3D Face Alignment in the Wild: A Landmark-Free, Nose-Based Approach

Flávio H. de Bittencourt Zavan(✉), Antônio C.P. Nascimento, Luan P. e Silva, Olga R.P. Bellon, and Luciano Silva

Departmento de Informática, Universidade Federal do Paraná, Curitiba, Brazil
{flavio,antonio.paes,luan.porfirio,olga,luciano}@ufpr.br

Abstract. We present a methodology for 3D face alignment in the wild, such that only the nose is required as input for assessing the position of the landmarks. Our approach works by first detecting the nose region, which is used for estimating the head pose. After that, a generic face landmark model, obtained by averaging all training images, is rotated, translated and scaled based on the size and localization of the nose. Because little information is needed and there are no refinement steps, our method is able to find suitable landmarks even in challenging poses. While not taking into account facial expressions and specific facial traits, our algorithm achieved competitive scores on the 3D Face Alignment in the Wild (3DFAW) challenge. The obtained results have the potential to be used as rough estimation of the position of the 3D face landmarks in the wild images, which can be further refined by specially designed algorithms.

Keywords: 3D face alignment · Head pose estimation · Faces in the wild

1 Introduction

Face alignment is defined as determining the position of a set of known facial points across different subjects, illuminations, expressions and poses [4]. 3D face alignment in the wild is defined as determining the position of these landmarks in the 3D space given only a 2D image acquired in unconstrained environments. This information can be used for several computer vision applications, such as face recognition [11], pose estimation [7], face tracking [14,15], 3D face reconstruction [1] and expression transfer [12,13].

Recent face alignment work can be subdivided into 2D and 3D methods. Zhu and Ramanan [19] use mixtures of trees with a shared pool of parts for sparsely aligning faces even in profile head poses, successfully calculating the position of the 2D landmarks. Ren *et al.* [8] uses regression local binary features to perform 2D sparse face landmark estimation at 3000 frames per second. Jeni *et al.* [4] is able achieve state-of-the-art real-time 3D dense face alignment by fitting a 3D model on images acquired in controlled environments. The use of cascaded

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 581–589, 2016.
DOI: 10.1007/978-3-319-48881-3_40

coupled-regressors, by integrating a 3D point distribution model was proposed by Jourabloo and Liu [5] for estimating sparse 3D face landmarks in extreme poses.

In this paper, we present our entry for the 3D Facial Alignment in the Wild (3DFAW) challenge. Our approach is landmark-free in the sense that it does not need any specific face information, only a detected nose region that is used to estimate the head pose. A generic face landmark model is rotated based on the head pose, translated and scaled to fit the detected nose. We choose to use the nose as basis of our work as it has been shown efficient for head pose estimation, does not deform easily when facial expressions are present, is not easily occluded by accessories and is visible even in extreme profile head poses [17]. Our method does not make use of any facial trait specific to the subject and does not take facial expression into account, yet it achieves competitive results. Our approach works well as an initial estimation for the position of the landmarks, since it only needs the nose region, which can be easily obtained with existing detection methods even in challenging environments.

The 3DFAW challenge presents a set of images and annotations for evaluating the performance of in the wild 3D sparse face alignment methods. Part of the data is from the MultiPIE dataset [2] or from images and videos collected on the internet, having its depth information been recovered through a dense 3D from 2D videos alignment method [4]. The rest of the data was synthetically generated by rendering the 3D models present in the BU-4DFE [16] and BP4D-Spontaneous [18] databases onto different backgrounds. The training data includes the face bounding box and the 3D coordinates of 66 facial landmarks, while the testing data only includes the face bounding box.

The results obtained on the challenge's dataset are evaluated using two different metrics: Ground Truth Error (GTE) (Eq. 1) and Cross View Ground Truth Consistency Error (CVGTCE) (Eq. 2), such that X is the prediction, Y is the ground-truth, d_i is the Euclidean distance between the corner of the eyes for the i-th image [10] and P is obtained using Eq. 3.

$$E(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\frac{\|x_k - y_k\|_2}{d_i} \tag{1}$$

$$E_{vc}(X,Y,P) = \frac{1}{N}\sum_{k=1}^{N}\frac{\|(sRx_k + t) - y_k\|_2}{d_i} \tag{2}$$

$$P = \{s,R,t\} = \underset{s,R,t}{\operatorname{argmin}}\sum_{k=1}^{N}\|y_k - (sRx_k + t)\|_2^2 \tag{3}$$

This paper is structured as follows: Sect. 2 explains our method in detail, with attention to each step; Sect. 3 presents and explains our results on the 3DFAW challenge; and Sect. 4 includes final remarks.

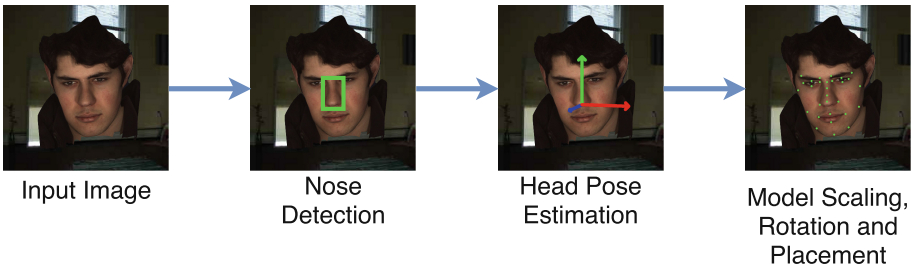


Fig. 1. Overview of our method

2 Our Approach

Our method is composed of seven steps, four offline and three online: 1. A nose detector is trained; 2. An average face model is generated to be used as template and for calibrating the pose; 3. Ground-truth head poses are extracted; 4. A head pose estimator is trained; 5. Nose detection is performed; 6. The detected region is used for estimating the head pose; 7. The face model is adjusted to the pose and fitted using the nose for alignment. The simplicity of the online steps is outlined in Fig. 1.

2.1 Landmark Model and Head Pose Ground-Truth Generation

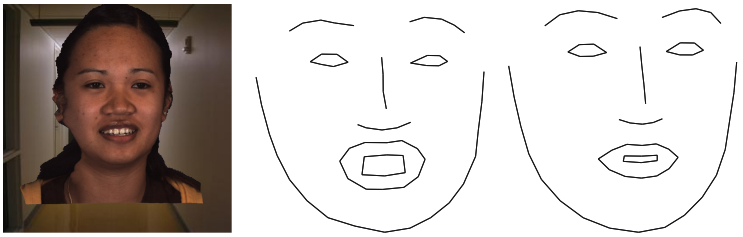
A near frontal image (0° head yaw, pitch and roll) from the training subset was chosen to be used for calibrating all other images (Fig. 2(a)), it is defined as not having any rotation on any of the three axes.

For generating the ground-truth head pose, an affine transformation is applied to the landmarks belonging to all images in the training subsets relative to the landmarks in the calibration image (Fig. 2(b)), such that a transformation matrix including translation, scale and rotation is generated. The Euler angles are extracted from this matrix and defined as the ground-truth head yaw, pitch and roll.

The landmark model (Fig. 2(c)) is generated by applying the aforementioned transformation to the training data, normalizing the scale based on the distance of landmarks number 32 and 36 on the base of the nose and averaging the position of each landmark. The resulting model roughly represents the average face with a neutral expression in the dataset.

2.2 Nose Detection

Ground-truth nose regions are extracted by cropping the training images around the nose landmarks. Faster R-CNN [9] is trained with all training images in the 3DFAW dataset and used for detection. The Faster R-CNN introduces the novel concept and use of a Region Proposal Network, that generates both candidate bounding-boxes and detection confidence scores.



(a) (b) (c)

Fig. 2. (a) Calibration image; (b) Calibration landmarks; (c) Landmark model viewed with the calibration pose

When processing the testing images, the detected region with the highest confidence score is selected. If Faster R-CNN yields no candidates, a region at the center of the image is selected.

2.3 Head Pose Estimation

For performing head pose estimation, the CNN variant of NosePose [17] is applied. It uses a network similar to those crafted for face recognition [3,6], but modified to be trained with a smaller number of images and to support the smaller nose regions.

Training was performed using all images in all training subsets. The ground-truth head pose was discretized for both yaw and pitch in steps of 7.5°, the yaw ranges from −60 to 60 and the pitch, from −52.5 to 52.5°, relative to the calibration image. The roll is not estimated as only small variations are present in the dataset. These values were all empirically determined and a single network is trained for estimating both the yaw and the pitch simultaneously.

2.4 Model Fitting

Given the detected nose region, an optional face bounding box and the estimated head pose, the landmark model, explained in Sect. 2.1, is fitted on the face according to Algorithm 1. While having the face bounding box allows the method to perform a slightly more precise scaling of the model, it is not required as it is possible to infer the size of the face from the size of the nose.

Figure 3 contains examples of this process, including the detected nose region and face bounding-box. If the face bounding box is given, when scaling the model, three constants for minimizing the error are used, one for each axis. The best position used for aligning the model with the nose region and the factor used when scaling according to it were also determined in a similar fashion.

3 Experimental Results

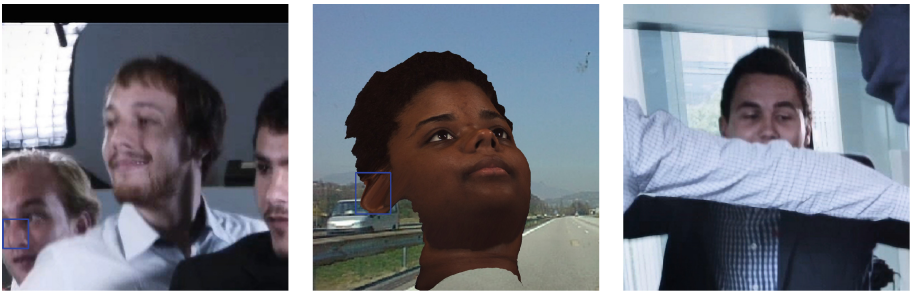
In order to assess the performance of our nose detection step, manual verification of the results was performed on all 4,912 images in the testing subset. Only

Algorithm 1. Model Fitting Algorithm

function FITMODEL($model, noseBB, headPose, faceBB$)
  $modelNoseBase \leftarrow average(model.noseBaseLeft, model.noseBaseRight)$
  $rotate(model, modelNoseBase, headPose)$
  if $isDefined(faceBB)$ **then**
    $xScale \leftarrow .975 * faceBB.width/model.width$
    $yScale \leftarrow .975 * faceBB.height/model.height$
    $zScale \leftarrow (xScale + yScale)/2 * .95$
    $scale(model, xScale, yScale, zScale)$
  else
    $modelNoseWidth \leftarrow l2Norm(model.noseBaseLeft, model.noseBaseRight)$
    $scale(model, .6 * nose.width/modelNoseWidth)$
  end if
  $noseBase \leftarrow \{nose.x + nose.width * .5, nose.y + nose.height * .9, 0\}$
  $translate(model, modelNoseBase - noseBase)$
  $translate(model, \{0, 0, -average(model).z\})$
  return model
end function

Fig. 3. Example results of the model fitting stage, showing the face bounding-box in red, the detected nose region in blue and the estimated position of the landmarks in green. (a) Near frontal good fit; (b) Bad fit caused by bad head pitch estimation; (c) and (d) Half-profile good fit; (e) Good fit in an image sourced from the MultiPIE dataset; (f) Modest fit in one of the most challenging images in the dataset (Color figure online)

nine images failed to yield detections, however, the wrong region was detected in four images and the nose belonging to the wrong subject was detected in one. The detection was accurate in 99.71 % of the images. This high rate is expected as a state-of-the-art detection method was used and all images are high resolution with very few of them including blur and variations in lighting. Figure 4 illustrates three cases where the detector failed.



(a) (b) (c)

Fig. 4. Images where the nose detection (blue box) failed: (a) Wrong nose detected; (b) False positive; (c) No detection (Color figure online)

Isolating and assessing the head pose estimation performance is not possible as the ground-truth landmarks were not made available and visually verifying the correctness of the head pose is a difficult problem for humans [17]. Its performance, however, reflects on the final landmark estimation score, which was calculated by the challenge's web system. Our method achieved 5.9093 CVGTCE and 10.8001 GTE when scaling the model according to the size of the face bounding box.

Due to the limited number of allowed submissions to the web system and not being able to locally assess our method's performance, we do not present a quantitative evaluation when the size of the nose region is used to infer the size of the face. However, we performed visual inspection of the results and concluded they are consistent with those obtained using the face bounding box. Figure 5 presents examples of our results using both scaling approaches, for comparison.

4 Final Remarks

We presented a nose region based approach for in the wild 3D landmark estimation, our entry for the 3D Face Alignment in the Wild Challenge. A generic face landmark model is generated using the information present in the training subset of the challenge. A high-performance, state-of-the-art nose detector and head pose estimator are trained using the nose regions extracted from the landmark annotations. The detected nose is used for estimating the head pose and

Fig. 5. Visual comparison between the two different model scaling methods. The results obtained using the detected nose region are on the left and the ones obtained with the face bounding box are on the right. (a) and (b) the model fitted with the nose is noticeably larger; (c) and (d) using the nose yielded better results

projecting the rotated landmark model according to the size of the face, either inferred from the size of the nose or from the available face bounding-box. Competitive results were achieved while taking neither facial expressions nor facial traits into account. Because only the nose region is needed for performing our estimation, it has the potential to be used for calculating useful initial landmark positions that can be refined for finer face alignment.

Acknowledgment. The authors would like to thank CNPq and CAPES for supporting this research.

References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999), pp. 187–194. ACM Press/Addison-Wesley Publishing Co., New York (1999). http://dx.doi.org/10.1145/311535.311556
2. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. In: 8th IEEE International Conference on Automatic Face Gesture Recognition (FG 2008), pp. 1–8 (2008)
3. Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S., Hospedales, T.: When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 142–150 (2015)
4. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3D face alignment from 2D video for real-time use. In: Image and Vision Computing (2016). http://www.sciencedirect.com/science/article/pii/S0262885616300877
5. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
7. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 607–626 (2009)
8. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment via regressing local binary features. IEEE Trans. Image Process. **25**(3), 1233–1245 (2016)
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 91–99. Curran Associates, Inc., Red Hook (2015). http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf
10. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. Image Vis. Comput. 47, 3–18 (2016), 300-W, The First Automatic Facial Landmark Detection in-the-Wild Challenge. http://www.sciencedirect.com/science/article/pii/S0262885616000147
11. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
12. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. (TOG) **34**(6), 183 (2015)
13. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of Computer Vision and Pattern Recognition (CVPR). IEEE (2016)
14. la Torre, F.D., Chu, W.S., Xiong, X., Vicente, F., Ding, X., Cohn, J.: Intraface. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8 (2015)
15. Yang, J., Deng, J., Zhang, K., Liu, Q.: Facial shape tracking via spatio-temporal cascade shape regression. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 994–1002 (2015)

16. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face Gesture Recognition (FG 2008), pp. 1–6 (2008)
17. Zavan, F.H.B., Nascimento, A.C.P., Bellon, O.R.P., Silva, L.: Nosepose: a competitive, landmark-free methodology for head pose estimation in-the-wild (2016)
18. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: BP-4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. Image Vis. Comput. 32(10), 692–706 (2014), Best of Automatic Face and Gesture Recognition 2013. http://www.sciencedirect.com/science/article/pii/S0262885614001012
19. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (2012)