EURASIP Journal on Image and Video Processing

REVIEW Open Access

Evaluating effects of focal length and viewing angle in a comparison of recent face landmark and alignment methods

Xiang Li[1*], Jianzheng Liu[2], Jessica Baron[1], Khoa Luu[3] and Eric Patterson[1]

*Correspondence:
xiang5@clemson.edu
[1]School of Computing, Clemson University, 304 McAdams Hall, 29630 Clemson, SC, USA
Full list of author information is available at the end of the article

Abstract

Recent attention to facial alignment and landmark detection methods, particularly with application of deep convolutional neural networks, have yielded notable improvements. Neither these neural-network nor more traditional methods, though, have been tested directly regarding performance differences due to camera-lens focal length nor camera viewing angle of subjects systematically across the viewing hemisphere. This work uses photo-realistic, synthesized facial images with varying parameters and corresponding ground-truth landmarks to enable comparison of alignment and landmark detection techniques relative to general performance, performance across focal length, and performance across viewing angle. Recently published high-performing methods along with traditional techniques are compared in regards to these aspects.

Keywords: Facial alignment and landmarking, Convolutional neural networks, Focal length, View angle, Comparison, Evaluation, Review

1 Introduction

Face detection, tracking, and recognition continue to be employed in a variety of ever more common-place biometric applications, particularly with recent integrations in mobile-device security and communication. Most of these applications, such as identity verification, pose tracking, expression analysis, and age or gender estimation, make use of landmark points around facial components. Correctly locating these key points is crucial as they often are used to abstract main features such as the jaw, eye-brows, eyes, nose shape, nostrils, and mouth [1]. Due to the complexity of head gestures , automatic localizing of canonical landmarks usually first involves face alignment to account for rotation, translation, and scale due to pose or view-direction differences [2–5]. Furthermore, 2D images photographically captured by cameras are affected by perspective and lens distortion, an important aspect considered in this work.

This review aims to compare performance of five notable facial landmark and alignment methods under the effects of different camera focal lengths and positions, particularly under conditions that have been ignored or difficult to test. Previously, Çeliktutan et

al. completed a thorough survey of facial landmark detection algorithms and comparative performance in 2013, which at the time primarily focused on 2D techniques such as Active Shape Model (ASM) and Active Appearance Model (AAM) variations [6]. In 2018, Johnston and Chazal published work that built on the earlier survey, noting the shift of interest to deep-learning methods due to potential performance increases as well as techniques that also perform 3D alignment [7]. Several strong-performing neural-network methods have been published since; however, and in general, no performance comparisons have included lens-perspective effects nor systematic evaluation across the range of viewing angles. This study is not an exhaustive survey of recent methods but rather an investigation in the effects of focal length and viewing angle on both traditional and more recent neural methods (published after the 2018 article). Focal-length-based perspective and viewing angle are both important considerations if designing a biometric or other system in order to account for the lens chosen, viewing angle, and proximity necessary for the system.

The effects a lens imparts on acquisition have often been ignored in face-related research. A fundamental technique in computer vision is estimating a camera projection matrix and has been regarded in many studies; however, the datasets used to train and test landmark detection do not usually include camera meta-data (particularly large datasets gleaned from the Internet for deep-learning approaches), or datasets have been captured in very controlled situations with a single lens. The most widely used databases in training recent deep networks are 300W [8], COFW [3], WFLW [9], and AFLW [10]. Those cover large variation over age, ethnicity, skin color, expression, and pose and have been used by top-performing deep neural networks [11–15]. None of them explicitly note focal-length as a parameter. In short, there is no dataset published online that has considered focal length/field of view versus proximity for training alignment or landmark detection methods. We assume perspective distortions caused by focal length will likely affect the final annotation results. If so, training sets including camera and lens parameters could increase accuracy of a system or at least aid in designing systems.

A few researchers have considered aspects of image distortions relative to face images for particular applications, but not what we present here. Damer et al. investigated state-of-the-art deep neural networks for facial landmark detection, but their main focus was perspective distortion due to distances between cameras and captured faces and did not consider the effects due to lenses and associated field of view [16]. Valente et al. investigated basic lens effects; however, they only analyzed these relative to simple mathematical algorithms for facial recognition (EIGENDETECT and SRC) and not those for facial alignment nor their effects on facial landmark detection [17]. Flores et al. also focused on perspective distortion caused by distance [18]. They estimated camera pose from facial images using Efficient Perspective n-Point(EPnP) rather than evaluating landmark location. In this work, we consider the effects of lens focal length and viewing angle in regards to some of the highest performing recent facial-landmarking techniques. Although the method of evaluation uses synthetic images, the question of performance relative to lens and viewing angle is also relative, and the goal is to demonstrate that all methods are affected to varying degree. Studying such effects without large datasets that include camera and lens meta-data would not currently be possible without either collecting such a dataset or creating test images synthetically as we have done. Future work could include design of a dataset, although it could be prohibitive to collect data on the

size order of Internet-driven datasets used for deep-learning training. Furture work also could consider improving synthetically rendered images for higher fidelity, style-transfer, or in-environment placement, etc.

The contributions of this work include evaluation of five different facial landmark detection methods in regards to varying lens choice and viewing angle. Three of them are from recently published deep-learning 3D facial annotating methods, and the remaining two are AAM implementations. We evaluate the performance of these methods across view angles and focal lengths by using face images synthesized from detailed 3D scans of individuals.We demonstrate that all are subject to particular performance degradation with lens-perspective distortion and viewing angle. This information may be used to guide design choices in biometric or other imaging systems as well as develop on methods that are more robust to lens choice and angle.

2 Method of comparison

2.1 Landmark schemes

There have been a variety of landmark schemes used in related projects, but a few have been most used in recent work and make a logical choice for comparative evaluation. Following the categories in [6], there are two major groups of facial landmarks schemes: primary landmarks and secondary landmarks. Primary landmarks usually define the eye corners, the mouth corners, and the nose tip. Those landmarks are located at "T" sections between boundaries or at high curvatures on a face which may be detected by image processing algorithms, e.g., multi-resolution shape models [19], Harris Corner Detection model [20], or Image Gradient Orientation (IGO) model [21]. Secondary landmarks outline the contour of main features that are guided by primary landmarks, such as the jaw line, eyebrows, and nostrils. Wu et al. [1] provide a thorough survey on facial landmark databases and their corresponding landmark schemes. A common 68-point landmark is supported by many face databases, e.g., AFLW [10], BU-4DFE [22], Helen [8, 23], etc. For easiest consistency, the 68-point scheme from Multi-PIE [24], and further popularized by iBUG's 300W [8], was chosen for this study.

Sagonas et al. and Johnston et al. [7, 8] state that primary landmarks are more easily detected than secondary landmarks while annotating the ground-truth reference. The "m7 landmarks" including the 4 eye corners, 1 nose tip, and 2 mouth corners are also included here in some comparisons with the idea that they provide higher importance information. Figure 1 shows the two landmarks schemes used in this paper.

In order to generate face images at controlled focal lengths and precise angle selections, we synthesized photo-realistic images using detailed 3D meshes captured from a structured-light 3dMD system. Our facial capture participants were asked to make different expressions following the Facial Action Coding System (FACS). FACS was created by the anatomist Carl-Herman Hjortsjö [25] and further developed by Ekman etc. [26] It provides a coding system which describes how to categorize facial expressions into Action Units (AUs) with muscle movements. We manually annotated the ground-truth landmarks in 3D for 84 faces from our participants, 64 from a set of FACS-capture expressions of two individuals and 20 of unique individuals with a range of ethnicity, age, and gender where the pose was neutral or a slight smile. Figure 2 shows an example of FACS and neutral faces in our dataset. Landmark variation often occurs between in datasets, particularly for areas such as the jawline or eyebrows. For consistency, we keep jawline

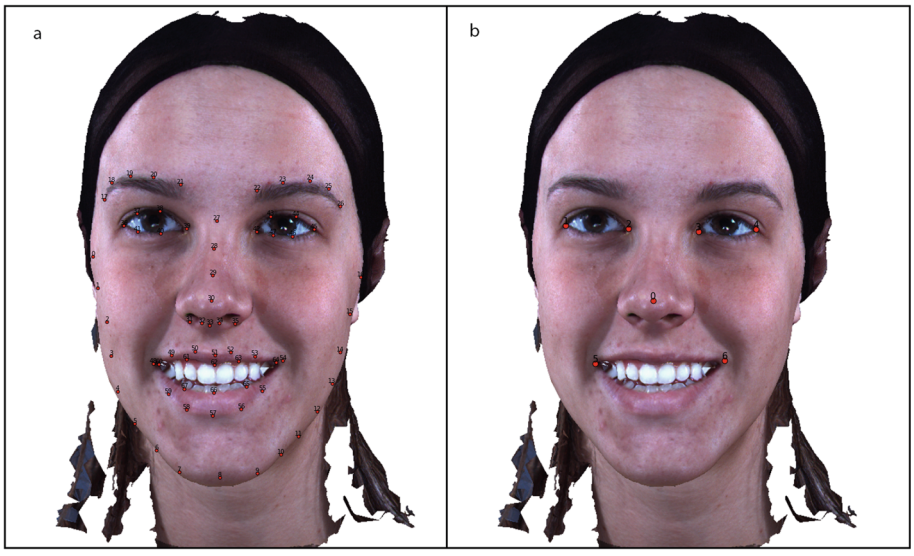

Fig. 1 Two different landmarks schemes. **a** iBUG-68 . **b** M7

points evenly distributed along the chin. In some projects, eyebrow points are placed at the center, bottom, or top of brow arcs. Good choices for landmarks points include those near high curvature or boundaries on objects. Here, eyebrows are marked anatomically at the supraorbital ridge or eyebrow ridge.

2.2 Evaluation metrics

We use ground-truth based localization error to evaluate performance in each case via root mean squared error (RMSE). Accurate landmarks are generated for each synthetic image by projecting manual 3D landmarks to match the rendered angle and field of view. We use the method proposed by Johnston et al. [7] for calculating the RMSE:

Fig. 2 FACS and NFACS faces. **a** It is a lip pucker expression and its corresponding action unit is 18 [26]. **b** The participant provides a neutral expression

$$\text{RMSE} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\left(x^k - \tilde{x}^k\right)^2 + \left(y^k - \tilde{y}^k\right)^2} \tag{1}$$

where x^k, y^k denote each of the K predicted landmark k in an image, and $\tilde{x}^k$, $\tilde{y}^k$ indicate the corresponding ground-truth landmark. Normalizing for face size in pixels is useful due to the variance across images. Previously, RMSE is normalized by the ground-truth outer corners of the left eye and right eye landmarks(Eq. 3)[8]. The error per landmark in image i is given as:

$$\epsilon_i^k = \frac{\sqrt{\left(x_i^k - \tilde{x}_i^k\right)^2 + \left(y_i^k - \tilde{y}_i^k\right)^2}}{{d_{\text{norm}}}^i} \tag{2}$$

$${d_{\text{norm}}}^i = \sqrt{\left(\tilde{x}_{le} - \tilde{x}_{re}\right)^2 + \left(\tilde{y}_{le} - \tilde{y}_{re}\right)^2} \tag{3}$$

where $(\tilde{x}_{le}, \tilde{y}_{le})$ and $(\tilde{x}_{re}, \tilde{y}_{re})$ are the ground-truth outer corners of the left eye and right eye in the image i. In our case, however, our synthetic images vary with camera positions. The distances of outer-eye corners may have small impacts at side angles due to perspective projection. Hence, we calculate Normalized Root Mean Squared Error (NRMSE) by normalizing per width of the head bounding box. We calculate the percentage of accepted points among all points to show the performance for each algorithm:

$$P(k) = 100\frac{1}{I}\sum_{i=1}^{I}\left[i : \epsilon_i^k < Th\right] \tag{4}$$

where $[\, i : \epsilon_i^k < Th]$ is a mask function that if the normalized distance ϵ is less than Th, it is acceptable, and i is set to 1. Otherwise, the result is not acceptable, and i's value is set to 0. So, the overall performance over K landmarks in each image for I image set is:

$$P = 100\frac{1}{K \times I}\sum_{k=1}^{K}\sum_{i=1}^{I}\left[i : \epsilon_i^k < Th\right] \tag{5}$$

2.3 Camera position and focal length

Our coordinate system follows the typical computer-graphics right-handed coordinate system convention, where the X-axis points to horizontal right, Y-axis points to vertical up, and Z-axis perpendicular to both X and Y points outward from the screen. In order to track the camera around each face, we use spherical coordinates to represent camera positions. Our interests are analyzing multiple viewing angles at a wide range of specific viewing angles. We define camera positions in spherical coordinates at (r, ϕ, θ), where ϕ is the polar angle (also known as zenith angle) from the positive Y-axis with $45° \leq \phi \leq 135°$, at 15° each. We define θ to be the azimuthal angle in the xy-plane from the positive X-axis with $180° \leq \theta \leq 0°$ at intervals 30°. Lastly, r varies for simulated focal length. Overall, we have 49 camera positions so that various front views of the face and some extreme camera positions could be tested. Figure 3 shows the position of spherical coordinates and samples of face images with different viewing angles.

Focal length, relative to the dimensions of the film or digital sensor, determines the field of view on a physical camera, and there are also radial distortion issues relative to physical lenses and typical of certain optical designs such as pincushion and barrel distortion (these are not specifically included here but could warrant a follow-up study). In photography, a common standard of comparison of focal length to express field of view is relative to the standard of the 35-mm-film frame size used for much of the twentieth century and

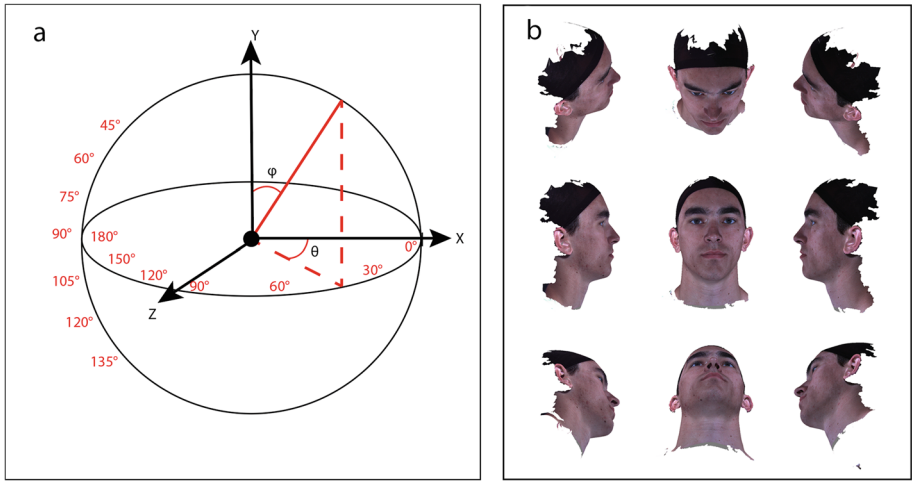

Fig. 3 Camera positions. **a** Right-handed system spherical coordinates. ϕ is the polar angle and θ is the azimuthal angle. **b** Samples for real face images taken from different viewing angles

carried forward into digital sensors. This "35-mm" frame size of 36 mm across by 24 mm down came to be a standard for still photography when Oskar Barnack doubled the individual frame from motion-picture film (standardized by Thomas Edison) to use in still cameras. The relation between angle of view and focal length is given by:

$$\alpha = 2 \arctan \frac{d}{2f} \tag{6}$$

where α is the angle of view, d denotes the size of film, and f is the focal length t. As can be seen by the relationship, shorter focal lengths widen the field of view and vice-versa. To maintain a face of a relative size in images captured with different focal lengths, the distance to the camera needs to be changed. Perspective effects are modified as this occurs, as can be noted in Fig. 4. Short focal lengths (wide-angle lenses) introduce a fair amount of facial distortion whereas longer lengths begin to approximate an orthographic projection that maintains relative distances among landmarks better. Although not tested here, these effects can be more pronounced near the edges of a capture frame. As mobile phone photography increases, some of the most common focal lengths relative to the standard of comparison noted would equate to the 28-mm to 35-mm range of focal lengths, or a relatively wide field of view. Interchangeable lens cameras or cameras with zoom lenses can vary the focal length. As can be seen from formula 6, a larger focal length lens has a narrower angle of view at the same camera-to-object distance which offers magnified, detailed photos. Focal lengths greater than 50 mm are often used in longer range photography, long range biometric acquisition, and especially in head-and-shoulder portrait photography. For this study, common focal lengths of prime lenses used in still photography were chosen as the range, from 24 mm (wide-angle on a 35-mm system) to 135-mm (slight telephoto on a 35-mm system), with the range covering typical focal lengths used in photography and not including extreme wide-angle lenses nor extreme telephoto lenses. We choose six different types of common lens focal lengths (24 mm, 28 mm, 35 mm, 50 mm, 85 mm, and 135 mm) as our test domains for comparison.

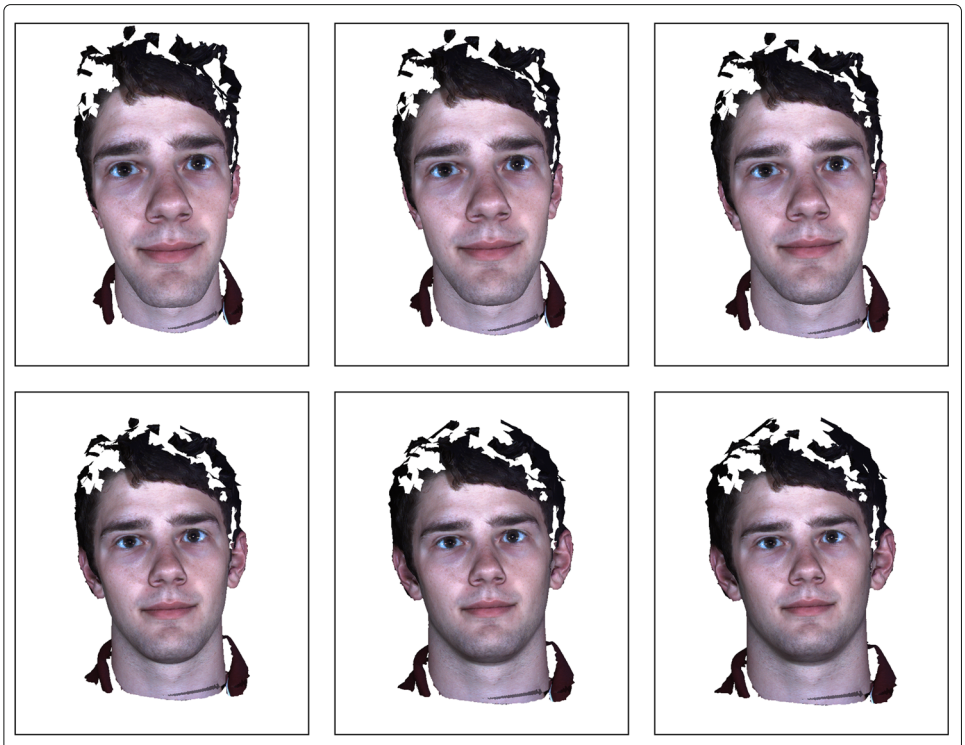

Fig. 4 Camera focal lengths: First row from left to right: 24 mm, 28 mm, and 35 mm; second row from left to right: 50 mm, 85 mm, and 135 mm

2.4 Face landmark and alignment methods

Wu et al.[1] mention classifying technology as holistic methods, constrained local model methods, and regression-based methods. Holistic methods treat a whole face image as the entire appearance and shape to train models. Constrained local models locate landmarks based on the global face but emphasizing local features around landmarks. Regression-based methods mostly are adopted for deep-learning, using regression analysis to map landmarks to images directly. Johnston et al. [7] believe that facial landmark detection methods can be divided into generative methods, discriminate methods, and statistical methods. Generative methods minimize the error between models and facial reconstructions. Discriminate methods use a dataset to train the regression models. Statistical models are a combination of generative methods and discriminate methods. Çeliktutan et al. classify facial landmark detection into model-based (using the entire face region) and texture-based (matching landmarks to local features) [6]. Here we consider landmarking algorithms based on either statistical methods or deep-learning methods. Statistical methods calculate the positions of landmarks using mathematical algorithms. Most of the traditional methods(e.g., AAM and ASM) can fall into this group. Deep learning methods feed facial images to train deep neural networks to locate landmarks.

ASM and AAM models have performed among some of the best landmark-detection algorithms for nearly two decades. ASM, first introduced by Cootes et al., attempts to detect and measure the expected shape of a target in an image. ASM requires a set of landmarked images for training the model. The first step is using Procrustes Analysis to align all object images. A mean shape is calculated by Principle Component Analysis

(PCA) which applied to find eigen vectors and eigen values [27]. All the objects' shapes can be approximated as:

$$x = \bar{x} + Pb \tag{7}$$

where $\bar{x}$ is the mean shape calculated over all overall training data. P is a set of eigen vectors derived from the covariance matrix calculated via PCA, and b is a set of shape parameters given by:

$$b = P^T (x - \bar{x}) \tag{8}$$

As an improvement of ASM, an active appearance model matches both shape and texture simultaneously and gives an optimal parameterized model. PCA is also applied for texture and once again for finding combined appearance parameters and vectors. Menpo provides five different AAM versions with two main groups: Holistic AAM (HAAM) and Patch AAM (PAAM) [28]. HAAM warps appearance information using a nonlinear function, such as Thin Plate Spline (TPS), and takes the whole texture into account when fitting, while the PAAM uses rectangular patches around each landmark as texture appearance. We test both HAAM and PAAM as separate techniques for comparison here. For building the AAM, we chose the widely used Helen Dataset which provides a high-resolution set of annotated facial images containing different ethnicities, ages, genders, head poses, facial expressions, and skin colors, similarly used by Johnston et al. [7]. In order to reduce error caused by facial detection, we extract faces from image using bounding boxes calculated from ground-truth landmarks and dilated by 5%.

In the past few years, deep-learning based neural-network methods have leveraged very large datasets for training and recently outperformed statistical shape and appearance models in many areas. We gathered three recent high-performing methods where implementations were available to compare in our various cases. The first method is called the Position Map Regression Network [29]. The main idea of PRNet is creating a 2D UV Position Map which contains the shape of an entire face to predict 3D positions. PRNet employs a convolutional neural network (CNN) trained 2D images along with ground truth 3D dense position clouds created via 3D morphable model (3DMM). 3D positions are projected to the UV texture-map format and used in training the CNN. The UV texture map preserves 3D information, even posed with occlusions.

The second method is the 3D Face Alignment Network (3D-FAN). Bulat and Tzimiropoulos use a 2D-to-3D Face Alignment Network combined with a stacked heat-map sub-network to predict Z coordinates along with 2D landmarks[30].

The third method from Bahagavatula et al. uses a 3D Spatial Transformer Network (3DSTN) to estimate a camera projection matrix in order to reconstruct 3D facial geometry. The method forms occluded faces with 2D landmark regression and predicts 3D landmark locations[31].

These methods were trained on 300W-LP except for 3D-FAN which was trained on the 230,000 + 300W-LP. It would be prohibitive to attempt to include all recent deep-learning methods in this comparison, but these were chosen based on strong performance in recent publications, and we believe other recent methods would very likely perform similarly based on similar overall performance on the same datasets.

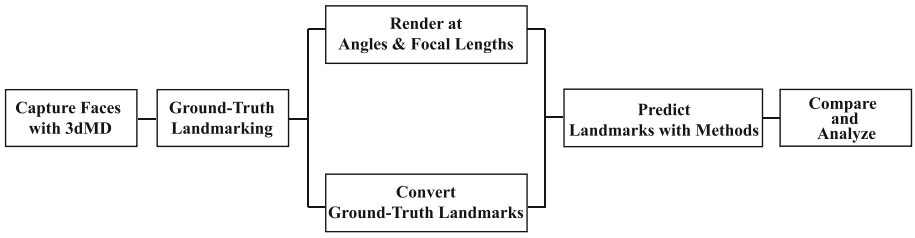

Fig. 5 Flow chart of the main process for measurement

3 Procedure

Figure 5 illustrates the main work flow of our approach to evaluate facial landmark and alignment algorithms.

To calculate the RMSE (1) and NRMSE (2), (3) on landmarks, all measurements require ground-truth as references. All facial meshes with texture were manually marked using landmarker.io to create these ground-truth landmarks. Figure 6 shows an example of 3D facial annotation in landmarker.io as performed on our dataset [28].

Using our own Python-, Qt-, and OpenGL-based lab application, Countenance Tool, we render 3D facial positions given varying angles and focal lengths. Since we compare how view angles and focal lengths affect landmark methods, we move the virtual camera to 49 different locations shown in Fig. 3. At each location, we rasterize faces with 6 different synthesized focal lengths (24 mm, 28 mm, 35 mm, 50 mm, 85 mm, and 135 mm) by changing the focal length parameter shown in equation 6 before rendering. Overall, there are images at 49 angles and 6 focal lengths for each face. At the same time, we use the same camera matrices (varying with view of angles and focal length parameters) to project the 3D ground-truth landmarks to yield the ground-truth 2D landmarks at image coordinates. Figure 7 shows a set of images with ground-truth landmarks of different focal lengths and viewing angles.

To summarize the workflow demonstrated in Fig. 5, we first performed facial geometry capture with a 3dMD system. The 3dMD system provided 3D meshes along with texture information. We then imported those into landmarker.io to annotate each face manually to generate 3D ground-truth landmarks. After getting the ground-truth, we rasterized each face at 49 angles and 6 focal lengths and calculated the ground-truth 2D landmark locations. Finally, we analyzed performance of each method by calculating NRMSE error between a method's predicted landmarks and the 2D ground-truth locations.

Fig. 6 3D ground-truth landmarking using landmarker.io

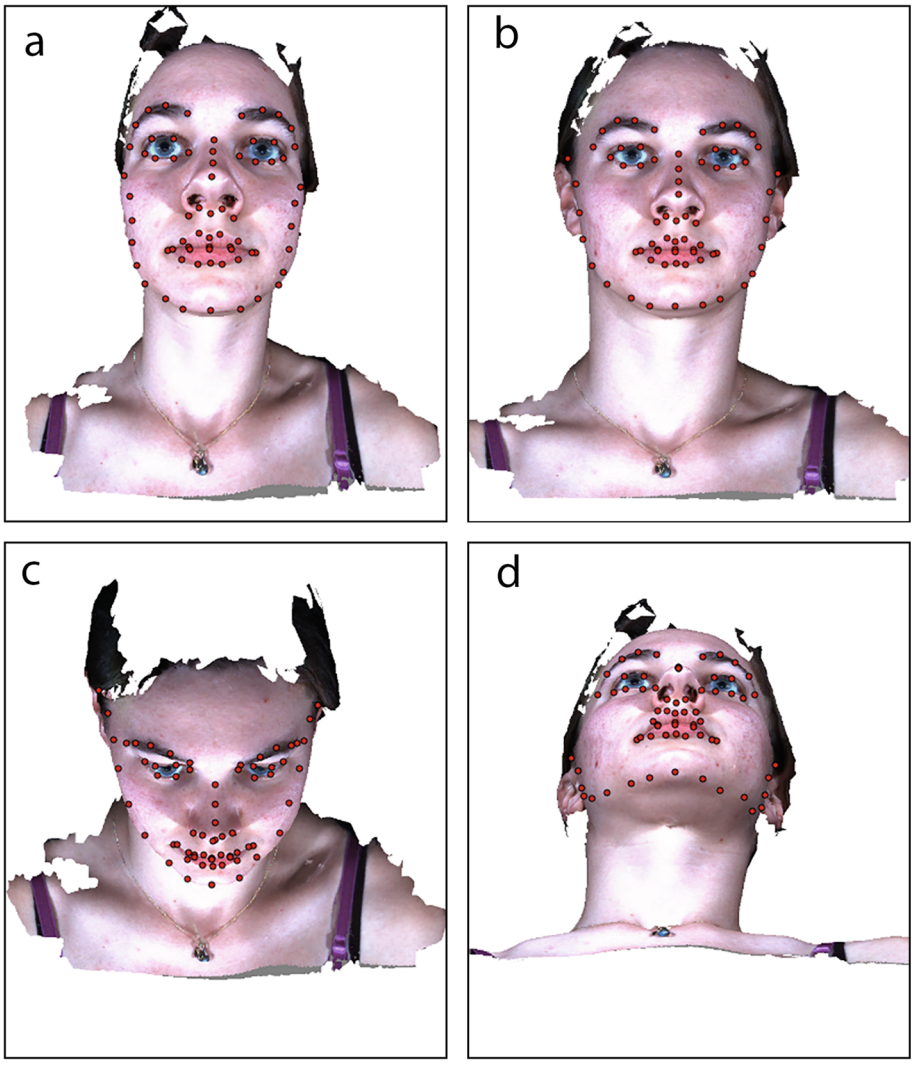

Fig. 7 Converted 2D ground-truth on images. **a** and **b** are ground-truth landmarks with 24 mm and 135 mm focal length at the center view, respectively. **c** and **d** are rendered landmarks from top view and down view with 135 mm focal length

4 Results and discussion

In this section, we compare the RMSE performance of the five methods with the full 68-points scheme and the reduced m7 scheme against 6 threshold levels. Figure 8 plots the percentage correctly accepted for each facial landmark and alignment method with both schemes. Generally speaking, as expected, the overall acceptance performance for each algorithm increases as the threshold widens. The m7 landmarking scheme tends to show better performance as a smaller set located at distinct "corners." In general, the CNN methods perform better, but all are still subject to performance effects due to focal lengths and viewing angles. It would be remiss to declare one method particularly better than another here, particularly since 3D-FAN was trained on an augmented dataset versus the others; we used the publicly available pre-trained networks. Compared to the neural-network techniques, the performance of traditional statistical methods is typically lower. As Cootes explains [32], the performance of ASM and AAM is dependent on the starting

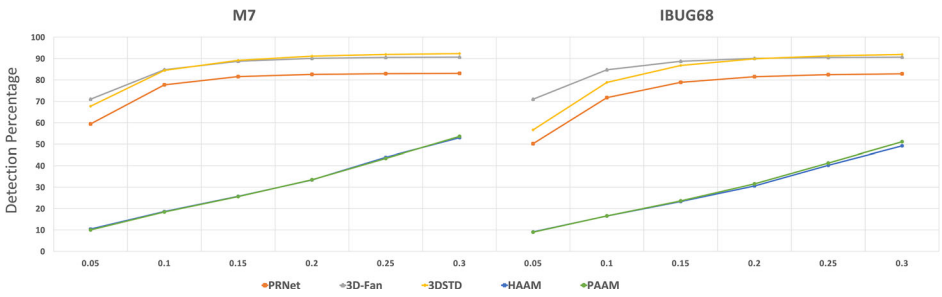

Fig. 8 Results with M7 (left) and iBUG (right) landmarks are both plotted for comparison. Performance respect to thresholds using 68 pts

position of landmark displacement. Higher accuracy of face detection tends to improve landmark detection.

One of the main contributions of this paper is demonstrating the effect of focal length on landmarking accuracy. Figure 9 demonstrates lower performance with a wider field-of-view, associated with strong perspective effects, and better performance as focal length increases. There is expected leveling in performance with focal length increase.

In order to visualize effects on specific landmarks at different focal lengths, we drew the 68-point landmarks located by each method and the average of the frontal view for the extremes (135 mm lens in blue circles based on RMSE and 24mm in red). This shows which landmarks are most affected by the focal-length perspective warping. Figure 10 also reflects the data depicted in the Fig. 9. The radius of the RMSE presents how far each predicted landmark is from the ground-truth. The result shows that all of the landmarks that are close to the center of faces have more accurate predictions, while landmarks along facial edges have lower accuracy predictions due to projective distortions; particularly, corners of eyes and lips seem affected.

The last consideration for this paper is systematic adjustment of the camera's viewing angle across the viewing hemisphere. We place the camera at 49 different positions with extreme poses included. When the camera views from the center ($\theta \cong 90°$ and $\phi \cong 90°$), the performance results are better than when the camera view from the sides. The landmark predictions at ϕ around 45° and 135° have the lowest performances due to extreme viewing angles. As expected, performance drops as the view moves to the more extreme angles, and the rate of effect for each method are shown in Figs. 11, 12, 13, 14, and 15.

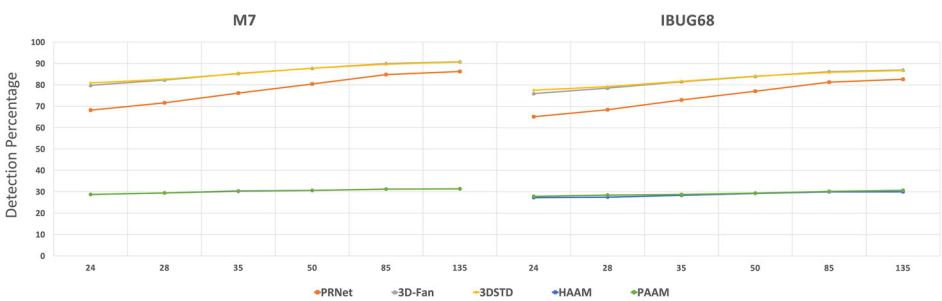

Fig. 9 Focal Length Varying: Left half of the chart uses M7, right half uses 68 points

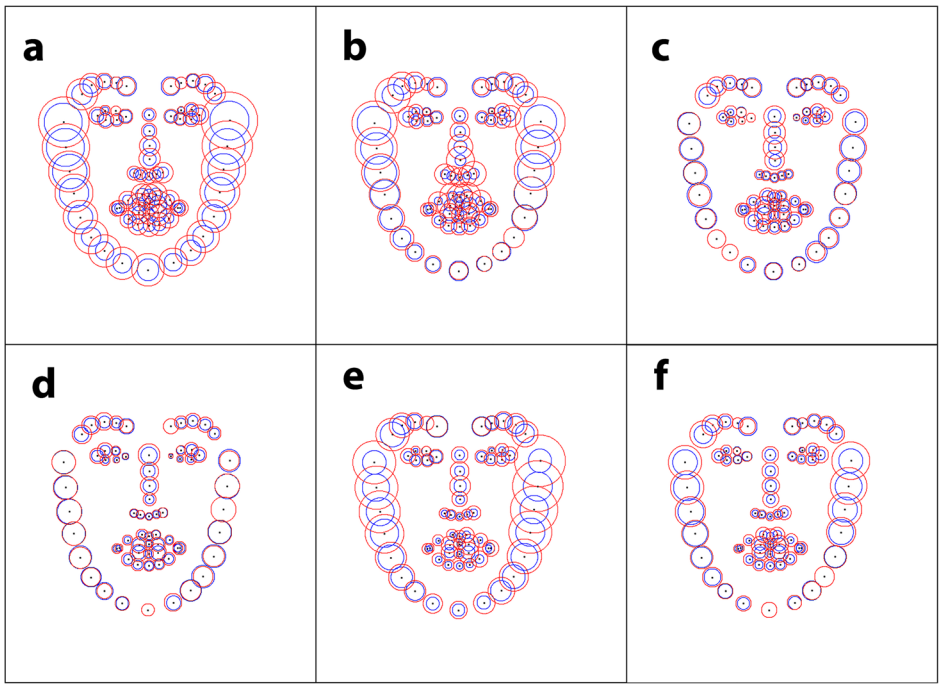

Fig. 10 RNSE distance. Blue circles are RMSE at 135 mm, red circles are 24 mm. **a** HAAM, **b** PAAM, **c** PRNet, **d** 3DSTN, **e** 3D-Fan, and **f** average result for all algorithms

Most of the facial landmark and alignment algorithms perform well at frontal views, and the detection precision relies on the training set variability. Attempting to delineate prediction differences between extreme-view cases and center view cases, we chose the most centered view image ($\phi = 90°$ and $\theta = 90°$), as well as 8 images surrounding by it, to be the frontal group (Fig. 16). The rest of the images are the outer group (Fig. 17). Front view detection can approach almost 100% accuracy especially at center view for deep neural networking methods. The precision rate drops more than 50% approaching extreme angles ($\theta = 0°$ and $\theta = 180°$).

Part of the set of the images used were also based on 3D captures of action units from FACS which taxonomizes individual physical expression of emotions. The results shown in Fig. 18 illustrate that in general the landmark-prediction methods work better on neutral faces due to FACS faces having more facial expressions which increase prediction difficulties. Performance decreases across wide field of view and view angle are consistent.

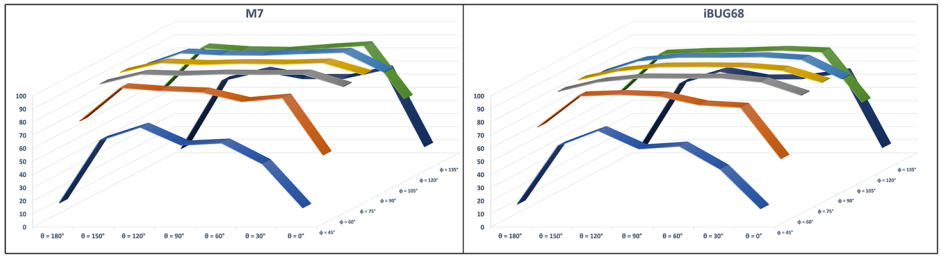

Fig. 11 PRNet: camera position varying across ϕ (from 45° to 135°) and θ (from 0° to 180°)

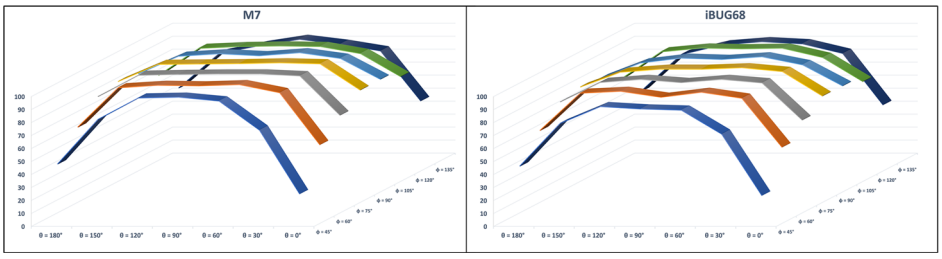

Fig. 12 3DFan: camera position varying across ϕ (from 45° to 135°) and θ (from 0° to 180°)

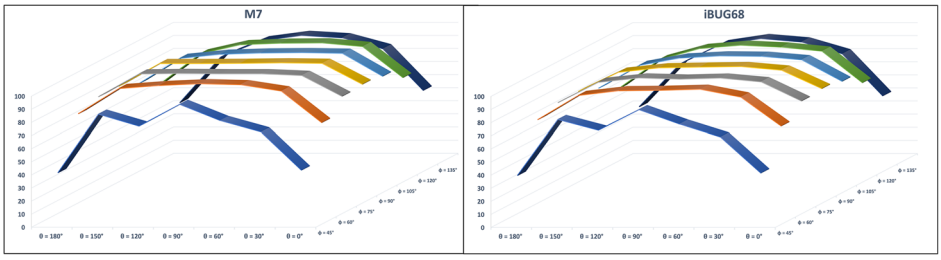

Fig. 13 3DSTN: camera position varying across ϕ (from 45° to 135°) and θ (from 0° to 180°)

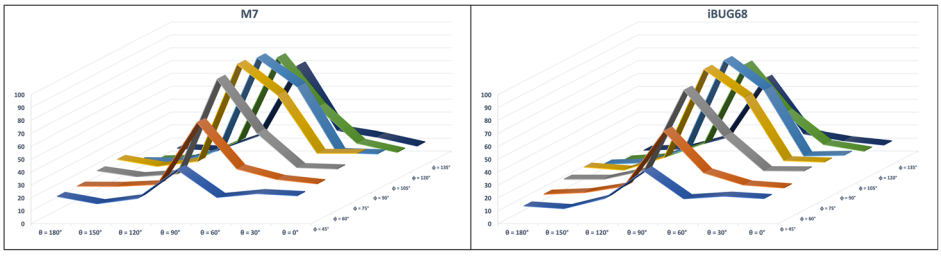

Fig. 14 HAAM: camera position varying across ϕ (from 45° to 135°) and θ (from 0° to 180°)

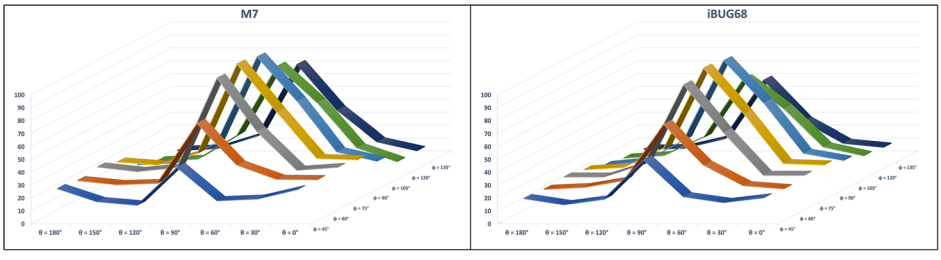

Fig. 15 PAAM: camera position varying across ϕ (from 45° to 135°) and θ (from 0° to 180°)

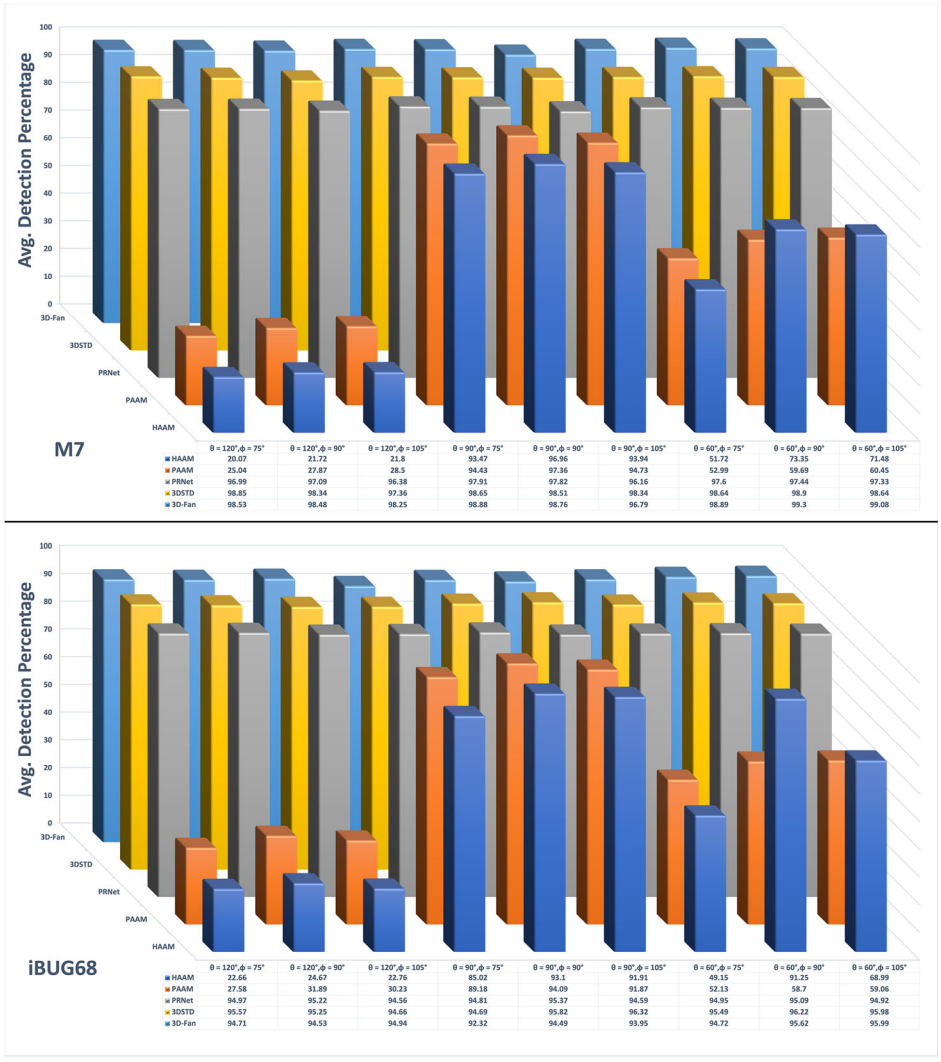

Fig. 16 Camera position with group of frontal view. The frontal view positions where $\phi = 75°, 90°$, and $105°$; $\theta = 60°, 90°$, and $120°$

5 Conclusion

In conclusion 3DSTN, PRNet, and 3D-FAN methods generally work better than traditional statistical methods. Deep-learning methods have become the prevalent research direction for the time being, but they are still subject to viewing angles and also, particularly, lens effects that have rarely been considered during any performance evaluations.

Increasing focal length tends to improve the landmark and alignment performances due to less projection distortion. This could inform design decisions for camera system and lens chosen for a biometric system, or it could be used to inform future algorithm design.

Given experimental results, all methods, as expected, work best from frontal-viewing angles. It is also interesting to note that the slope of fall-off for the performance decrease introduced by shorter focal lengths (wider field of view) is less for the AAM based methods and the 3DSTN approach. This is likely due to the AAM methods being based on image features, and the PAAM more specifically emphasizing local image features.

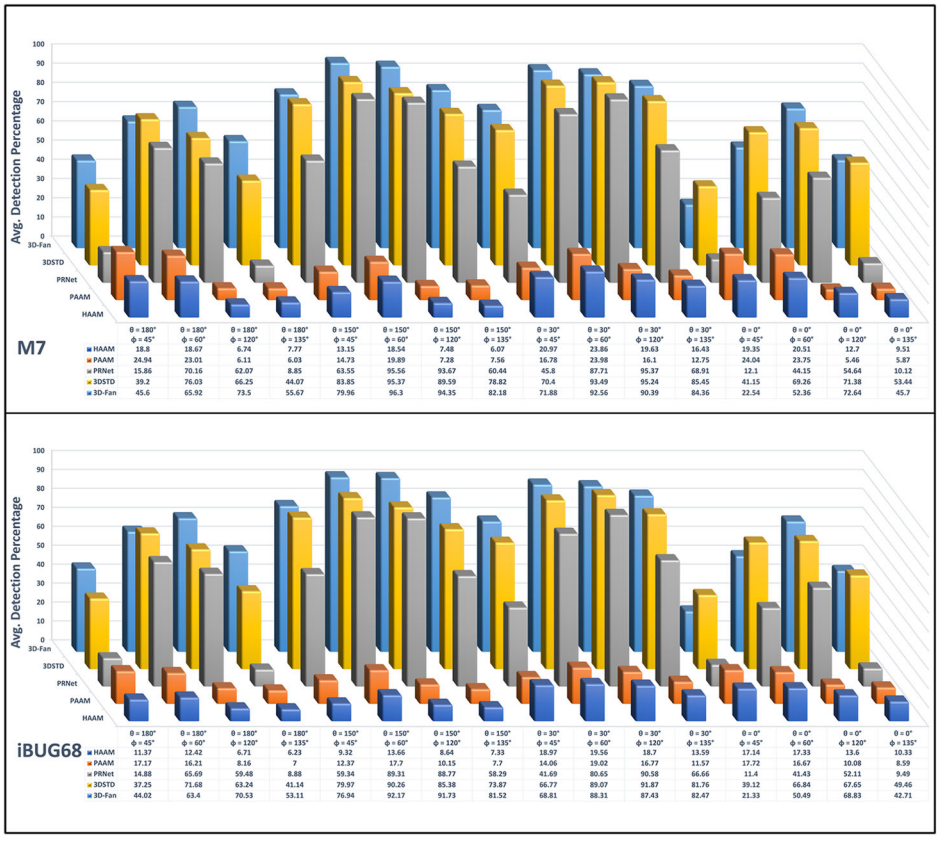

Fig. 17 Camera position with group of outer view. The outer view position where $\phi = 45°, 60°, 120°$, and $135°; \theta = 0°, 30°, 150°$, and $180°$

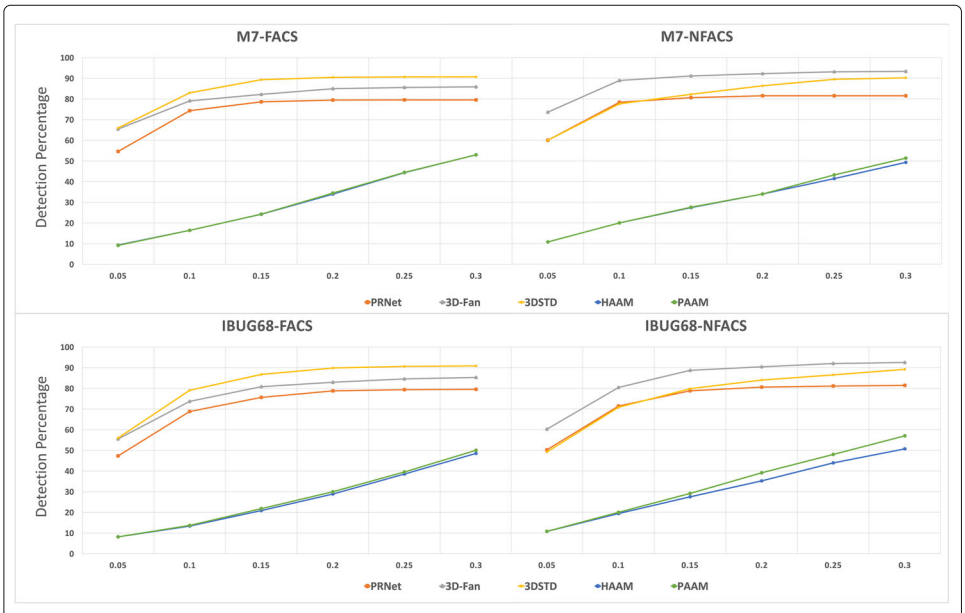

Fig. 18 FACS and non-FACS comparison. **a** FACS NRMSE and non-FACS NRMSE with M7 and **b** is iBUG-68 landmark schemes

3DSTN likely does well as part of the method specifically estimates a camera projection matrix, which in some sense should help counteract some of the focal length introduced perspective issues. PRNET and 3D-FAN methods using more general 3D data are likely more affected, and the larger training set for 3D-FAN likely assists its performance here.

One limitation of statistical algorithms is the landmark detection performance is tied to the head pose variation in the training set. When applying PCA, the first N eigen vectors are chosen as the main components. Typically, these are chosen based on representing $\pm$ 3 standard deviations from the mean value. Based on this limitation, the landmarking performance for extreme view angles, as often shown, drops. However, the CNN methods that all incorporate some system of 3D reference tend to do better as viewing angles move from the center; however, they still suffer performance drops and are still affected by shorter focal lengths.

Since focal length variance does affect final face landmark and alignment performance, future work could include use of this to augment training data. This could be done through data collection or use of synthetic data.

Meta-data from capture lenses stored in digital photographs is often removed by the time images reach large datasets, but it would be interesting to note such effects from in-the-wild photographs. In the meantime, training with synthetic data that includes controlled variance of viewing angle ranges as well as varying focal length, added to photographic datasets, should likely improve results.

In the future, image acquisition should not only cover pose, illumination, expression, ethnicity, skin color, etc., but also include consideration of full camera and lens parameters when possible.

Abbreviations

ASM: Active shape model; AAM: Active appearance model; FACS: Facial Action Coding System; Efficient Perspective n-Point(EPnP); NFACS: Non Facial Action Coding System; AUs: Action Units; RMSE: Root mean squared error; NRMSE: Normalized root mean squared error; PCA: Principal component analysis; HAAM: Holistic active appearance model; PAAM: Patch active appearance model; TPS: Thin plate spline; PRNet: Position map regression network; CNN: convolutional neural network; 3DMM: 3D morphable model; 3D-FAN: 3D face alignment network; 3DSTN: 3d spatial transformer network

Acknowledgements

No additional acknowledgements.

Authors' contributions

Xiang Li was primary author of the experimental rendering and data analysis as well as alignment performed by AAM in Menpo. Marcus Liu was responsible for training and implementation of the PRNet and 3D-FAN models used for alignment. Khoa Lua was responsible for alignment of the data using the 3DSTN method. Jessica Baron was responsible for portions of the software used to create the rendering as well as some of the numerical analysis. Eric Patterson was the coordinator and designer of the experiment and final editor for the paper after primary writing by Xiang Li.

Funding

This work was not funded by any external body and thus was not affected by any aspects of such.

Availability of data and materials

Please contact authors for data requests.

Consent for publication

We have consent to publish for images of individuals from 3D scans.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] School of Computing, Clemson University, 304 McAdams Hall, 29630 Clemson, SC, USA. [2] College of Computer Science & Information Engineering, Tianjin University of Science & Technology, 13th St, Binhai Xinqu, 300457 Tianjin, China. [3] Department of Computer Science and Computer Engineering, University of Arkansas, JBHT #521, 72701 Fayetteville, AR, USA.

Received: 28 February 2020 Accepted: 2 February 2021
Published online: 29 March 2021

References

1. Y. Wu, Q. Ji, Facial landmark detection: A literature survey. Int. J. Comput. Vis. **127**(2), 115–142 (2018). https://doi.org/10.1007/s11263-018-1097-z
2. X. P. Burgos-Artizzu, P. Perona, P. Dollár, in *Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV '13*, Robust Face Landmark Estimation under Occlusion (IEEE Computer Society, USA, 2013), pp. 1513–1520. https://doi.org/10.1109/ICCV.2013.191
3. J. Shi, A. Samal, D. Marx, How effective are landmarks and their geometry for face recognition?. Comp. Vision Image Underst. **102**(2), 117–133 (2006). https://doi.org/10.1016/j.cviu.2005.10.002
4. A. Kae, Incorporating Boltzmann Machine Priors for Semantic Labeling in Images and Videos (2014). https://doi.org/10.7275/37zj-rc94
5. X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014). https://doi.org/10.1007/s11263-013-0667-3
6. O. Çeliktutan, S. Ulukaya, B. Sankur, A comparative study of face landmarking techniques. EURASIP J. Image Video Process. **2013**(1), 13 (2013). https://doi.org/10.1186/1687-5281-2013-13
7. B. Johnston, P. d. Chazal, A review of image-based automatic facial landmark identification techniques. EURASIP J. Image Video Process. **2018**(1), 86 (2018). https://doi.org/10.1186/s13640-018-0324-4
8. C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: Database and results. Image vision Comput. **47**, 3–18 (2016)
9. W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, Q. Zhou, Look at Boundary: A Boundary-Aware Face Alignment Algorithm. arXiv (2018). https://arxiv.org/abs/1805.10483
10. M. Köstinger, P. Wohlhart, P. M. Roth, H. Bischof, in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization, (2011), pp. 2144–2151. https://doi.org/10.1109/ICCVW.2011.6130513. https://ieeexplore.ieee.org/document/6130513
11. X. Wang, L. Bo, L. Fuxin, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression, (2019), pp. 6970–6980. https://doi.org/10.1109/ICCV.2019.00707. https://ieeexplore.ieee.org/document/9010657
12. R. Valle, J. M. Buenaposada, A. Valdes, L. Baumela, in *Proceedings of the European Conference on Computer Vision (ECCV)*, A Deeply-initialized Coarse-to-fine Ensemble of Regression Trees for Face Alignment, (2018). https://openaccess.thecvf.com/content_ECCV_2018/html/Roberto_Valle_A_Deeply-initialized_Coarse-to-fine_ECCV_2018_paper.html
13. J. Su, Z. Wang, C. Liao, H. Ling, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Efficient and Accurate Face Alignment by Global Regression and Cascaded Local Refinement, (2019), pp. 267–276. https://doi.org/10.1109/CVPRW.2019.00036. https://ieeexplore.ieee.org/document/9025428
14. M. Kowalski, J. Naruniec, T. Trzcinski, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment, (2017), pp. 2034–2043. https://doi.org/10.1109/CVPRW.2017.254. https://ieeexplore.ieee.org/document/8014988
15. J. Lv, X. Shao, J. Xing, C. Cheng, X. Zhou, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, A Deep Regression Architecture with Two-Stage Re-initialization for High Performance Facial Landmark Detection, (2017), pp. 3691–3700. https://doi.org/10.1109/CVPR.2017.393. https://ieeexplore.ieee.org/document/8099876
16. N. Damer, Y. Wainakh, O. Henniger, C. Croll, B. Berthe, A. Braun, A. Kuijper, in *2018 24th International Conference on Pattern Recognition (ICPR)*, Deep Learning-based Face Recognition and the Robustness to Perspective Distortion, (2018), pp. 3445–3450. https://doi.org/10.1109/ICPR.2018.8545037. https://ieeexplore.ieee.org/document/8545037
17. J. Valente, S. Soatto, in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Perspective distortion modeling, learning and compensation, (2015), pp. 9–16. https://doi.org/10.1109/CVPRW.2015.7301314. https://ieeexplore.ieee.org/document/7301314
18. A. Flores, E. Christiansen, D. Kriegman, S. Belongie, in *Advances in Visual Computing*. ed. by G. Bebis, R. Boyle, B. Parvin, D. Koracin, B. Li, F. Porikli, V. Zordan, J. Klosowski, S. Coquillart, X. Luo, M. Chen, and D. Gotz, Camera distance from face images (Springer, Berlin, Heidelberg, 2013), pp. 513–522
19. P. J. Burt, *The Pyramid as a Structure for Efficient Computation*. (A. Rosenfeld, ed.) (Springer Berlin Heidelberg, Berlin, 1984), pp. 6–35. https://doi.org/10.1007/978-3-642-51590-3_2. https://link.springer.com/chapter/10.1007/978-3-642-51590-3_2
20. C. Harris, M. Stephens, in *Proceedings of the Alvey Vision Conference 1988*, A Combined Corner and Edge Detector (Alvey Vision Club, 1988). https://doi.org/10.5244%2Fc.2.23. https://core.ac.uk/display/21892060
21. G. Tzimiropoulos, S. Zafeiriou, M. Pantic, in *2011 International Conference on Computer Vision*, Robust and efficient parametric face alignment, (2011), pp. 1847–1854. https://doi.org/10.1109/ICCV.2011.6126452. https://ieeexplore.ieee.org/document/6126452
22. X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. Image Vis. Comput. **32**(10), 692–706 (2014). https://doi.org/10.1016/j.imavis.2014.06.002. https://www.sciencedirect.com/science/article/pii/S0262885614001012?via%3Dihub
23. V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang, S. Lazebnik, P. Perona, Y. Sato, C. Schmid, in *Computer Vision – ECCV 2012*. ed. by A. Fitzgibbon, Interactive Facial Feature Localization (Springer Berlin Heidelberg, Berlin, 2012), pp. 679–692. https://link.springer.com/chapter/10.1007/978-3-642-33712-3_49
24. R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie. Image Vision Comput. **28**(5), 807–813 (2010)
25. C. H. Hjortsjö, *Man's Face and Mimic Language*. (Studentlitteratur, 1969). https://books.google.com/books?id=BakQAQAAIAAJ. https://books.google.com/books/about/Man_s_Face_and_Mimic_Language.html?id=BakQAQAAIAAJ
26. CMU, *FACS - Facial Action Coding System*, (2002). https://www.cs.cmu.edu/~face/facs.htm

27. T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Active shape models-their training and application. Comput. vision image Underst. **61**(1), 38–59 (1995)
28. J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, S. Zafeiriou, in *Proceedings of the 22nd ACM International Conference on Multimedia*, Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models (Association for Computing Machinery, New York, 2014), pp. 679–682. https://doi.org/10.1145/2647868.2654890
29. Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, in *Computer Vision – ECCV 2018*. ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network (Springer International Publishing, Cham, 2018), pp. 557–574. https://github.com/YadiraF/PRNet. https://link.springer.com/chapter/10.1007%2F978-3-030-01264-9_33
30. A. Bulat, G. Tzimiropoulos, in *2017 IEEE International Conference on Computer Vision (ICCV)*, How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks) (IEEE. https://doi.org/10.1109/iccv.2017.116. https://arxiv.org/abs/1703.07332
31. C. Bhagavatula, C. Zhu, K. Luu, M. Savvides, in *2017 IEEE International Conference on Computer Vision (ICCV)*, Faster than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses, (2017), pp. 4000–4009. https://doi.org/10.1109/ICCV.2017.429
32. T. F. Cootes, G. Edwards, C. J. Taylor, in *Proc. British Machine Vision Conf*, Comparing Active Shape Models with Active Appearance Models (BMVA Press, Durham, 1999), pp. 173–182. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.524

Publisher's Note