes the i^{th} landmark which is labeled an inlier with large probability, ψ_j denotes the j^{th} landmark which is labeled as an outlier with large probability. In the process of eliminating outliers, landmarks belonged to $\mathbf{\Phi}$ are selected to calculate the pose using *LMA* optimization with less probability, landmarks belonged to $\mathbf{\Psi}$ are selected as inliers with more chance. The process of flexible extrinsic camera calibration is summarized in Algortihm 1. First, we use all of landmarks to estimate the initial $[\mathbf{R}, \mathbf{t}]$. Second, we project the 3D model into the 2D image and calculate the distance

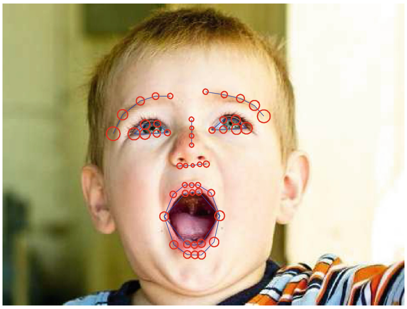

Fig. 3. Probability distribution of each landmark which are labeled as outliers in dataset. All landmarks are drawn by red circles with different sizes. The larger size of circle represents that the current landmark is labeled as an outlier with larger probability. (Color figure online)

between each projected landmark and the corresponding real landmark. Third, landmark noises are eliminated by comparing the threshold and the normalized distance. We control the opportunity of elimination by setting the threshold θ_1 for landmarks belong to $\mathbf{\Phi}$ larger than threshold θ_2 for landmarks belong to $\mathbf{\Psi}$.

Algorithm 1. Flexible Extrinsic Camera Calibration

```
Input: 2D aligned facial landmarks $\hat{\mathbf{s}}$, 3D face model $\mathbf{S}$, instrinsic camera parameter
    $\hat{\mathbf{A}}$, index pool of inliers and outliers $\mathbf{\Phi}$, $\mathbf{\Psi}$.
Output: rotation matrix $\mathbf{R}^*$, translation vector $\mathbf{t}^*$.
while not converged do
    Calculate $\mathbf{R}$ and $\mathbf{t}$ by using $LMA$ with $\hat{\mathbf{s}}$ and $\mathbf{S}$.
    Project $\mathbf{S}$ to 2D landmarks $\mathbf{s_{proj}}$ by using $\hat{\mathbf{A}}$, $\mathbf{R}$ and $\mathbf{t}$.
    Cacluate distance $\mathbf{D} = \{d_1, d_2, ..., d_N\}$ between each landmark of $\hat{\mathbf{s}}$ and $\mathbf{s}_{proj}$.
    Obtain $\mathbf{D_{sort}} = \{d_{sort,1}, d_{sort,2}, ..., d_{sort,L}\}$ by sorting elements of $\mathbf{D}$ in a
    descending order.
    Find the index $\mathbf{E} = \{e_1, e_2, ..., e_N\}$ of the first $L$ elements of $\mathbf{D_{sort}}$ in $\hat{\mathbf{s}}$.
    for $i = 0 \rightarrow L - 1$ do
        if $(e_i \in \mathbf{\Phi}$ and $d_{sort,i} > \theta_1)$ or $(e_i \in \mathbf{\Psi}$ and $d_{sort,i} > \theta_2)$ then.
            Eliminate the landmark of index $e_i$.
        end if
    end for
end while
Generate the final parameters $[\mathbf{R}^*, \mathbf{t}^*]$.
```

In our experiments, we set $\theta_1 = 0.08$, $\theta_2 = 0.05$, $L = 10$ when $N = 49$, outlier elimination quickly converges in only 1 or 2 stages.

2.3 3D Alignment and Face Recognition

After the extrinsic parameters are calculated, we caculate the normalization transformation based on the estimated poses $[\hat{\mathbf{A}}, \mathbf{R}^*, \mathbf{t}^*]$. Then, we get the 3D aligned faces (more details can be found in [15, 25]) and use them to train models for face recognition. An example of our 3D alignment result can be seen in Fig. 4.

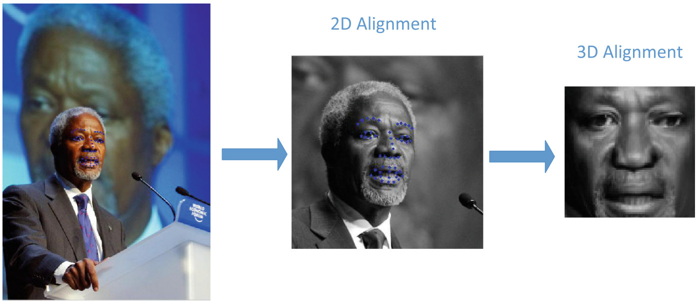

Fig. 4. An example of our 3D alignment result.

3 Experiments

In this section, we present experimental results of our proposed method on face recognition.

3.1 Database

LFW: LFW dataset consists of more than 13000 faces of 5749 celebrities. Each face has been labeled with the name of the person pictured. The number of faces varies from 1 to 530 for one person. 1680 of the people pictured have two or more distinct photos in the dataset. It is the most commonly used database for unconstrained face recognition.
CASIA-WebFace [26]**:** It contains 10575 subjects and 49414 images, which are collected from Internet by a semi-automatical way. CASIA-WebFace is prepared for training, LFW is used to evaluate our alignment compared with other alignment methods.

3.2 2D Alignment and 3D Alignment

After detecting faces [27] and landmarks [19] in an image, we use facial landmarks to normalize faces. 2D affine transformation is often used to align faces for improving face recognition. It is used to approximately scale, rotate and translate the image into a new warped image. It is also called 2D-alignment, pose normalization is often called 3D-alignment, which can be applied to compensate out-of-plane rotation. In this paper, we compare the three methods of alignment in face recognition: 2D alignment, 3D alignment of [15], and our proposed method.

3.3 Performance Analysis

We get aligned faces by applying the three alignment methods for training and test datasets, then we train three DCNN models on the training dataset. To evaluate the discriminative capability of the face representation, we compare the cosine distance of a pair of a normalized features which are transformed by PCA. The comparison of face recognition results on LFW by applying standard protocols and BLUFR protocols [26] are listed in Tables 1 and 2. The results show that our method is better than the other two normalization methods. We train models with the *BN-inception v1* network [28] on *Caffe* platform [29] from scratch for DCNN models training.

Because the limitation of GPU resources and the scale of training set, and our goal is only to show that face recognition can benefit from our 3D alignment method, we do not get the best result compared with the recent results on LFW. We believe that we can get state-of-the-art face recognition performance using our proposed method if we continued to adjust parameters, enlarge dataset and train deeper models.

Table 1. The performance of our proposed method compared with other methods on LFW under standard protocol.

Method	Accuracy ± SE
2D alignment	0.9623 ± 0.0107
3D alignment of [15]	0.9660 ± 0.01
Proposed method	**0.9673 ± 0.0081**

Table 2. The performance of our proposed method compared with other methods on LFW under standard protocol.

Method	VR@FAR = 0.1 %	DIR@FAR = 1 %, Rank = 1
2D alignment	68.64 %	34.74 %
3D alignment of [15]	73.27 %	37.47 %
Proposed method	**74.72** %	**38.57** %

4 Conclusion and Future Work

In this paper, we present a flexible camera calibration for 3D alignment to improve pose-invariant face recognition. Compared with previous normalization work, our method based on RANSAC and facial unique characters is insensitive to outliers of landmarks caused by landmark detection or variant of person and expressions. Experiments show that it the best performance on recognition of faces under complicated environment.

In the future, we will continue to improve our 3D alignment method to overcome the difficulty brought by various poses and expressions of faces. We will also get a further study to solve this problem by applying the deep learning method.

Acknowledgement. This work was supported by the Project from National Science Foundation for Youths of China (No. 61502444), National Natural Science Foundation of China (No. 61472386), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA06040103). Two Titan X GPUs applied for this research were donated by the NVIDIA Corporation.

References

1. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
2. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483 (2013)

3. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014)
4. Liu, J., Deng, Y., Huang, C.: Targeting ultimate accuracy: face recognition via deep embedding (2015). arXiv preprint arXiv:1506.07310
5. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07–49, University of Massachusetts, Amherst (2007)
6. Huang, G.B., Learned-Miller, E.: Labeled faces in the wild: updates and new reporting procedures. Technical report 14–003, Department of Computer Science, University Massachusetts Amherst, Amherst, MA, USA (2014)
7. Wiskott, L., Fellous, J.M., Kuiger, N., Von Der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 775–779 (1997)
8. Gross, R., Matthews, I., Baker, S.: Appearance-based face recognition and light-fields. IEEE Trans. Pattern Anal. Mach. Intell. **26**(4), 449–465 (2004)
9. Lai, J.H., Yuen, P.C., Feng, G.C.: Face recognition using holistic fourier invariant features. Pattern Recogn. **34**(1), 95–109 (2001)
10. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition. IEEE Trans. Image Process. **16**(7), 1716–1725 (2007)
11. Hu, Y., Jiang, D., Yan, S., Zhang, L., Zhang, H.: Automatic 3D reconstruction for face recognition. In: 2004 Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 843–848. IEEE (2004)
12. Wang, C., Yan, S., Li, H., Zhang, H., Li, M.: Automatic, effective, and efficient 3D face reconstruction from arbitrary view image. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3332, pp. 553–560. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30542-2_68
13. Asthana, A., Marks, T.K., Jones, M.J., Tieu, K.H., Rohith, M.: Fully automatic pose-invariant face recognition via 3D pose normalization. In: Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV), pp. 937–944. IEEE (2011)
14. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 787–796 (2015)
15. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4295–4304 (2015)
16. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
17. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)
18. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
19. Xiong, X., Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539 (2013)
20. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692 (2014)

21. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
22. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (TOG) **33**(4), 43 (2014)
23. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1330–1334 (2000)
24. Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. Int. J. Comput. Vis. **15**(1–2), 123–141 (1995)
25. http://www.openu.ac.il/home/hassner/projects/frontalize. Accessed 12 Feb 2015
26. Liao, S., Lei, Z., Yi, D., Li, S.Z.: A benchmark study of large-scale unconstrained face recognition. In: 2014 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2014)
27. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
28. Laurent, C., Pereyra, G., Brakel, P., Zhang, Y., Bengio, Y.: Batch normalized recurrent neural networks (2015). arXiv preprint arXiv:1510.01378
29. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding (2014). arXiv preprint arXiv:1408.5093