Facial Landmark Detection Evaluation on MOBIO Database

Na Zhang

Abstract—**MOBIO is a bi-modal database that was captured almost exclusively on mobile phones. It aims to improve research into deploying biometric techniques to mobile devices. Research has been shown that face and speaker recognition can be performed in a mobile environment. Facial landmark localization aims at finding the coordinates of a set of pre-defined key points for 2D face images. A facial landmark usually has specific semantic meaning, e.g. nose tip or eye centre, which provides rich geometric information for other face analysis tasks such as face recognition, emotion estimation and 3D face reconstruction. Pretty much facial landmark detection methods adopt still face databases, such as 300W, AFW, AFLW, or COFW, for evaluation, but seldomly use mobile data. Our work is first to perform facial landmark detection evaluation on the mobile still data, i.e., face images from MOBIO database. About 20,600 face images have been extracted from this audio-visual database and manually labeled with 22 landmarks as the groundtruth. Several state-of-the-art facial landmark detection methods are adopted to evaluate their performance on these data. The result shows that the data from MOBIO database is pretty challenging. This database can be a new challenging one for facial landmark detection evaluation.**

Index Terms—**Facial landmark detection, detection performance, deep learning**

I. Introduction

The mobile biometrics database, MOBIO [1], is an audio-visual database captured almost exclusively using mobile phones. It is taken from 152 persons with 100 males and 52 females, and collected from August 2008 until July 2010 in six different sites from five different countries with both native and non-native English speakers. This mobile phone database consists of over 61 hours of audio-visual data with 12 distinct sessions usually separated by several weeks. One special point is that the acquisition device is given to the user, rather than being in a fixed position, which makes this database unique and now being used in an interactive and uncontrolled manner. The MOBIO database provides a challenging test-bed for face verification, speaker verification, and bi-modal verification.

Facial landmark detection, also known as face alignment or facial landmark localization, is a mature field of research. In recent years, facial landmark detection has become a vary active area, due to its importance to a variety of image and video-based face analysis systems, such as face recognition [2]–[5], facial expression analysis [6]–[9], human-computer interaction, video games and 3D face reconstruction [10]–[15]. Hence, accurate face landmarking and facial feature detection is an important intermediary step for many subsequent face processing operations that range from biometric recognition to the understanding of mental states, which have an impact on subsequent tasks focused on the face, such as coding, face recognition, expression and/or gesture understanding, gaze detection, animation, face tracking, etc.

Since face alignment is essential to many face applications, the requirement for the efficiency of facial landmark detection becomes higher and higher, especially when more face images and videos captured in the wild appear. Hence, the large visual variations of faces, such as occlusions, large pose variations and extreme lightings, impose great challenges for face alignment in real world applications. For facial landmark detection evaluation, still face images, like 300W, AFW, AFLW, are universally used. However, mobile face data, e.g., the MOBIO database, is seldomly adopted for facial landmark evaluation so far. In this work, we try to perform facial landmark detection on the mobile still face data using up-to-date methods, and check their performance on this type of faces.

A total of 20,600 still face images are extracted from MOBIO database and labelled manually with 22 facial feature points as groundtruth. Seven state-of-the-art facial landmark localization methods are chosen to perform facial landmark detection on these face images. And several measure metrics, e.g., Normalized Mean Error (NME), Cumulative Error Distribution (CED), Area-Under-the-Curve (AUC) and failure rate, are calculated or drawn for evaluation. The experimental result shows that these mobile still face images are pretty challenging for existing facial landmark detection technology and could be a new database for facial landmark localization. This evaluation could establish baseline performance for the MOBIO mobile face images.

The contributions of our work includes:

- generate a still mobile face database with a total of 20,600 images based on video-visual database MOBIO with 22 manually labelled facial landmarks as groundtruth;
- adopt seven state-of-the-art facial landmark detection methods to evaluate their performance on these 20,600 face images;
- the result shows that the mobile faces in MOBIO is pretty challenging which can be used as a new database for facial landmark detection evaluation in mobile condition.

This paper is organized as follows. In section II, we briefly describe facial landmark detection technique. In section III, the still faces based on MOBIO are generated and labeled with 22 facial landmarks. Our approach procedure is given in section IV. And experimental results are provided in section V. In section VI, some interesting discussion and conclusions are drawn.

Na Zhang is with Lane Department of Computer Science and Electrical Engineering at West Virginia University, Morgantown, WV 26506-6109.

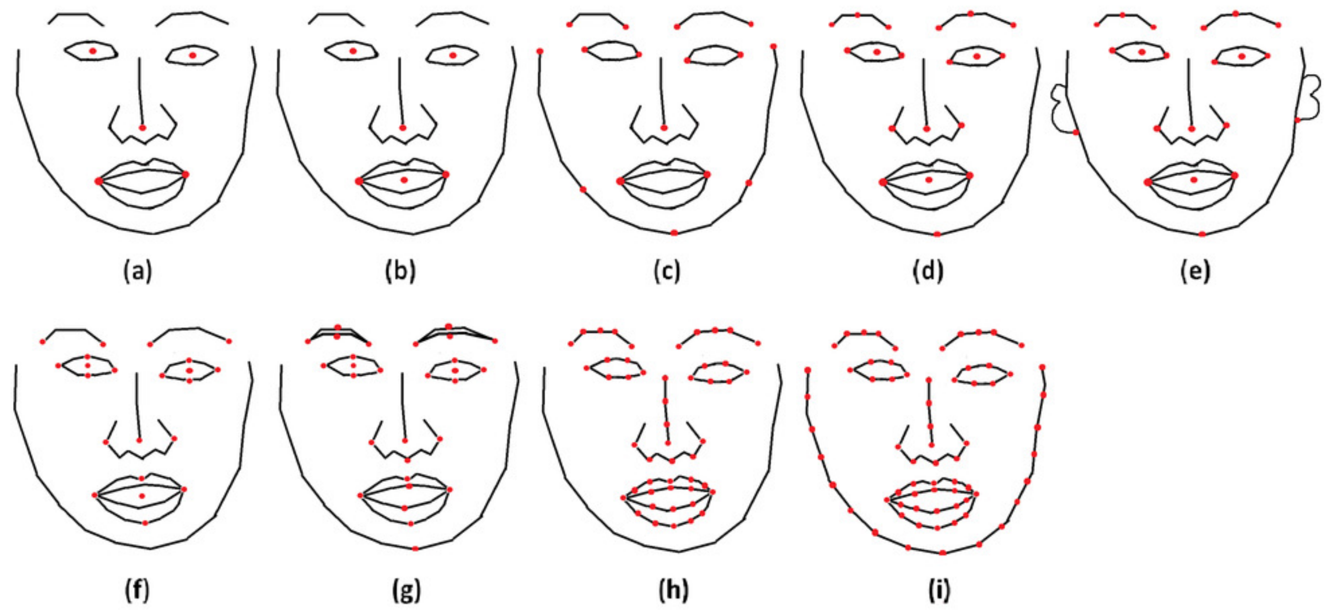

Fig. 1. Types of facial landmarks with (a) 5, (b) 6, (c) 16, (d) 19, (e) 21, (f) 22, (g) 29, (h) 49, (i) 68 points.

II. Facial Landmark Detection

In this section, we talk about some basic information about facial landmark detection, the challenges it faces, several categories of detection methods, types of facial points and measure metrics.

A. What is Facial Landmark Detection?

Facial landmark detection, or facial landmark localization, or face alignment, is to automatically localize a set of predefined semantic key points including eyes, nose, mouth and other points on the 2D face images. A facial landmark usually has specific semantic meaning, e.g. nose tip or eye center, which provides rich geometric information for other face analysis tasks such as face recognition [2]–[5], emotion estimation [6]–[9] and 3D face reconstruction [10]–[15]. It is a fundamental problem in computer vision study and an essential initial step for a number of research areas, and plays a key role in many face processing applications, including head pose estimation [16], [17], facial expression analysis and emotion recognition [18]–[20], face attribute analysis [21], [22], face alignment in 2D [23], [24] and 3D (e.g., frontalization [25], face 3D modeling, video games, multimodal sentiment analysis [26], person identification, and, of course, face recognition (see, e.g., Sun et al. [27] and many others).

B. What is the Challenge?

Due to its relevance to many facial analysis tasks, facial landmark detection has attracted increasing interests in the past couple of years. It is a well-researched problem with large amounts of annotated data, and impressive progress has been made too. Current methods could provide reliable results for near-frontal face images [23], [24], [28]–[31]. Thanks to the successive developments in this area of research during the past decades, facial landmark localization can be performed very accurately in constrained scenarios, even using traditional approaches such as Active Shape Model (ASM) [32], Active Appearance Model (AAM) [33] and Constrained Local Model (CLM) [22]. As the rapid development of deep learning technology, facial landmark detection gains a pretty good performance in unconstrained environment.

Though great strides have been made in this field, facial landmark detection is particularly daunting considering the real-world, unconstrained imaging conditions. In an uncontrolled setting, face is likely to have large out-of-plane tilting, occlusion, illumination and expression variations. Robust facial landmark detection remains a formidable challenge in the presence of partial occlusion and large head pose variations. Images often portray faces in myriads of poses, expressions, occlusions and more, any of which can affect landmark appearances, locations or even presence. Therefore, it is still a challenging problem for localizing landmarks in face images with partial occlusions or large appearance variations due to illumination conditions, poses, and expression changes.

C. Categories of Methods

In general, existing facial landmark detection methods can be divided into two categories: (1) traditional approaches, e.g., ASM [32] and AAM [33] based methods, which fit a generative model by global facial appearance; (2) cascade regression based methods, which try to estimate the facial landmark positions by a sequence of regression models. In recent year, deep learning based cascade regression models have performed robust facial landmark localization using deep neural networks.

ASM and AAM based methods. This kind of methods is traditional approaches, which usually perform accurately in constrained scenarios. They rely on a generative PCA-based shape model. However, these methods require expensive iterative steps and rely on good initialization. The mean shape is often used as the initialization, which may be far from the target position and hence inaccurate.

Cascade regression based methods. In cascade regression framework, a set of weak regressors are cascaded to form a strong regressor [23], [24], [34]–[37]. It tries to obtain the coarse location first, and the following steps are to refine the initial estimate, yielding more accurate results. Cascade regression directly positions facial landmarks on their optimal locations based on image features. The shape update is achieved in a discriminative way by constructing a mapping function from robust shape related local features to shape updates. However, these methods need to train individual systems for each group of the landmarks, the computational burden grows proportional to the group numbers and cascade levels. For example, the cascaded Convolutional Neural Network (CNN) method [38] needs to train 23 individual CNN networks. However, the capability of cascaded regression is nearly saturated due to its shallow structure. After cascading more than four or five weak regressors, the performance of cascaded regression is hard to improve further [39].

More recently, as deep neural networks have been put forward as a more powerful alternative in a wide range of computer vision and pattern recognition tasks, facial landmark localization gains large development too. Different network types have been explored, such as Convolutional Neural Network (CNN), Auto-Encoder Network and Recurrent Neural Network, to perform robust facial landmark localization. In our work, most of methods adopted belong to deep learning based models, such as Wing loss based method WingLoss [40].

D. Types of Facial Landmarks

Existing facial landmark detection methods can figure out different numbers of facial feature points, e.g., 5, 6, 16, 19, 21, 22, 29, 49, 68, etc. Figure 1 gives several typical facial landmarks. Figure 1(a) consists five feature points(i.e., left eye center, right eye center, nose tip, left mouth corner, and right mouth corner). Figure 1(b) is a face with six points with one more landmark, mouth center, than Figure 1. Besides feature points of eye area, nose, and mouth, Figure 1(c) considers five face contour landmarks. Figure 1(d) and (e) share similar landmarks, the only difference is Figure 1(e) have two extra points on two ears. Figure 1(f) (i) provide more points to describe geometric information of face.

E. Landmark Performance and Evaluation Metrics

There are two different metrics to evaluate landmark detection performance, task-oriented performance and ground-truth based localization performance. For task-oriented performance, one can measure the impact of the landmark detection accuracy on the performance scores of a task. For ground-truth based localization performance, a straightforward way is to use manually annotated ground-truths.

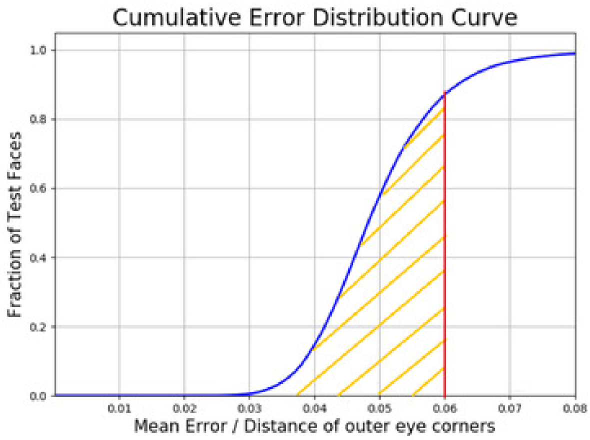

Fig. 2. An illustration of CED and AUC.

In practice, ground-truth based localization performance is commonly used in facial landmark detection for thorough analysis. If the ground-truth positions are available, the localization performance can be expressed in terms of the Normalized Mean Error (NME), Cumulative Error Distribution (CED) curve, Area-Under-the-Curve (AUC) and failure rate.

Normalized Mean Error (NME) is a primary metric in facial landmark detection evaluation. It is calculated first by the distances between the estimated landmarks and the groundtruths, and then normalized with respect to the inter-ocular distance, i.e. Euclidean distance between two eye centres, the distance of outer or inner corners of the eyes. The mean error has three types: landmark-wise, sample-wise or overall. Landmark-wise face alignment error is first normalized in the following way to make it scale invariant: $e_x = \frac{\|\hat{x}-x^{GT}\|}{D_{IOD}}$ where $\|\hat{x}-x^{GT}\|$ is the Euclidean distance between the estimated location $\hat{x}$ and the true location x^{GT}. D_{IOD} is the inter-ocular distance (IOD). Normalizing landmark localization errors by dividing with IOD makes the performance measure independent of the actual face size or the camera zoom factor. Sample-wise mean error could be calculated as $\frac{1}{n}\sum_{i=1}^{n}\frac{\|\hat{x_i}-x_i^{GT}\|}{D_{IOD}}$, where n is the number of facial landmarks involved in the evaluation. The error is normalized by the distance of outer or inner corners of the eyes. NME can be averaged over all the landmarks to produce a global precision figure, which is a overall mean error. In recent years, with the rapid progress of face alignment, most of the recent approaches report a error level of e at around 0.05 or smaller, which is close to human performance.

Using NME is very straightforward and intuitive given its single value form. However, this measure is heavily impacted by the presence of some big failures such as outliers, in particular when the average error level is very low. In other words, the mean error measure is very fragile even if there are just a few images with big errors. Thus though the mean error is widely used for face alignment evaluation [28], [38], [41]–[43], it does not provide a big picture on which cases the errors occur, e.g., minor big alignment error, many inaccuracies.

Since using overall mean error as an evaluation criterion is too sensitive to big erroneous samples, Cumulative Error Distribution (CED) curve and Area-Under-the-Curve (AUC) are adopted as two better metrics (see Figure 2). CED curve is the cumulative distribution function of the normalized error as shown by the blue line in Figure 2. The x-axis is error value, and y-axis is the fraction of test faces. In terms of outliers handling, CED is a better way. However, it is not intuitive given its curve representation. It is also hard to use it in sensitivity analysis. Therefore, AUC is adopted. It is calculated from the CED curve. The AUC stands for the value of area under the curve of CED. It is defined as: $AUC_\alpha = \int_0^\alpha f(e)de$ where e is the normalized error, f(e) is the cumulative error distribution function and α is the upper bound that is used to calculate the definite integration. In Figure 2, α is 0.06 (red line). Given the definition of the CED function, the value is AUC_α lies in the range of [0, α], the area with yellow titled lines. The value of AUC_α will not be influenced by points with error bigger than α.

Landmark detection statistics can be characterized by the exceedance probability of the localization error. A general agreement in the literature is that e ¡ 0.1 is an acceptable error criterion so that a landmark is considered detected whenever it is found within proximity of one tenth of the inter-ocular distance from its true position. In our experiment, α is set to 0.08 and 0.1.

Failure rate is calculated with the threshold for the normalized mean error. It computes the fraction of test faces that the error value of which is larger than the threshold. In our experiment, the thresholds are set to 0.1 and 0.08.

F. Common Databases Used

Annotated databases are extremely important in computer vision. Therefore, a number of databases containing faces with different facial expressions, poses, illumination and occlusion variations have been collected in the past. Most evaluation experiments are conducted on commonly used benchmark datasets, such as the 300 Faces in the Wild (300W) [44], Annotated Facial Landmarks in the Wild (AFLW) [45], the Annotated Faces in-the-wild (AFW) [17], the Labeled Face Parts in-the-wild (LFPW) [46], HELEN [47], the Caltech Occluded Faces in the Wild (COFW) [48]. The aforementioned databases, cover large variations including: different subjects, poses, illumination, occlusion, etc.

The 300 Faces in the Wild (300W) [44] dataset is a commonly used benchmark for facial landmark localization problem. It contains near-frontal face images in the wild and provides 68 semi-automatically annotated points for each face. It is created from existing datasets, including LFPW [46], AFW [17], HELEN [47], XM2VTS [49] and IBUG [44]. The 300W training set contains 3,148 training images from AFW, LFPW and HELEN. The common subset of 300W contains 554 test images from LFPW and HELEN. The challenging subset of 300W contains 135 test images from IBUG. The fullset of 300W is the union of the common and challenging subset. The 300W test set contains 600 test images which are provided officially by the 300W competition [44] and said to have a similar distribution to the IBUG dataset. IBUG subset is extremely challenging due to the large variations in face pose, expression and illumination. As the rapid progresses of the study in facial landmark localization in recent years, several methods have reported close-to-human performance on this dataset of which the images that are acquired from unconstrained environments. Figure 3 shows annotated (a) indoor and (b) outdoor images of 300W.

The Annotated Faces in-the-wild (AFW) [17] database is a popular benchmark for facial landmark detection, containing 205 images with 468 faces. A detection bounding box as well as up to 6 visible landmarks are provided for each face. An example of an image taken from the AFW database along with the corresponding annotated landmarks is depicted in Figure 3 (c).

The Annotated Facial Landmarks in-the-wild (AFLW) [45] is a very challenging dataset that has been widely used for benchmarking facial landmark localization algorithms. It contains 25,000 images of 24,686 subjects downloaded from Flickr. The images contain a wide range of natural face poses in yaw (from -90 to 90) and occlusions. Facial landmark annotations are available for the whole database. Each annotation consists of 21 landmark points. The AFLW-Full protocol contains 20,000 training and 4,386 test images, and each image has 19 manually annotated facial landmarks. Figure 3 (f) depicts an annotated image from AFLW.

The COFW [48] dataset contains in-the-wild face images with heavy occlusions, including 1,345 face images for training and 507 face images for testing. For each face, 29 landmarks and the corresponding occlusion states are annotated in the COFW dataset. An example of an image taken from the COFW database along with the corresponding annotated landmarks is depicted in Figure 3 (g).

The Labeled Face Parts in-the-wild (LFPW) [46] database contains 1,432 images downloaded from google.com, fickr.com, and yahoo.com. The images contain large variations including pose, expression, illumination and occlusion. The provided ground truth consists of 29 landmark points. An example of an image taken from the LFPW database along with the corresponding annotated landmarks is depicted in Figure 3 (d).

The HELEN [47] database consists of 2,330 annotated images collected from the Flickr. The images are of high resolution containing faces of size sometimes greater than 500*500 pixels. The provided annotations are very detailed and contain 194 landmark points. Figure 3 (e) depicts an annotated image from HELEN.

III. Face Images Based on MOBIO Database

The MOBIO is an audio-video database of human faces and voice captured almost exclusively on mobile phones. This database is originally used to evaluate the performance of face and speaker recognition in the context of a mobile environment [1], [50].

So far, facial landmark detection technique has been seldomly evaluated in the context of a mobile environment. So it is meaningful to perform facial landmark detection on the mobile faces. The mobile environment was chosen as it provides a realistic and challenging test-bed for face points detection techniques to operate. For instance, the environment is quite complex and there is limited control over the illumination conditions and the pose of the subject for the video.

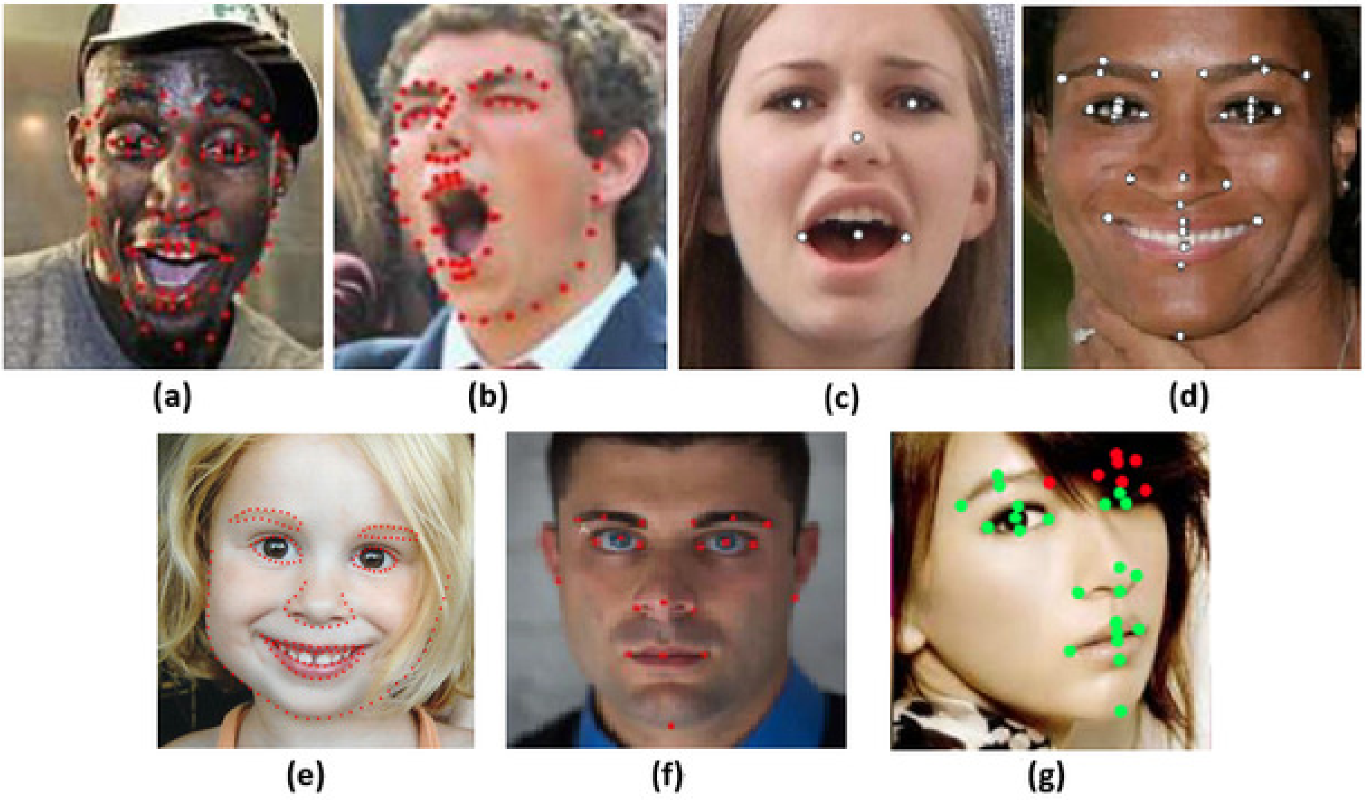

Fig. 3. Annotated images from (a) 300W (Indoor), (b) 300W (Outdoor), (c) AFW, (d) LFPW, (e) HELEN, (f) AFLW and (g) COFW.

In our work, we extract still face frames from the MOBIO videos and generate a face images database. This section briefly describes the MOBIO database first, and then introduces the face images extracted from the MOBIO video data, finally talks about how to generate the groundtruth of the faces with 22 feature points.

A. MOBIO Database

MOBIO [1] database is an unique diverse bi-modal database (audio + video) that was captured almost exclusively on mobile phones. It consists of over 61 hours of audio-visual data with 12 distinct sessions usually separated by several weeks. There are a total of 192 unique audio-video samples for each of the 152 participants. Female-Male ratio is 1:2. This data was captured at 6 different sites over one and a half years with people speaking English. Capturing the data on mobile phones makes this database unique because the acquisition device is given to the user, rather than being in a fixed position. This means that the microphone and video camera are no longer fixed and are now being used in an interactive and uncontrolled manner. This database was captured almost exclusively using mobile phones and aims to improve research into deploying biometrics techniques to mobile devices.

The database was acquired primarily on mobile phones. 12 sessions were captured for each participant. 6 sessions for Phase I and 6 sessions for Phase II. In Phase I, the participants are asked to answer a set of questions which are classified as set responses, read speech from a paper, and free speech. Each session consists of 21 questions: 5 pre-defined set response questions, 1 read speech question and 15 free speech questions. Phase II consists of 11 questions with the question types ranging from short response questions, set speech, and free speech.

All videos are recorded using two mobile devices: one mobile phone (NOKIA N93i) and one laptop computer (standard 2008 MacBook). The laptop was only used to capture part of the first session. The first session consists of data captured on both the laptop and the mobile phone.

The publicly-available mobile phone database MOBIO (Source download link: https://www.idiap.ch/dataset/mobio) presents several challenges, including: (1) high variability of pose and illumination conditions, even during recordings, (2) high variability in the quality of speech, and (3) variability in the acquisition environments in terms of acoustics as well as illumination and background.

B. Extracted Face Images and Facial Landmark Groundtruth

Based on the video data in MOBIO, a few face frames are extracted for each subject. A total of 20,600 still face images with size of 640*480 are generated finally. The average number of images for each subject is about 136. Figure 4 gives several face samples. Since all images are captured in unconstrained conditions, it contains big variations in head pose, illumination, occlusion (e.g., hair, glass), which makes it a challenging database.

Much work have been done to generate the groundtruth of these face images via manually labeling 22 facial feature

Fig. 4. Face samples of MOBIO database.

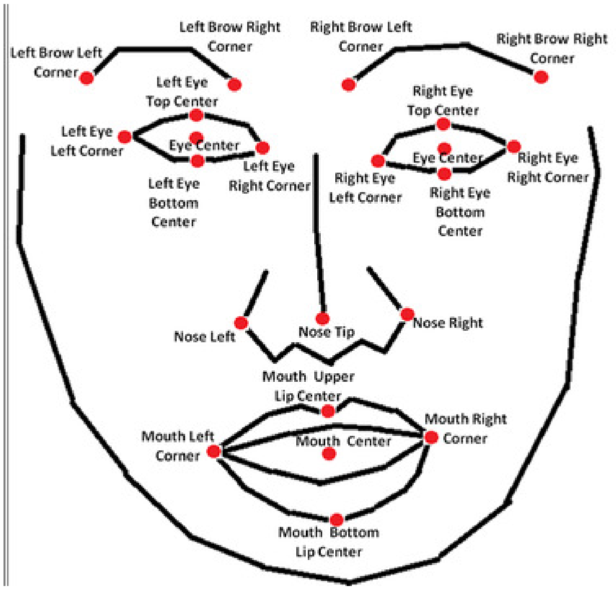

Fig. 5. 22 facial landmarks.

points. Figure 7 shows the 22 facial landmarks of the face, including 4 points describing brow (left brow left corner, left brow right corner, right brow left corner, right brow right corner), 10 points describing eyes (left eye left corner, left eye top center, left eye right corner, left eye bottom center, left eye center, right eye left corner, right eye top center, right eye right corner, right eye bottom center, right eye center), three points describing nose (nose tip, nose left, nose right), and five points describing mouth (mouth left corner, mouth upper lip center, mouth right corner, mouth bottom lip center, and mouth center).

In order to label these faces conveniently and efficiently, a labeling tool named 'Face Label App' as shown in Figure 6 was developed, which can run on Windows system. The users need to load face images first, and then click the 22 facial feature points on each face in a pre-defined order. The app can automatically capture the position (with x, y values) of each facial landmark when mouse moves on it and clicked. All facial landmark position information are saved in .txt files.

C. Preprocess Mobile Still Face Data

Some facial landmark detection methods are able to handle faces with any sizes, like DAC-CSR [51], PA-CNN Model [52], CE-CLM [53], etc. However, some detection methods, such as Tweaked CNN [54], WingLoss [40], and ECT [55], require that their input must be square faces with fixed size (e.g., 256*256). Hence, face detection, cropping and resizing are executed to preprocess the faces for these methods. MTCNN [56] model which is a pretty good face detector is adopted in our work for face detecting. Based on the bounding box of faces, all detected faces are cropped into square shape and then resized to fixed size.

IV. Our Approach

In this section, we choose several facial landmark detection methods (Tweaked CNN [54], PA-CNN Model [52], WingLoss [40], CE-CLM [53], ECT [55], TCDCN [43], DAC-CSR [51]) to perform face alignment task and analyze their performance on the mobile still faces and other commonly used face databases like 300W, AFW, AFLW, COFW. Normalized mean error (NME), Cumulative Error Distribution curve (CED), Area Under the error Curve (AUC) and failure rate are adopted as our measure metrics for evaluation.

A. Facial Landmark Detection Methods

We choose seven facial landmark detection methods to detect face feature points on mobile still face images. They

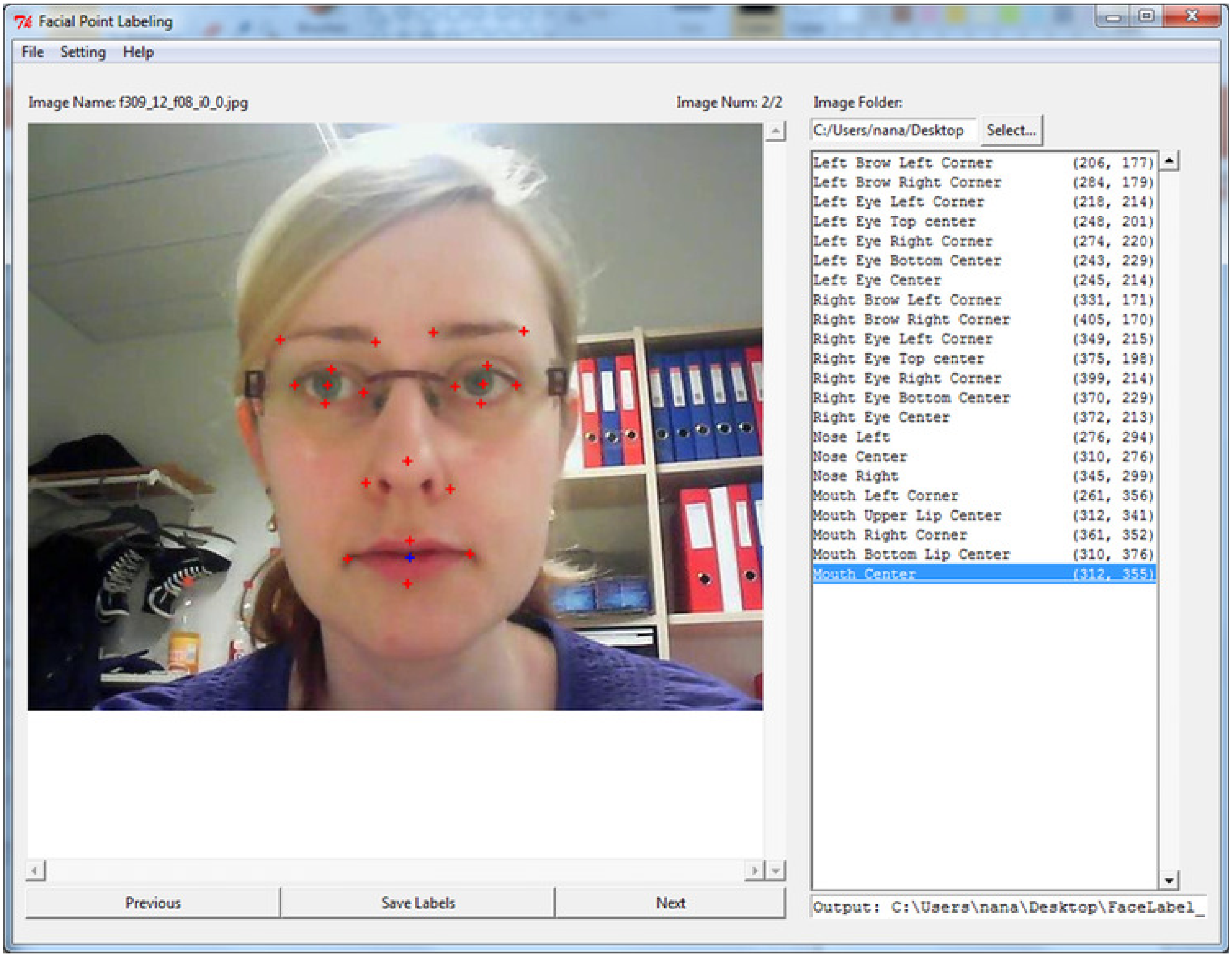

Image Name: f309_12_f08_i0_0.jpg

Image Num: 2/2

Image Folder:

C:/Users/nana/Desktop

Left Brow Left Corner	(206, 177)
Left Brow Right Corner	(284, 179)
Left Eye Left Corner	(218, 214)
Left Eye Top center	(248, 201)
Left Eye Right Corner	(274, 220)
Left Eye Bottom Center	(243, 229)
Left Eye Center	(245, 214)
Right Brow Left Corner	(331, 171)
Right Brow Right Corner	(405, 170)
Right Eye Left Corner	(349, 215)
Right Eye Top center	(375, 198)
Right Eye Right Corner	(399, 214)
Right Eye Bottom Center	(370, 229)
Right Eye Center	(372, 213)
Nose Left	(276, 294)
Nose Center	(310, 276)
Nose Right	(345, 299)
Mouth Left Corner	(261, 356)
Mouth Upper Lip Center	(312, 341)
Mouth Right Corner	(361, 352)
Mouth Bottom Lip Center	(310, 376)
Mouth Center	(312, 355)

```
Output: C:\Users\nana\Desktop\FaceLabel_
```

Fig. 6. Face Label App

are Tweaked CNN [54], WingLoss [40], DAC-CSR [51], PA-CNN Model [52], OpenPose [53], ECT [55], and TCDCN [43]. Most of them are deep learning based methods.

Among these deep learning methods, Tweaked CNN [54] detects 5 facial points, WingLoss [40] and DAC-CSR [51] detect 19 points, and the others detect 68 points. Some methods of them can do facial landmark detection directly on face images with any sizes. Tweaked CNN [54], WingLoss [40] and ECT [55] need face cropping with size of 256*256 before facial landmark detection. MTCNN [56] is adopted in our work to do face detection and cropping due to its efficiency.

B. Methods with 5 Points

Figure 7 (a) shows the five facial landmarks (left eye center, right eye center, mouth left corner, mouth right corner, nose tip) that are detected by Tweaked CNN [54] models.

1) Tweaked CNN Model: Based on the analysis that the features produced at intermediate layers of a convolutional neural network can be trained to regress facial landmark coordinates, face images can be partitioned in an unsupervised manner into subsets containing faces in similar poses (i.e., 3D views) and facial properties (e.g., presence or absence of eyewear). Therefore, Tweaked CNN (TCNN) [54] specializes in regressing the facial landmark coordinates of faces in specific poses and appearances. It is shown to outperform existing landmark detection methods in an extensive battery of tests on the AFW, ALFW, and 300W benchmarks.

C. Methods with 19 Points

Figure 7 (b) shows the 19 facial landmarks containing 6 points on brow (left brow left corner, left brow center, left brow right corner, right brow left corner, right brow center, right brow right corner), 6 points on eyes (left eye left corner, left eye center, left eye right corner, right eye left corner, right eye center, right eye right corner), 3 points on nose (nose left, nose tip, nose right), 3 points on mouth (mouth left corner, mouth center, mouth right corner), and 1 point on chin (lower chin center) used by WingLoss [40] and DAC-CSR [51] models.

1) WingLoss Model: WingLoss [40] method presents a piece-wise loss function, namely Wing loss, for robust facial landmark localization in the wild with Convolutional Neural

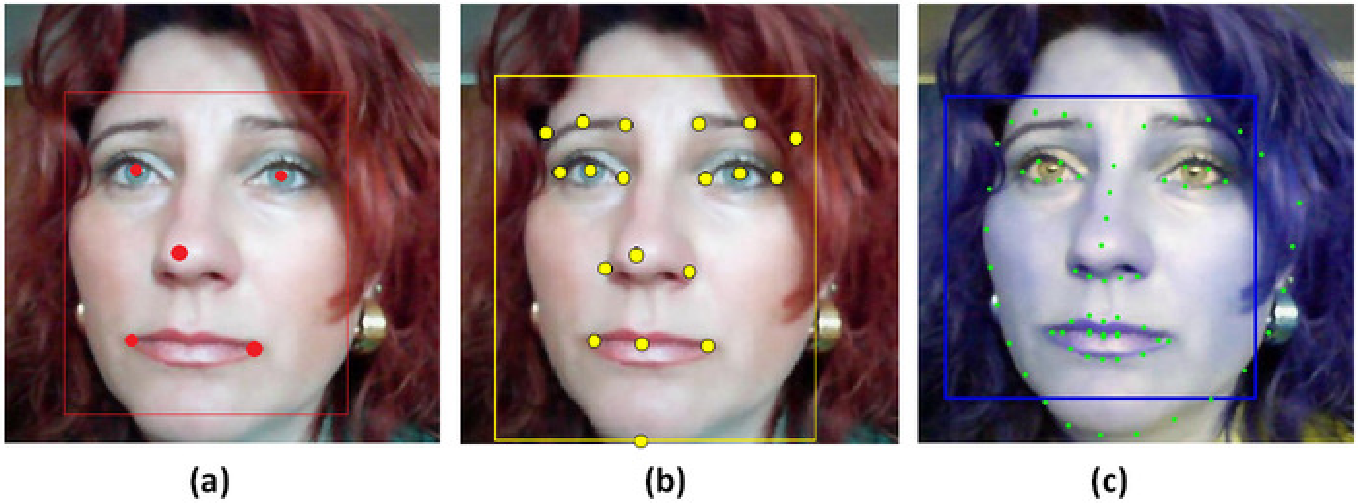

Fig. 7. 5, 19, and 68 Facial landmarks.

Networks (CNNs). The loss function pays more attention to small and medium range errors and amplifies the impact of errors from the interval (-w,w) by switching from L_1 loss to a modified logarithm function. The experimental results obtained on the AFLW (AFLW-Full protocol) and 300W datasets demonstrate the merits of the Wing loss function, and prove the superiority of the proposed method over the state-of-the-art approaches.

2) DAC-CSR Model: DAC-CSR [51], namely Dynamic Attention-Controlled Cascaded Shape Regression architecture, is for robust facial landmark detection on unconstrained faces. It divides facial landmark detection into three cascaded sub-tasks: face bounding box refinement, general cascaded shape regression and attention-controlled cascaded shape regression. The first two stages refine initial face bounding boxes and output intermediate facial landmarks. Then, an online dynamic model selection method is used to choose appropriate domain-specific cascaded shape regressions for further landmark refinement. The key innovation of the DAC-CSR is the fault-tolerant mechanism, using fuzzy set sample weighting, for attention-controlled domain-specific model training. It uses two challenging face datasets, AFLW and COFW, to evaluate the performance of the DAC-CSR architecture.

D. Methods with 68 Points

Figure 7 (c) shows the 68 facial landmarks including 17 contour landmarks and 51 inner landmarks. As shown in Figure 8, number 1-17 of points denote the contour landmarks and number 18-68 points denote the inner landmarks.

1) PA-CNN Model: PA-CNN Model [52] is short for Part-Aware Deep Convolutional Neural Network. It is an end-to-end regression framework for facial landmark localization. It encodes images into feature maps shared by all landmarks. Then, these features are sent into two independent sub-network modules to regress contour landmarks and inner landmarks, respectively. It incorporates the contour landmark sub-network and the inner landmark sub-network into a unified architecture. Contrary to others, this method does not involve multiple individual models or require auxiliary labels. More importantly, the framework treats landmarks on different facial part differently which helps to learn discriminative features. This method can directly detect landmarks on original images. It does not need face detection, cropping, and resizing. Extensive evaluations are conducted on 300W benchmark dataset.

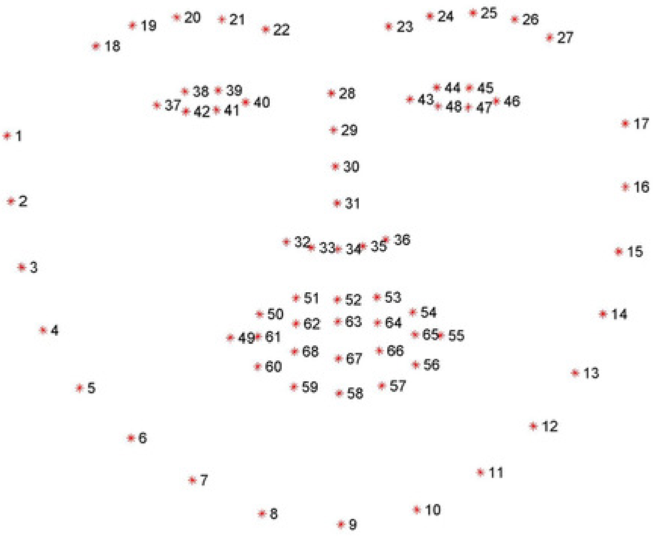

Fig. 8. 68 facial landmarks.

2) CE-CLM Model: Constrained Local Models (CLMs) are a well-established family of methods for facial landmark detection. However, they have recently fallen out of favor to cascaded regression-based approaches. This is in part due to the inability of existing CLM local detectors to model the very complex individual landmark appearance that is affected by expression, illumination, facial hair, makeup, and accessories.

CE-CLM [53] introduced a member of CLM family, Convolutional Experts Constrained Local Model (CE-CLM), in which it uses a local detector called Convolutional Experts Network (CEN). CEN brings together the advantages of neural architectures and mixtures of experts in an end-to-end framework. It is able to learn a mixture of experts that capture different appearance prototypes without the need of explicit attribute labeling, and is able to deal with varying appearance of landmarks by internally learning an ensemble of detectors, thus modeling landmark appearance prototypes. This is achieved through a Mixture of Expert Layer, which consists of decision neurons connected with non-negative weights to the final decision layer. Convolutional Experts Constrained Local Model (CE-CLM) algorithm consists of two main parts: response map computation using Convolutional Experts Network and shape parameter update. CE-CLM is able to perform well on facial landmark detection and is especially accurate and robust on challenging profile images.

3) ECT Model: The three-step framework named ECT (Estimation-Correction-Tuning) [55] is an effective and robust approach for facial landmark detection by combining data- and model-driven methods. Firstly, a Fully Convolutional Network (FCN) which is a data-driven method is trained to compute response maps of all facial landmark points, which makes full use of holistic information in a facial image for global estimation of facial landmarks. After that, the maximum points in the response maps are fitted with a pre-trained Point Distribution Model (PDM) to generate the initial facial shape. This model-driven method is able to correct the inaccurate locations of outliers by considering the shape prior information. Finally, a weighted version of Regularized Landmark Mean-Shift (RLMS) is employed to fine-tune the facial shape iteratively.

This Estimation-Correction-Tuning process perfectly combines the advantages of the global robustness of data-driven method (FCN), outlier correction capability of model-driven method (PDM) and non-parametric optimization of RLMS. The method is able to produce satisfying detection results on face images with exaggerated expressions, large head poses, and partial occlusions.

4) TCDCN Model: Facial landmark detection has long been impeded by the problems of occlusion and pose variation. Instead of treating the detection task as a single and independent problem, TCDCN [43] investigate the possibility of improving detection robustness through multi-task learning. This tasks-constrained deep model can facilitate learning convergence with task-wise early stopping. It optimizes the facial landmark detection together with heterogeneous but subtly correlated tasks, e.g.head pose estimation and facial attribute inference.

E. Measure Metric

In our experiment, we adopt four mainly used measure metric: Normalized Mean Error(NME), Cumulative Error Distribution Curve (CED), Area Under the error Curve (AUC) and Failure rate.

Normalized Mean Error is calculated using the Euclidean Distance (L_2 norm) between estimated points and groundtruth, and being normalized by the distance of two outer eye corners. For each point, landmark-wise normalized error is calculated by:

$$e_i = \frac{\|x^e_{(i)} - x^g_{(i)}\|_2}{d_{io}}$$

where, e_i is the i-th error value, $x^e_{(i)}$ is the i-th estimated points, $x^g_{(i)}$ is the i-th ground truth, and d_{io} is IOD, the inter-ocular distance, i.e. Euclidean distance between two outer eye corners.

For every face, sample-wise NME is calculated by summarizing the normalized errors of all facial points by:

$$e = \sum_{i=1}^{n} e_i$$

where, e is the error value of face, n is the number of facial landmarks. Since the groudtruth of MOBIO faces contains 22 landmarks, the facial landmark detection methods we choose can detect different numbers of facial feature points (i.e., 5, 19, and 68), in our experiment, we choose the overlapped facial landmarks of groundtruth points and detected points. As shown in Figure 9, there are 5 (in 5), 16 (in 19), and 15 (in 68) overlapped points (blue dots) are used for evaluation.

The overall normalized mean error is computed by:

$$error = \frac{\sum_{i=1}^{m} e}{m}$$

where, $error$ is the overall normalized mean error, m is the number of faces.

CED is the cumulative distribution function of normalized errors, which evaluates the fraction of facial landmarks changes as error threshold changes. It is a better way to handle outliers. In our experiment, we set the error value threshold as 0.08 and 0.1. We partition the error value range [0, 0.08] or [0, 0.1] into 80 or 100 segments with equal step size 0.001. For each error value point X, the fraction of face images whose error value is ¡= X is calculated.

AUC means the area under the error curve CED:

$$AUC_\alpha = \int_0^\alpha f(e)de$$

where, e is normalized error, $f(e)$ is cumulative error distribution function, α is the upper bound used to calculate the define integration. In our experiment, α is set as 0.08 and 0.1.

Failure Rate is to count the fraction of faces whose error value is greater than error value threshold, in our experiment, 0.08 and 0.1 too.

V. Experimental Results

This section, we describe the details of experiment implementation first, including details of each method, and then give a through evaluation result of these method on the generated mobile face images, finally provide a thorough comparison of these methods on other databases, e.g., 300W, AFW, AFLW, COFW.

A. Implementation Details

Seven facial landmark detection methods are selected. Table I gives the detailed experiment information of these models. Some models [43], [51]–[53], [56] can deal with original images directly, and some others need inputting square faces [40], [54], [55] with fixed size. Most models adopt MTCNN as face detector before facial landmark detection. Different models

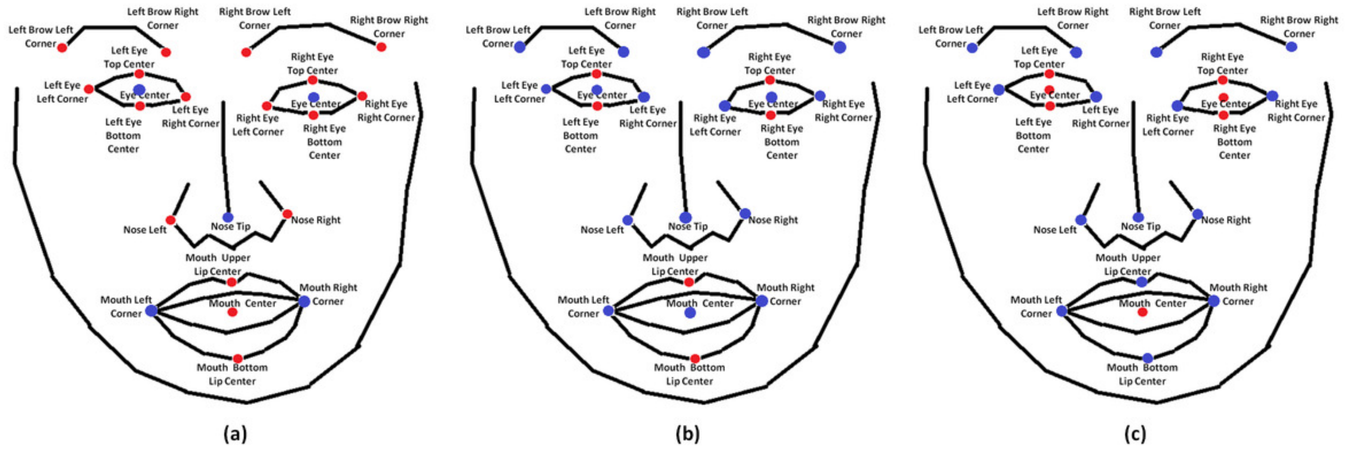

Fig. 9. Chosen (a)5, (b)16, (c)15 overlapped facial landmarks (blue dots) from 5, 19, 68 points for evaluation.

TABLE I
INFORMATION OF SELECTED FACIAL LANDMARK MODELS.

Method	Face Detector	Input Size	Output Size	#Detected Landmarks	#Used Landmarks	#Detected Faces
Tweaked CNN [54]	-	256*256	40*40	5	5	20,481
WingLoss [40]	MTCNN	256*256	256*256	19	16	20,481
DAC-CSR [51]	MTCNN	640*480	640*480	19	16	20,481
PA-CNN [52]	-	640*480	640*480	68	15	19,905
CE-CLM [53]	MTCNN	640*480	640*480	68	15	20,487
ECT [55]	-	256*256	256*256	68	15	20,481
TCDCN [43]	MTCNN	640*480	640*480	68	15	20,481

can detect different numbers of faces. So during testing, only the visible landmarks are involved in the evaluation. For each comparison we use the biggest set of overlapping landmarks. For example, Tweaked CNN [54] detects 5 landmarks, and the biggest overlapping set with groundtruth (22 landmarks) is 5. WingLoss [40] and DAC-CSR [51] detect 19 landmarks, and finally 16 landmarks are used. PA-CNN [52], CE-CLM [53], ECT [55], and TCDCN [43] adopt 15 landmarks in 68 for evaluation.

B. Evaluation on Mobile Still Face Images

Seven models are performed facial landmark detection on our generated face data based on MOBIO. Table II provides the normalized mean error, AUC and failure rate when the thresholds are set to 0.08 and 0.1. One can see WingLoss [40] gains the lowest mean error and greatest AUC as the threshold is equal to 0.08 and 0.1. TCDCN [43] gains the greatest mean error and the lowest AUC as the threshold is equal to 0.08 and 0.1. CE-CLM [53] obtains the smallest failure rate when the failure rate is defined by the percentage of test images with more than 8% detection error. ECT [55] obtains the smallest failure rate when the failure rate is defined by the percentage of test images with more than 10% detection error. Figure 10 gives the CED curve of all models. Figure 10(a) shows the curves with threshold as 0.08, and (b) as 0.1. One can see the performance in Figure 10(a) is similar to those in Figure 10(b).

Although there is much ongoing research in computer vision approaches for face alignment, varying evaluation protocols, lack descriptions of critical details and the use of different experimental setting or datasets makes it hard to shed light on how to make an assessment of their cons and pros, and what are the important factors influential to performance. We try our best to make a through performance comparison of the selected methods on our data and other databases, such as 300W, AFLW, AFW, and COFW.

Tweaked CNN [54] computes NME values on AFW and AFLW by normalizing the mean distance between predicted to ground truth landmark locations to a percent of the inter-ocular distance. It detects 5 facial feature points for each dataset. Tweaked CNN also detects 49 and 68 points on 300W and calculates AUC and failure rate with threshold as 0.1. For 49 points, AUC is 0.817 and failure rate is 1.17%. For 68 points, the AUC is 0.771 and failure rate is 1.95%. Both values of AUC are greater than that on our data (39.29%), and both failure rates are less than that on our data(9.46%).

WingLoss [40] is evaluated on AFLW and 300W via calculating NME. For the AFLW dataset, AFLW-Full protocol is adopted, and the width (or height) of the given face bounding box as the normalization term. 1.65% is gained finally, which is much lower than that on our data. For 300W dataset, the NME uses the inter-pupil distance as the normalization term, and the face images involved in the 300W dataset have been semi-automatically annotated by 68 facial landmarks. The final size of the test set is 689. The test set is further divided into two subsets for evaluation, i.e. the common and challenging subsets. The common subset has 554 face images from the LFPW and HELEN test subsets and the challenging subset

TABLE II
EVALUATION RESULTS OF FACIAL LANDMARK DETECTION ON DEEP MODELS

Method	Normalized Mean Error	Threshold=0.08		Threshold=0.10	
	(10^{-2})	AUC (%)	Failure Rate (%)	AUC (%)	Failure Rate (%)
Tweaked CNN [54]	6.4739049	27.533598	19.334993	39.288243	9.462429
WingLoss [40]	**3.8777522**	**54.384399**	1.904204	**63.232557**	1.010693
DAC-CSR [51]	4.6757547	45.475898	5.849324	55.507959	3.251794
PA-CNN [52]	5.7171261	29.574564	4.029736	43.333145	0.630608
CE-CLM [53]	4.7493759	42.482611	**0.990872**	53.840948	0.536926
ECT [55]	5.0704699	38.226405	1.079049	50.450906	**0.502905**
TCDCN [43]	6.5829441	21.545304	13.290367	35.863483	3.071139

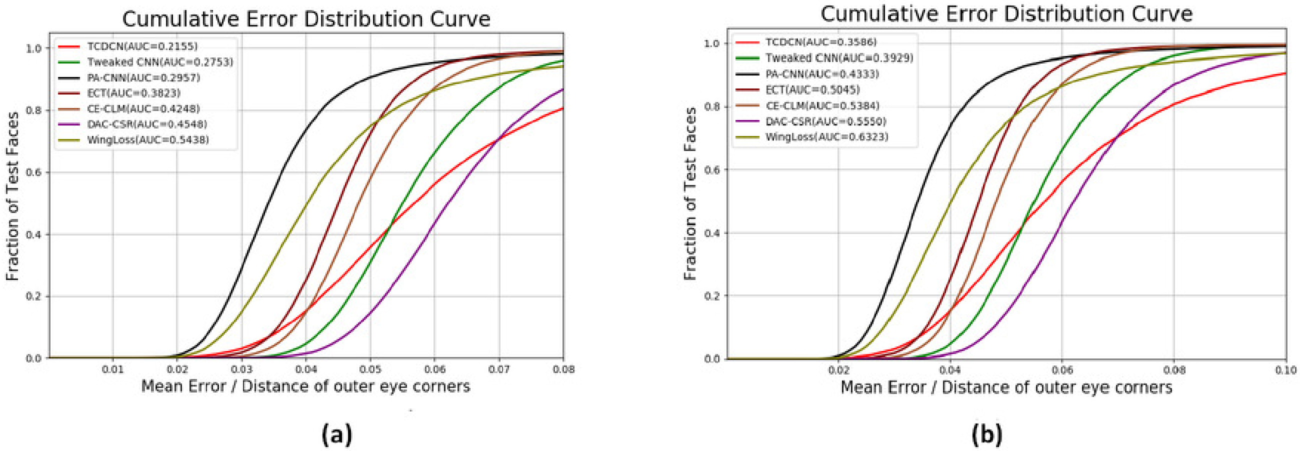

Fig. 10. CED comparison of all models as error threshold is (a) 0.08, and (b) 0.10.

constitutes the 135 IBUG face images. 3.27% is gained finally on Common set which is lower than our faces (3.88%)

DAC-CSR [51] is evaluated on AFLW and COFW. For AFLW, 19 landmarks per image without the two ear landmarks are opted. And two protocols (i.e., AFLW-full, AFLW-frontal) are used. AFLW-full uses 4,386 images for test and AFLW-frontal uses 1,165. The performance is measured in terms of the average error, normalized by face size. 2.27% and 1.81% are obtained on AFLW-full and AFLW-frontal, which are lower than that on MOBIO (4.68%).

PA-CNN [52] evaluates the alignment accuracy on 300w by the mean error, which is measured by the distances between the predicted landmarks and the groundtruth, normalized by the inter-pupil distance. The 300w is divided into three sets, i.e., Common, Challenging and Fullset, and 4.82%, 9.80%, and 5.79% mean errors are gained. In them, the mean error on Common set is lower than ours (5.72%).

CE-CLM [53] is also evaluated on the typical split Common set of 300w by NME, and gains 3.14% and 2.30% with outline (68) and without outline (49) separately, which are much lower than ours (4.75%).

ECT [55] evaluated its performance on four databases (300w, AFLW, AFW, and COFW). NME (%), AUC, and/or Failure Rate (%) are calculated. The evaluation on 300W consists of two parts. The first part is conducted on the 300W test set provided officially by the 300W competition. The second part of the evaluation is performed on the fullset of 300W which is widely used in the literature. The error is normalized by the distance of outer corners of the eyes. Failure rate is calculated with the threshold set to 0.08 for the normalized point-to-point error. The AUC and failure rate are 45.98% and 3.17% with 68 points, which are higher than ours (38.23%, 1.08%), and 58.26% and 1.17% with 51 points on the test set of 300W competition which are higher too. And the NME are 4.66%, 7.96%, and 5.31% on the Common subset, Challenging subset, and Fullset of 300W. In them, the mean error on Common set is lower than ours (5.07%).

The evaluation on AFLW-PIFA is performed by NME, which are 3.21% and 3.36% with 21 and 34 points. Both are lower than ours(5.07%).

ECT [55] picked out 6 visible landmarks for evaluation on AFW. For NME, the normalized distance is the square root of the bounding box size provided in the AFW dataset. Finally, 2.62% is obtained, which is lower too (5.07%).

TCDCN [43] is evaluated on 300W, AFLW, AFW and COFW using NME and failure rate. The mean error is measured by the distances between estimated landmarks and the ground truths, normalizing with respect to the inter-ocular distance. Mean error larger than 10% is reported as a failure. And the NME on Common Subset, Challenging Subset, and Fullset of 300W are 4.80%, 8.60%, and 5.54%. In them, the NME on Common Subset and Fullset are lower than ours (6.58%).

Based on the abovementioned comparison, one can see it is a little bit difficult to tell clearly on which database the selected method perform better due to different measure metrics, and settings. However, in most cases, our mobile face images are more challenging than existing still face images. Our face data can be a new database for facial landmark detection evaluation with 22 facial landmarks as grountruth.

VI. Discussion and Conclusion

MOBIO is a mobile biometrics database captured almost exclusively using mobile phones, which provides a challenging test-bed both for face verification, speaker verification, and bi-modal verification. In this paper, we generate a mobile still face database with 20,600 images based on the MOBIO database and manually label all faces with 22 facial landmarks as groundtruth. Seven state-of-the-art facial landmark detection methods are adopted to evaluate their performance on these 20,600 face images. A thorough analysis about the result and the comparison on other databases are given too. The result shows that our dataset is a pretty challenging one for facial landmark detection.

VII. Acknowledgments

This work was partly supported by a NSF-CITeR grant and a WV HEPC grant. The authors would like to thank the editors and the anonymous reviewers for the comments and suggestions to improve the manuscript.

References

[1] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernockỳ, N. Poh, J. Kittler, A. Larcher, C. Levy *et al.*, “Bi-modal person recognition on a mobile phone: using mobile phone data,” in *2012 IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2012, pp. 635–640.

[2] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[3] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, “Pose-aware face recognition in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4838–4846.

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[5] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, “Neural aggregation network for video face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4362–4371.

[6] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.

[7] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[8] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, “Copula ordinal regression for joint estimation of facial action unit intensity,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4902–4910.

[9] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[10] P. Dou, S. K. Shah, and I. A. Kakadiaris, “End-to-end 3d face reconstruction with deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5908–5917.

[11] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, “3d morphable face models and their applications,” in *International Conference on Articulated Motion and Deformable Objects*. Springer, 2016, pp. 185–206.

[12] P. Huber, P. Kopp, W. Christmas, M. Rätsch, and J. Kittler, “Real-time 3d face fitting and texture fusion on in-the-wild videos,” *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 437–441, 2016.

[13] G. Hu, F. Yan, J. Kittler, W. Christmas, C. H. Chan, Z. Feng, and P. Huber, “Efficient 3d morphable face model fitting,” *Pattern Recognition*, vol. 67, pp. 366–379, 2017.

[14] J. Roth, Y. Tong, and X. Liu, “Adaptive 3d face reconstruction from unconstrained photo collections,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4197–4206.

[15] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin, “Gaussian mixture 3d morphable face model,” *Pattern Recognition*, vol. 74, pp. 617–628, 2018.

[16] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, “Hierarchical temporal graphical model for head pose estimation and subsequent attribute classification in real-world videos,” *Computer Vision and Image Understanding*, vol. 136, pp. 128–145, 2015.

[17] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2879–2886.

[18] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, and Q. Wang, “Facial action unit event detection by cascade of tasks,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2400–2407.

[19] B. Martinez and M. F. Valstar, “Advances, challenges, and opportunities in automatic facial expression recognition,” in *Advances in face detection and facial image analysis*. Springer, 2016, pp. 63–100.

[20] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[22] D. Cristinacce and T. F. Cootes, “Feature detection and tracking with constrained local models.” in *Bmvc*, vol. 1, no. 2. Citeseer, 2006, p. 3.

[23] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[24] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[25] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[26] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” *arXiv preprint arXiv:1606.06259*, 2016.

[27] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[28] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment at 3000 fps via regressing local binary features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[29] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment,” in *European conference on computer vision*. Springer, 2014, pp. 1–16.

[30] S. Zhu, C. Li, C. Change Loy, and X. Tang, “Face alignment by coarse-to-fine shape searching,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4998–5006.

[31] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Learning deep representation for face alignment with auxiliary attributes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 918–930, 2015.

[32] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models-their training and application,” *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[33] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[34] Y. Wu, C. Gou, and Q. Ji, “Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3471–3480.

[35] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 160–169.

[36] Y. Wu and Q. Ji, “Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3400–3408.

[37] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X.-J. Wu, “Random cascaded-regression copse for robust facial landmark detection,” *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 76–80, 2014.

[38] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[39] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu, “Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3425–3440, 2015.

[40] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2235–2245.

[41] A. Jourabloo and X. Liu, “Pose-invariant 3d face alignment,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3694–3702.

[42] D. Lee, H. Park, and C. D. Yoo, “Face alignment using cascade gaussian process regression trees,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4204–4212.

[43] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in *European conference on computer vision*. Springer, 2014, pp. 94–108.

[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[45] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,” in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2144–2151.

[46] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, “Localizing parts of faces using a consensus of exemplars,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[47] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, “Interactive facial feature localization,” in *European conference on computer vision*. Springer, 2012, pp. 679–692.

[48] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, “Robust face landmark estimation under occlusion,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[49] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, “Xm2vtsdb: The extended m2vts database,” in *Second international conference on audio and video-based biometric person authentication*, vol. 964, 1999, pp. 965–966.

[50] S. Marcel, C. McCool, P. Matějka, T. Ahonen, J. Černockỳ, S. Chakraborty, V. Balasubramanian, S. Panchanathan, C. H. Chan, J. Kittler *et al.*, “On the results of the first mobile biometry (mobio) face and speaker verification evaluation,” in *International Conference on Pattern Recognition*. Springer, 2010, pp. 210–225.

[51] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, “Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2481–2490.

[52] K. He and X. Xue, “Facial landmark localization by part-aware deep convolutional network,” in *Pacific Rim Conference on Multimedia*. Springer, 2016, pp. 22–31.

[53] A. Zadeh, T. Baltrusaitis, and L.-P. Morency, “Convolutional experts network for facial landmark detection,” in *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, vol. 3, no. 5, 2017, p. 6.

[54] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, “Facial landmark detection with tweaked convolutional neural networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 3067–3074, 2017.

[55] H. Zhang, Q. Li, Z. Sun, and Y. Liu, “Combining data-driven and model-driven methods for robust facial landmark detection,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2409–2422, 2018.

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