A semi-automatic methodology for facial landmark annotation

Christos Sagonas[1], Georgios Tzimiropoulos[1,2], Stefanos Zafeiriou[1] and Maja Pantic[1,3]
[1] Comp. Dept., Imperial College London, UK
[2] School of Computer Science, University of Lincoln,U.K.
[3] EEMCS, University of Twente, The Netherlands
{c.sagonas, gt204, s.zafeiriou, m.pantic}@imperial.ac.uk

Abstract

Developing powerful deformable face models requires massive, annotated face databases on which techniques can be trained, validated and tested. Manual annotation of each facial image in terms of landmarks requires a trained expert and the workload is usually enormous. Fatigue is one of the reasons that in some cases annotations are inaccurate. This is why, the majority of existing facial databases provide annotations for a relatively small subset of the training images. Furthermore, there is hardly any correspondence between the annotated landmarks across different databases. These problems make cross-database experiments almost infeasible. To overcome these difficulties, we propose a semi-automatic annotation methodology for annotating massive face datasets. This is the first attempt to create a tool suitable for annotating massive facial databases. We employed our tool for creating annotations for MultiPIE, XM2VTS, AR, and FRGC Ver. 2 databases. The annotations will be made publicly available from http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/. *Finally, we present experiments which verify the accuracy of produced annotations.*

1. Introduction

Various aspects of face analysis (face detection, facial point detection, face and facial expression recognition etc) are among the most popular and well-studied areas in computer vision. Face alignment plays arguably the most important role [6]. In the last decade the most successful face alignment methods are based on deformable models. In general, deformable models represent the variation in shape or appearance of the target object (e.g. human face). Deformable models involve the representation of a template and fitting the template to a new image. Various deformable models have been proposed for model-based face analysis. The most well-known are the Active Shape Models (ASMs) [2], the Active Appearance Models (AAMs) [1, 8], and the Constrained Local Models (CLMs) [3, 12].

Training the aforementioned methods requires a facial database to be carefully developed and annotated. Existing facial databases [5, 9, 10, 7], cover large variations including: different subjects, poses, illumination, occlusions etc. However, the provided annotations appear to have several limitations (Figure 1). **(1)** The majority of existing databases provide annotations for a relatively small subset of the overall images (MultiPIE [5], AR [7]). **(2)** The accuracy of provided annotations in some cases is not so good (probably due to human fatigue, XM2VTS [9]). **(3)** The annotation model of each database consists of different number of landmarks (MultiPIE [5], XM2VTS [9], AR [7] FRGC Ver.2 [10]).

One way to extent and correct the annotations is by manual labour. However, this is not a trivial task, as it requires a human trained to perform this task. Finally, due to human factors (such as fatigue etc) the produced annotations could be still inaccurate. Therefore, (semi-)automatic annotation systems are needed.

Let us assume that we have a cohort of annotated images. In order to create annotations for the non-annotated images, we need to combine their instances. This can be done using generative models such as AAMs. One of the main advantages of generative models is that they can be naturally used to generate novel instances. For example training with images from one view with 'Neutral' expression (e.g. pose $+15^0$) and expressive images from another view (e.g. 'Frontal'), one can generate a model and use this model to fit expression in different view. However, the fitting procedure is a very tedious task and most methods for fitting deformable model do not generalize well to previously unseen images. One of the AAM variants that generalizes well to unseen variations is the recently introduced Active Orientation Models (AOMs). In [13] it was shown that AOMs outperform, by a large margin, the state-of-the-art discriminative deformable models proposed in [12] and [14]. The success of AOMs in generative face fitting motivated the

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.132

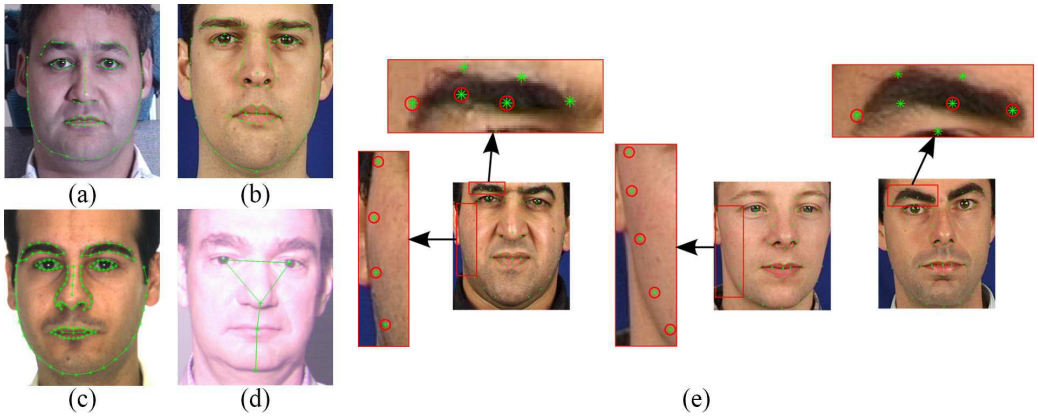

Figure 1. (a)-(d) Annotated images from MultiPIE, XM2VTS, AR, FRGC Ver.2 database, and (e) examples from XM2VTS with inaccurate annotations.

development of the proposed tool.

In this paper we propose a semi-automatic annotation tool which can be applied for annotating in a time efficient manner massive facial databases. The proposed tool was applied to create the annotations for MultiPIE [5], XM2VTS [9] and FRGC Ver. 2 [10] and AR [7] databases and is based on the recently introduced AOMs.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of the existing facial databases. Section 3 gives a brief review of AOMs and details about the proposed tool. The produced results are presented and disgusted in Section 4.

2. Existing Databases

The past twenty years the research community has collected a number of facial databases the most popular ones of which are MultiPIE (used for face recognition, facial expressions recognition, facial deformable models), XM2VTS (used for face recognition / verification and building facial deformable models), FRGC Ver. 2 (used for face recognition), and AR (used for face and facial expressions recognition). In the following section we provide an overview of the above databases, and comment on the available mark-ups they provide. As it becomes evident, the majority of these databases provide annotations for a relatively small subset of the overall images. The accuracy of the provided annotations might be limited and the annotation model of each database consists of different mark-ups. We address all of these limitations in our work.

2.1. CMU MultiPIE

The CMU Multi Pose Illumination, and Expression (MultiPIE) Database [5] contains around 750,000 images of 337 subjects captured under laboratory conditions in four different sessions. For each subject there are available images for 15 different poses, 19 illumination conditions and 6 different expressions (Neutral, Scream, Smile, Squint, Surprise, Disgust). The accompanying facial landmark annotations consist of a set of 68 points for images in the range $-45^0 : 45^0$ (Figure 1(a)) and 39 points for profile images. The provided annotations correspond to $9,3\%$ of the available images, only.

2.2. XM2VTS

The Extended Multi Modal Verification for Teleservices and Security applications (XM2VTS) [9] database contains 2,360 frontal images of 295 different subjects. Each subject has two available images for each of the four different sessions. All subjects are captured under the same illumination conditions and in the majority of images the subject displayed a neutral expression. Facial landmark annotations are available for the whole database [11]. Each annotation consists of 68 points (Figure 1(b)). As we may see, the accuracy of the annotations in some cases is limited. Also, the provided points do not correspond to the same points provided by MultiPIE.

2.3. AR

AR [7] contains 4,000 images of 126 different subjects captured in two sessions. Each subject has up to 26 images taken in two sessions. The first session contains 13 images with different illumination conditions, facial expressions (Neutral, Smile, Anger, Scream) and occlusions (sun glasses and scarf). The provided annotations contain 130 landmark points (Figure 1(c)) and correspond to a subset of only 896 images from 60 subjects without occlusions nor different illumination conditions.

2.4. FRGC Ver. 2

The Face Recognition Grand Challenge (FRGC) Version 2.0 database [10] consists of 4,950 face images of 466 different subjects. Each subject session consists of images

taken under well-controlled conditions (i.e., uniform illumination, high resolution) and images taken under fairly uncontrolled ones (i.e., non-uniform illumination, poor quality). The provided annotations consist of 5 landmark points (Figure 1 (d)) only.

3. A tool for semi-automatic database annotation

We begin with a brief review of Active Orientation Models (AOMs). We describe how they are constructed and describe the fitting procedure. Finally, we present the proposed tool for semi-automatic database annotation.

3.1. Active Orientation Models

AOMs are a variant of AAMs [8]. They are different in the appearance model used, as well as in the fitting and parameter estimation procedures. AOMs were chosen due to their good generalization properties with unseen images.

An AOM is defined by the shape, appearance and motion model. A shape model in AOMs represents the point distribution variance observed in a training set. Let us suppose that the training set consists of D annotated images. N fiducial points are provided for each image. Each mesh of the object can be then represented as:

$$\mathbf{s} = (x_1, y_1, ..., x_N, y_N)^T, \tag{1}$$

where x, y are the coordinates of each fiducial point. By applying 'Procrustes Analysis' we align the training shapes into a common frame and remove the similarity transformations (scale, rotation, translation). Finally a Principal Component Analysis (PCA) is applied on the aligned shapes. The result is n eigen-shapes $\mathbf{s}_i$, corresponding to the n largest eigenvalues λ_S^i. Therefore, every shape $\mathbf{s}$ can be expressed as a mean shape $\mathbf{s}_0$ plus a linear combination of n eigen-shapes $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{\Phi}_S \mathbf{p}, \tag{2}$$

where $\mathbf{\Phi}_S = (\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$ and the coefficients $\mathbf{p} = (p_1, p_2, ..., p_n)$ are the eigen-shapes and the shape model parameters, respectively.

The appearance model of an AOM is defined in the mean shape $\mathbf{s}_0$ using a warping function (motion model). To this end, all training textures are warped to the mean shape $\mathbf{s}_0$. In the majority of cases the motion model is a piece-wise affine warp. The results of the warping of each image is an image $\mathbf{g}(\mathbf{x})$ containing the warped pixels $\mathbf{x}' \in \mathbf{x}_0$. Subsequently, the normalized gradients of the warped textures $\mathbf{z}(\mathbf{x})$ are computed. Finally, a PCA is applied on the matrix $\mathbf{Z}$ the columns of which are the shape-free normalized gradients of the training images. PCA produces m eigen-images corresponding to the m largest eigenvalues λ_Z^i. Every appearance $\mathbf{z}$ can then be expressed as:

$$\mathbf{z} = \mathbf{\Phi}_Z \boldsymbol{\lambda}, \tag{3}$$

where $\mathbf{\Phi}_Z = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m)$ are the appearance eigen-images and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)$ are the coefficients of the appearance model parameters.

The fitting procedure of AOMs is modelled as the following optimization problem:

$$\{\mathbf{p}_0, \boldsymbol{\lambda}_0\} = arg \max_{\{\mathbf{p}, \boldsymbol{\lambda}\}} \frac{\mathbf{z}[\mathbf{p}]^T \mathbf{Z}^T \boldsymbol{\lambda}}{||\mathbf{\Phi}_Z \boldsymbol{\lambda}||}. \tag{4}$$

In [8] it was shown how optimization problem (4) can be solved efficiently using both inverse compositional alternating optimization and project out algorithms.

3.2. Proposed Tool

Let $\mathcal{DB}$ be a database consisting of N_{subj} different subjects. We assume that for each subject, images from different expressions $E_j, j \in \{1, ..., N_{exp}\}$, and poses $P_k, k \in \{1, ..., N_{pos}\}$ are available. Let $\mathcal{V}$ be a subset of annotated images and $\mathcal{U}$ a subset of non-annotated images. The goal of our tool is to **(1)** produce annotations for the subjects in $\mathcal{V}$ but having different expressions and poses in $\mathcal{U}$ and **(2)** produce annotations for subjects outside $\mathcal{V}$. For example in MultiPIE the annotations for subjects with expressions 'Disgust' at 0^0 and 'Neutral' at 15^0 are provided and we want to produce the annotations for subjects with expression 'Disgust' at 15^0. In this case the annotated and non-annotated subsets are defined as: $\mathcal{V} = \{E_j, P_k, E_{j+1}, P_{k+1}\} = \{\text{'Disgust'}, 0^0, \text{'Neutral'}, 15^0\}$ and $\mathcal{U} = \{E_j, P_{k+1}\} = \{\text{'Disgust'}, 15^0\}$ respectively.

Algorithm 1 Semi-automatic database annotation tool

Require: Annotated subset $\mathcal{V}$
Non-annotated subset $\mathcal{U}$
Ensure: Annotations of $\mathcal{U}$
1: Train an AOM using the set $\mathcal{V}$
2: Apply the landmark detector on images of $\mathcal{U}$.
3: Use the results from detector as initialization and fit the AOM to $\mathcal{U}$
4: Classify manual the fittings to 'Good' $\overline{\mathcal{U}}$ and 'Bad' $\mathcal{B} = \mathcal{U} - \overline{\mathcal{U}}$
5: **while** $\mathcal{B} \neq \emptyset$ **do**
6: Train an AOM using the $\overline{\mathcal{U}}$.
7: Fit the AOM to $\mathcal{B}$.
8: Classify manual the fittings to 'Good' $\overline{\mathcal{U}}$ and 'Bad' $\mathcal{B} = \mathcal{U} - \overline{\mathcal{U}}$
9: **end while**
10: Check and correct manually the created fittings of $\mathcal{U}$.

In order to produce annotations for the images in $\mathcal{U}$ we employ AOMs. First, the annotated subset $\mathcal{V}$ is used as the

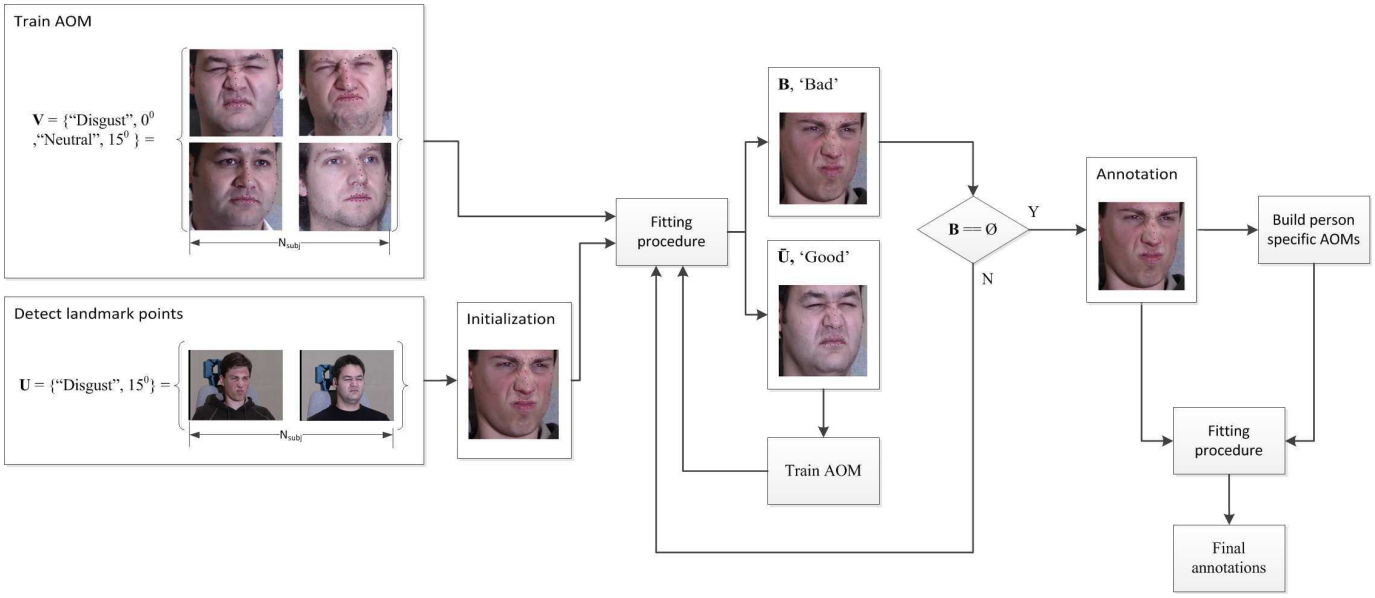

Figure 2. Flowchart of the proposed tool.

training set. Subsequently, an AOM is learned from $\mathcal{V}$ and the model is fitted in $\mathcal{U}$. The model is initialized by applying the landmark detector in [14].

The results from the fitting procedure are classified manually by the user to 'Good' and 'Bad' denoted by $\overline{\mathcal{U}}$ and $\mathcal{B} = \mathcal{U} - \overline{\mathcal{U}}$, respectively. A new AOM is learnt using $\overline{\mathcal{U}}$ and is subsequently fitted to the images of $\mathcal{B}$. This procedure is repeated until the set of 'Bad' fittings is empty, $\mathcal{B} = \emptyset$. Finally, the created fittings are checked and manually corrected (if necessary) by the user.

Finally, in order to increase the accuracy of the created annotations we built Person Specific Models (PPM) for each subject of the database. First we constructed an AOM with the annotated images of each subject. We then refit the AOM to the same images and use the vertex locations of the fitting results as new annotations [4].

The algorithm can be readily applied for providing annotations in different databases. In short, we train a model in $\mathcal{DB}_1$ (for which annotation are assumed to exist) and we fit $\mathcal{DB}_2$. Then, fittings are manually clustered into 'Bad' and 'Good'. A model for $\mathcal{DB}_2$ is trained from 'Good' fittings and the remaining are fit until convergence.

4. Results

The proposed method was used to create annotations for the databases presented in Section 2. The same landmark configuration provided by MultiPIE was adopted for all databases.

4.1. MultiPIE

The available annotations from MultiPIE cover only expressions at 'Frontal' pose and 'Neutral' expression for some subjects at $-90^0 : 90^0$ poses. Figure 2 shows the flow chart of Algorithm 1 for the case of 'Disgust' at 15^0. A total of 12.570 annotations for all expressions and poses $-30^0 : 30^0$ were generated. Examples of annotated images are depicted in Figure 5.

4.2. XM2VTS

In order to fit 'Frontal' with 'Neutral' expression images of XM2VTS we used the images with 'Neutral' expression and poses between $-15^0 : 15^0$ from MultiPIE. We first created the annotations for the first session using the proposed method. Subsequently, the produced annotations from the first session were used as input to Algorithm 1 to produce annotations for the next session. This procedure was repeated for the remaining sessions. Finally, 2.360 annotations were produced. In Figure 6 we show an example of a set of 68-points annotated images.

4.3. AR

A procedure similar to the one used for XM2VTS was used to generate annotations for the neutral images of AR. For images having a specific expression E_j, we used the annotated neutral images of AR and the images with the corresponding expression and frontal pose from the MultiPIE. Some examples of annotated images are depicted in Figure 7.

4.4. FRGC Ver. 2

Firstly, the images from MultiPIE with six expressions and poses $-15^0 : 15^0$ were used to build an AOM. For each subject of FRGC we used two images with two different illumination conditions in order to create a subset with information from all subjects. Subsequently, the annotations for this subset are created and based on these we built a new AOM. Thus, by using this model, we created the annotations for the remaining images. Figure 8 depicts some examples from the annotated images.

4.5. Evaluate the accuracy of created annotations

In order to evaluate the accuracy of the produced annotations we conducted the following experiment on MultiPIE. We randomly selected 200 images with all expressions with 0^0 pose to form the train set while a set of 50 image with all expression of unseen subjects formed the test set. The trained AOM was used to find the location of landmark points on test images. Subsequently, the test images are annotated from four expert human annotators. We refer and use as the ground truth the average of locations supplied by the first two annotators.

The basic error measurement used in this experiment is the normalized mean euclidean distance. It is defined as follows: The fitting results of ith image are represented as $2 \times n$ matrix $\mathbf{A}_i$, while the ground truth as $\mathbf{GT}_i$ corresponding to the 2D coordinates of the N fiducial points. Then, the fitting error e_A^i for the ith image is computed by:

$$e_A^i = \frac{1}{Ns_i} \sum_{j=1}^{N} ||\mathbf{A}_i(j) - \mathbf{GT}_i(j)||_2 \tag{5}$$

where s is the ground truth face size, $N = 68$, and $|| \cdot ||_2$ is the 2-norm of a vector. The error e_M^i for the manual annotated image ith is computed as:

$$e_M^i = \frac{1}{2Ns_i} \sum_{j=1}^{n} ||\mathbf{M}_i^3(j) - \mathbf{GT}_i(j)||_2 + ||\mathbf{M}_i^4(j) - \mathbf{GT}_i(j)||_2, \tag{6}$$

where M^3, and M^4 are the annotations from the last two annotators.

In order to show the power of AOMs we compared the error of AOMs in the 50 images with regards to the first two annotators. This error was then compared with the one from two independent annotators with regard to the first two. Figure 3 shows that AOM is as accurate in these 50 images as independent annotators. The mean error of annotators is 0.0262 where the corresponding mean error of AOM is 0.017. The mean error by the annotators is higher than AOM's mean error, due to annotators' fatigue and lack of characteristic feature points in some areas. For example,

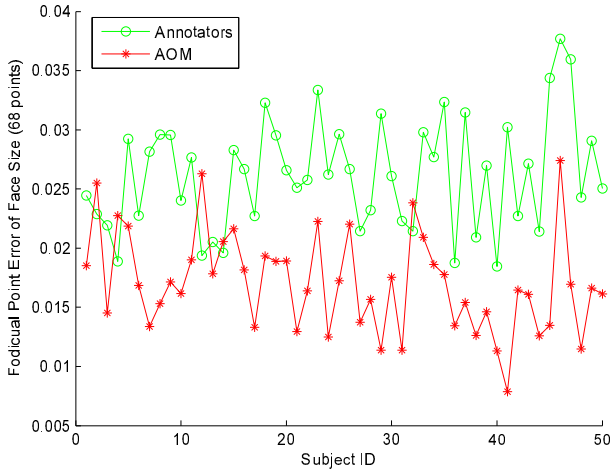

Figure 3. Fitting errors of AOM and annotators for the 50 subjects of MultiPIE.

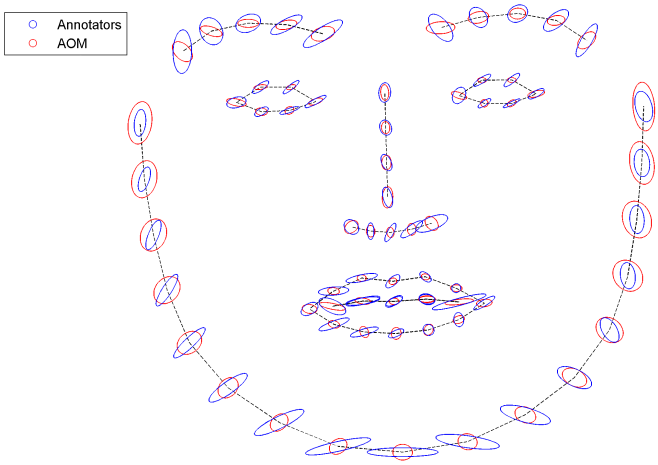

Figure 4. Each ellipse denotes the variance of each point with regards to the ground truth.

there are not obvious feature points on the chin as opposed to landmarks around the eye. Figure 4 shows the variance (plotted as ellipses) of each point for the test set with regards to the ground truth (first two annotators), both in cases of AOM and annotators (the last two annotators). As it can be seen for the majority of the points the variance from AOM is smaller from the one of the last two annotators.

5. Conclusion

We proposed a semi-automatic tool for database annotation. Using the proposed tool we produced annotations with the same model for the MultiPIE, XM2VTS, AR and FRGC databases. The annotations obtained by this tool are

Figure 5. Created annotations for subject of MultiPIE with id 2 for 6 expressions and poses $-30^0 : 30^0$.

so accurate that can be used as ground truth.

6. Acknowledgements

This work has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Georgios Tzimiropoulos is funded by the European Communitys 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG).

Figure 6. Annotations for different subjects of XM2VTS.

Figure 7. Annotated images from AR with neutral expression.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, 2001. 1

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, et al. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 1

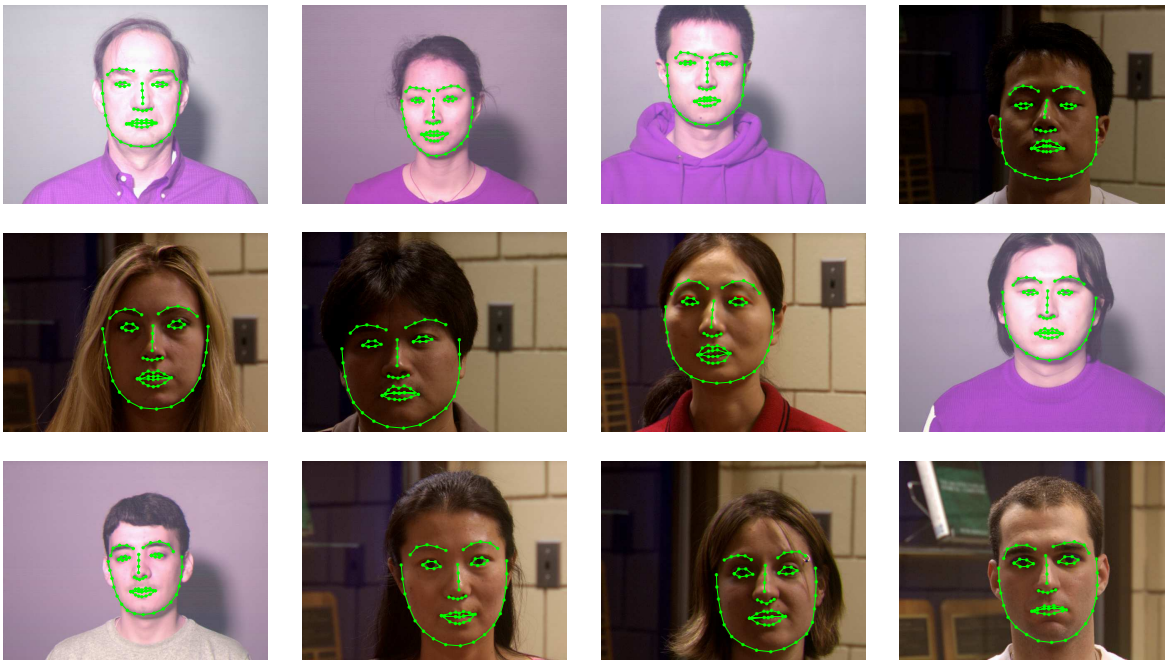

Figure 8. Annotated images from FRGC Ver. 2.

[3] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *Proc. British Machine Vision Conference*, volume 3, pages 929–938, 2006. 1

[4] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005. 4

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 1, 2

[6] B. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981. 1

[7] A. Martinez and R. Benavente. The ar face database. *CVC Technical Report*, 24, 1998. 1, 2

[8] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. 1, 3

[9] 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*, volume 964, pages 965–966. Citeseer, 1999. 1, 2

[10] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005. 1, 2

[11] E. U. projects: UFACE and FGNET. Xm2vts 68pt markup. 2

[12] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 1

[13] G. Tzimiropoulos, J. A. i medina, S. Zafeiriou, and M. Pantic. Generic active appearance models revisited. In *11th Asian Conference on Computer Vision (ACCV 2012)*, pages 650–663, Daejeon, Korea, November 2012. 1

[14] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 1, 4