2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

CG Benefited Driver Facial Landmark Localization Across Large Rotation

Liang Shi[1], Jiguang Yue[1], Yanchao Dong[1]*, Minjing Lin[1], Senbo wang[1], Runjie Shen[1] and Zhiming Chang[2]

Abstract— **Facial landmark localization is a crucial initial step Driver Inattention Monitoring. The aim of this paper is to localize driver facial landmarks across large rotation, say [-90°, +90°] in yaw rotation, to cope with real driving conditions. The paper proposes a flexible pipe-line for creating automatically labeled face image to supply wanted dataset. The benefits of CG (Computer Graphics) techniques such as 3D face modelling and morphing, photorealistic rendering and ground truth generation are utilized. To the best of our knowledge this is the first time to combine CG rendering and automatic ground truth labelling techniques with face landmark localization algorithms. The effectiveness of the CG rendered data is proved by cross validation with Multi-PIE dataset. Landmark localization across large rotation is obtained by a system simply integrating the off-the-shelves algorithms and trained with the CG rendered data. The experiments of the implemented system on Multi-PIE and real persons show that it could localize facial landmarks across large rotation accurately and in real time.**

I. Introduction

Driving inattention is a major contributor to highway crashes. Because the inattention contains many forms and it is a sophisticated internal state of the driver, it is quite difficult to detect. Driving Inattention Monitoring System has been an active research field for the last few decades. Facial landmark localization is a crucial initial step in Driver Inattention Monitoring. In the wild environment, the discriminative shape regression method for facial landmark localization has gained approval due to its good accuracy and real-time performance[1,2,3,4]. Given a test image with an initial shape, discriminative shape regression uses a cascade of pre-learned regression functions to update the shape stage by stage. Though the discriminative shape regression methods make significant progress for frontal facial landmark localization, its performance degrades dramatically when partial landmarks become invisible because of occlusions or large pose variation, which occurs very often during driving. One major reason is lack of standard training dataset across large pose variations. Such a training dataset must be general enough to represent all appearance of faces under any pose and illumination condition. But creating datasets with ground truth labels is of high cost and needs too much manual labor for labelling. The Panoptic Studio[5], a massively multiview system for social motion capture, is equipped with more than 500 synchronized cameras. The Panoptic Studio outputs more than 1T bytes data per minute. Based on the huge amount of dataset created by Panoptic Studio the OpenPose[6] achieves state-of-the-art performance, which shows the vital importance of a large scale and high quality dataset.

There are other facial image datasets in the facial image analysis community. Some datasets collect images from web and label them manually, such as ALFW [7,8] and Helen [9] datasets. AFLW provides a large-scale collection of annotated face images, exhibiting a large variety in appearance as well as general imaging and environmental conditions. In total about 25k faces are annotated with up to 21 landmarks per image. Helen consists of 2000 training and 330 test images which are high resolution face images collected from Flickr. The primary facial components are annotated accurately by hand.

Some other datasets are created in laboratory with controlled equipment. The Multi-PIE database [10], contains 755,370 images from 337 different subjects captured under 15 views and 19 illumination conditions. 6152 of the images are annotated with 39 or 68 landmark points. Other datasets created with controlled equipment include CASIA 2.0[11], Chicago face database [12], MUCT[13] and PUT[14], etc. This kind of datasets could be designed so as to meet the particular requirement. But it requires high cost hardware environment and consumes longtime to do manual labelling.

The aim of this paper is to localize driver facial landmarks across large rotation to cope with real driving conditions. The paper proposes a flexible pipe-line for creating automatically labeled face image to supply wanted dataset. The benefits of CG (Computer Graphics) techniques such as 3D face modelling and morphing, photorealistic rendering and ground truth generation are utilized. To the best of our knowledge this is the first time to combine CG rendering and automatic ground truth labelling techniques with face landmark localization algorithms. The effectiveness of the CG rendered data is proved by cross validation with Multi-PIE dataset. Landmark localization across large rotation is obtained by a system simply integrating the off-the-shelves algorithms and trained with the CG rendered data.

This work is partially supported by the Natural Science Foundation of Shanghai (No. 18ZR1442500), the Fundamental Research Funds for the Central Universities and the Science and Technology Project (No. 10.11) from China Datang Corporation Renewable Power Co., Limited.

Liang Shi (1531794@tongji.edu.cn), Jiguang Yue (yuejiguang@tongji.edu.cn), Yanchao Dong (*Corresponding Author, dongyanchao@tongji.edu.cn), Minjing Lin (1730756@tongji.edu.cn) , Senbo Wang (1410472@tongji.edu.cn) and Runjie Shen (shenrunjie@tongji.edu.cn) are with Tongji University, No.4800 Cao'an Road, Shanghai, P. R. China.

Zhiming Chang is with the Datang Guoxin Binhai Offshore Wind Power Co., Ltd.

978-1-5386-4451-5/18/$31.00 ©2018 IEEE

The rest of the paper is organized as follows: Section 2 presents the CG rendering pipe-line, where the 3D morphable model, the photorealistic rendering process and the ground truth generation are presented in details and describes the dataset capture setups, including the setups of camera, illumination and expression; Section 3 reviews the state-of-the-art landmark localization algorithms; Section 4 makes thorough experiments for evaluating the effectiveness of the proposed CG rendering pipe-line and gives a system for localizing landmarks across large rotation.

II. CG Imaging Pipeline

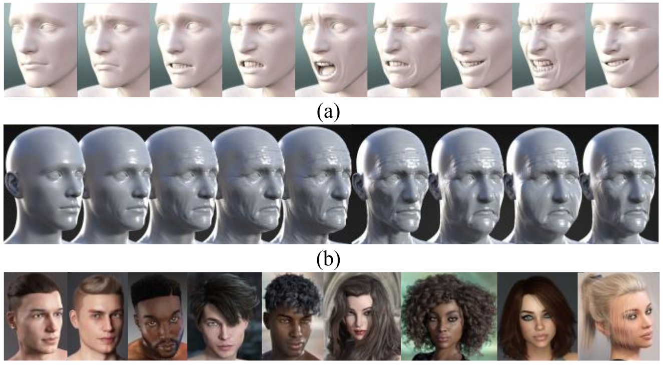
Numerous face analysis articles acknowledge the fact that 3DMM (3D Morphable Model) based face image analysis constitutes the state of the art, but note that the main obstacle resides in the complications of their construction [15,23,24]. A 3DMM consists of a parameterized generative 3D shape, and a parameterized albedo model together with an associated probability density on the model coefficients. A set of shape and albedo coefficients describes a face. Together with projection and illumination parameters a rendering of the face can be generated. This section firstly describes the 3D face models that are adopted for creating the data sets, then presents the rendering process.

A. The 3D Face Models

Daz3D is an ecosystem full of ready-to-pose 3D human figures, hair, clothing, props, scenery, lighting and cameras. The Genesis figure of Daz3D is more than just a figure but a true character engine that allows to modify and enhance them to meet various requirements. The more detailed a character is, the more realistic and life-like it becomes. The movable jaw and increased facial polygons and eye reflection mesh makes for incredibly life-like characters that truly convey emotions. Providing bases in both female and male forms gives even more control and the power to create more realistic characters. Daz3D meets the demand for generating photorealistic face images with various shapes, textures and expressions. Fig. 1 shows some samples from Daz3D.

B. Rendering

Daz3D provides excellent human models together with detailed morph controls, realistic texture and diffusion maps.



(a)

(b)

(c)

Fig. 1. The Daz3D Model. (a) A list of life-like expressions. (b) Aging morphing process. (c) Rendered photorealistic images. The Daz3D is the wanted model.

However, it is not easy to derive the ground truth label of the facial landmarks in Daz3D environment. In contrast, Blender[16], an free and open source 3D creation suite, supports the entirety of the 3D pipeline. Therefore, in our project the human models are firstly created in Daz3D with plenty of expression and appearance variations, then imported into the Blender suite where the final image is rendered and the facial landmark ground truth is labelled automatically.

Blender utilizes the Cycles[22] as a render engine. Cycles is an unbiased, physically based, path tracing rendering engine, which means that it produces an image by tracing the paths of rays through the scene. Specifically, Cycles is a backwards path tracer, which means that it traces light rays by sending them from the camera instead of sending them from light sources. Path tracing is a computer graphics method of rendering images of three dimensional scenes such that the global illumination is faithful to reality. Fundamentally, the algorithm integrates over all the illuminance arriving to a single point on the surface of an object. This illuminance is then reduced by a surface reflectance function to determine how much of it will go towards the viewpoint camera. This integration procedure is repeated for every pixel in the output image. When combined with physically accurate models of surfaces, accurate models of real light sources, and optically correct cameras, path tracing can produce photorealistic images. Path tracing naturally simulates many effects such as soft shadows, depth of field, motion blur, caustics, ambient occlusion, and indirect lighting.

Materials in Cycles define the appearance of meshes, curves and other objects. The surface shader defines the light interaction at the surface of the mesh. It specifies if incoming light is reflected back, refracted into the mesh, or absorbed. The BSDF (Bidirectional Scattering Distribution Function) of the surface shader defines how light is reflected and refracted at a surface. In this project the basic Diffuse BSDF Subsurface Scattering shaders are combined together into a more complex shading group to create realistic human skin materials. The shape of the surface may be altered by the displacement shaders in which way textures can then be used to make the mesh surface more detailed. A facial bump map is used to create displacement by using light and shadow effects. Cycles comes along with a dedicated Hair BSDF shader. This Hair BSDF shader contains two separate hair characteristics: Reflection and Transmission. Reflection controls the way hair reflects specular light. Being long and thin, hair follicles have an anisotropic-type reflection. Transmission controls the way light penetrates our hair follicles. It's an important factor in making hair believable and soft.

As illumination, the point, spot, area and sun lamps are available in Cycles and the lamp strength is specified in Watts . Cycles uses a physically correct light falloff, which makes the indoor and outdoor lighting configuration easy and life-like.

C. Capture Setup and Ground Truth Labelling

The capture setup was inspired by the system used in the Multi-PIE[10] system. To systematically capture images with varying poses and illuminations, we used a system of 13x3 cameras and 13 lights. Fig. 2 illustrates the system setup. The cameras' resolution is 500x500 pixels and its FOV is 35 degrees.

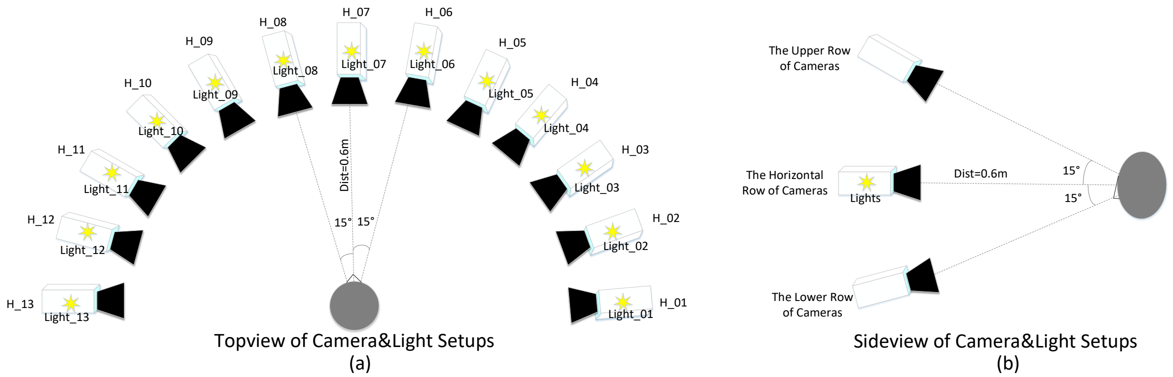


Fig. 2. Capture System Setup. (a) Top view of the cameras and lights, where lights are denoted as yellow stars. Only the horizontal row of cameras are shown in the top view subfigure. (b) Side view of the cameras and lights. The upper row and lower row of cameras point to the head and are 15° away from the horizontal row.

A row of 13 cameras locate at head height, spaced in 15° intervals horizontally, the second and third row of 13 cameras locate below and above the head height respectively. All of the cameras point to the head and the distance from the camera to the head is about 0.6 meter. Hence, the rendered face area occupies about two thirds of the whole image. The lights locate at the positions of the horizontal row cameras. The yellow stars in Fig. 2 indicate the corresponding lights.

Table.1 Expression setups. Compared with Multi-PIE datasets additional expressions such as continuous eye open, pupil movement, angry, fatigued and nervous are included.

Expressions	**Ours**	**Multi-PIE**
Continuous Eye Open	X	
Pupil movement	X	
Neutral	X	X
Smile	X	X
Surprise	X	X
squint	X	X
Disgust	X	X
Scream	X	X
Angry	X	
Fatigued	X	
Nervous	X	

A total of 200 figures of European, Asian and African are created using Daz3D for generating face images, half of whom are male and half are female. Each figure is instructed to display different facial expressions. In addition to the Multi-PIE's expression setup, we extended the expressions with continuous eye open/closure, pupil movement, angry, fatigued and nervous. The total expressions used during capture are listed in Table 1. Fig. 5 and Fig. 6 present the samples of expression setups of our dataset and the Multi-PIE dataset respectively.

During capturing one of the 200 created figure models is placed into the virtual environment with facial expression fixed. The light is turn on one by one sequentially. For each expression and illumination combination all of the cameras capture their own images respectively. Therefore, the captured images cover the variations of face shape, expression, pose and illumination. The captured image samples are shown in Fig. 3 to Fig. 5

Traditionally the facial landmarks are labelled manually, which needs long time cost and can't guarantee the label accuracy since human mistakes happen occasionally. The paper leverages the benefit of the CG virtual environment to derive the facial landmarks' ground truth automatically. As shown in Fig. 7, the CG virtual environment, the accurate 3D coordinate of every vertex of the figure model can be obtained and the poses of the cameras are known at hand. Therefore, transforming the landmark's vertex from world coordinate system to the camera's coordinate system obtains the 3D coordinate of the landmark in the camera's coordinate system. Then the landmark's 2D position in the image can be derived by projecting it onto the camera image plane using the pinhole camera model. Fig. 8 gives sample images overlaid with automatically generated landmark ground truth.

To increase the diversity of the data set various hair styles, glasses accessories and beards are added onto the figure. When placing the figure in the CG environment some random noise between [-7°, +7°] is added in to the head yaw, roll and pitch rotation respectively to make a better generalization.

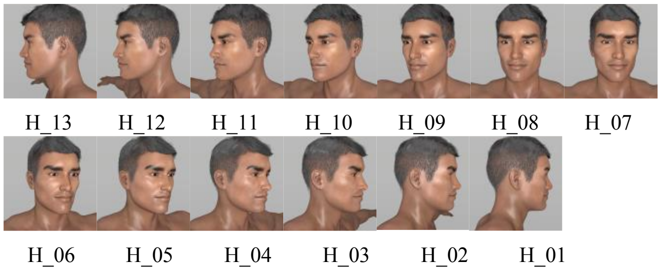


Fig. 3. Cameras' views. The "H_" prefixe stand for the horizontal row of cameras respectively. The followed number is the index of the camera in a row as shown in Fig. 2(a). The views of the upper and lower row of cameras are not depicted to save space.

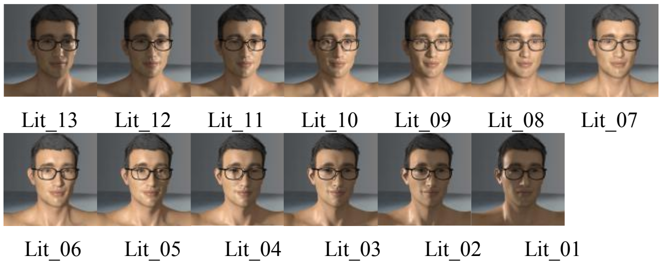


Fig.4. Illuminations effect. The images are captured while the 13 lights are turned on one by one. The name under the image is the corresponding light turned on.

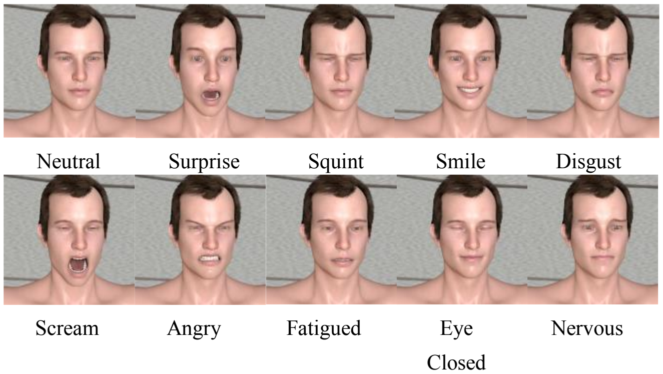


Fig. 5. Expression setup of our dataset.

Fig. 6. Expression setup of the in Multi-PIE dataset.

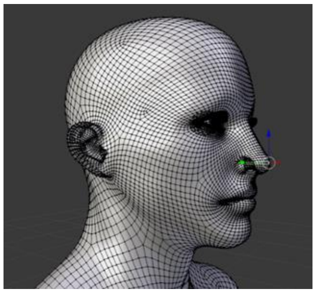


Fig. 7. The vertex of the nose tip landmark. The landmark ground truth is generated by projecting the 3D vertex onto the image plane.

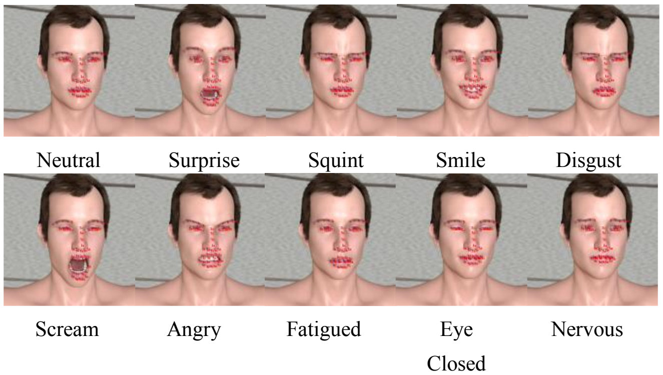


Fig. 8. Images overlaid with automatically generated landmark ground truth Landmark Localization

III. LANDMARK LOCALIZATION

A. Training of DSR

This section presents the DSR (discriminative shape regression) procedure for facial landmark localization due to its execution speed and portability.

Given N training images $\{I_i\}_{i=1}^{N}$ and the corresponding annotated shapes $\{\mathbf{S}_i^*\}_{i=1}^{N}$ with $\mathbf{S}_i^* \in \mathbb{R}^{1\times N_p}$, where N_p is the number of landmarks. The training procedure of DSR can be summarized as following steps, for details please refer to [17]:

Step 1. Training Data Augmentation. Each image in the training data is initialized by randomly sampling multiple shapes of other annotated images, the training samples can be expressed by triplets of face image, initial shape estimation and target shape. The triplet can be represented as $(I_{\pi_i}, \mathbf{S}^{(0)}_{(\pi_i,l)}, \mathbf{S}^*_{\pi_i})$, where $\pi_i \in \{1,\dots,N\}$ and $\mathbf{S}^{(0)}_{(\pi_i,l)} \in \{\mathbf{S}^*_1,\dots,\mathbf{S}^*_N\}\backslash \mathbf{S}^*_{\pi_i}$ ($l = 1,\dots,L$, where L is the number of augmentation). By randomly selecting other annotated shapes as the initial training shapes, one training image can produce different L triplets, this can be viewed as an augmented process, and the total number of these augmented samples is $N_{aug} = N \times L$.

Step 2. Feature Mapping. The shape-indexed feature is generated by the feature mapping function $\Phi^{(t-l)}_{(\pi_i,l)} = f(I_{\pi_i}, \mathbf{S}^{(t-l)}_{(\pi_i,l)})$, where $\Phi^{(t-l)}_{(\pi_i,l)} \in \mathbb{R}^{1\times f}$, f is the feature dimensionality. SIFT features is adopted as the feature mapping function in DSR. It is a kind of local feature that extract only the local region feature coordinated with the facial landmarks.

Step 3. Regressor Learning. A regressor in stage t is learned by minimizing the error between the estimated shape $\mathbf{S}^{(t-l)}_{(\pi_i,l)}$ and ground truth shape $\mathbf{S}^*_{\pi_i}$ in image I_{π_i} as:

$$r_t = \arg\min \sum_{\pi_i=1}^{N}\sum_{l=1}^{L}\left\| \mathbf{S}^*_{\pi_i} - \left(\mathbf{S}^{(t-l)}_{(\pi_i,l)} + r(\Phi^{(t-l)}_{(\pi_i,l)})\right)\right\|^2 \quad (1)$$

The form of $r(\cdot)$ depends on specific implementation.

Step 4. Shape Update.

$$\mathbf{S}^{(t)} = \mathbf{S}^{(t-1)} + r_t\left(\tilde{\Phi}^{(t-l)}\right) \quad (2)$$

where $\mathbf{S}^{(t-1)} = \{\mathbf{S}^{(t-l)}_{(\pi_i,l)}\}_{\pi_i=1,\dots,N;l=1,\dots,L} \in \mathbb{R}^{N_{aug}\times 2N_p}$.

Step 2–Step 4 iterate in a gradient boosting framework until $S^{(t)}$ converged to the target shapes.

B. Testing of DSR

Given a new image I with an initial shape $S^{(0)}$, the landmark is regressed by the learned cascaded regressor r_t stage by stage as:

$$\mathbf{S}^{(t)} = \mathbf{S}^{(t-1)} + r_t(f(I, \mathbf{S}^{(t-1)})) \quad (3)$$

where I is the input image, $\mathbf{S}^{(t)} = [x_1^{(t)}, x_2^{(t)}, \dots, x_{N_p}^{(t)}, y_1^{(t)}, y_2^{(t)}, \dots, y_{N_p}^{(t)}]^T$ is the shape with N_p facial landmarks in I at stage t, $f(\cdot)$ is the shape-indexed feature mapping function depends on both image I and previous estimated shape $\mathbf{S}^{(t-1)}$, r_t is the regression function in stage t and $t = 1,\dots,T$ is the number of cascade level.

The success of discriminative regression method is mainly due to the following properties: (1) the shape-indexed feature in each stage makes a re-sampling at the previous estimation of the landmark location. This feature extract method compensates the effect of large appearance variations and increases the robustness and accuracy; (2) gradient boosting framework is incorporated in the training procedure of regression functions. In each stage the regression function is learned based on the previously estimated shape error and the shape-indexed feature. Thus, the output error in each stage monotonically decreases and converges in 4 or 5 stages; (3) the output of the regressor is a linear combination of training shapes which inherently guarantees the output is a reasonable face shape without any extra constraints.

In this paper, two regression-based, Ensemble of Regression Trees (ERT)[1] and Explicit Shape Regression (ESR)[18] , are used to evaluate the CG rendered dataset. ERT is a tree-based regression algorithm. During the learning process, the updated shape of each layer is calculated and the result is added to the average shape to return the estimated shape. In the concrete realization we use a 15-layer regression structure, each layer contains 500 trees, the depth of each tree is 5; The decision at each node is based on thresholding the difference of intensity values at a pair of pixels, which is a rather simple but much more powerful test since it is relative insensitive to changes in global lighting. Closer pixels are encouraged to form the pair. ESR adopts the Ferns as regression function. It is a shape-constrained algorithm that deals with large-area shape changes and small-area accuracy adjustments by using two-level boosted regression. In the implementation, we use 10 regressors in external-level and 500 regressors in internal-level, with a 400 random initial features.

IV. Experiments

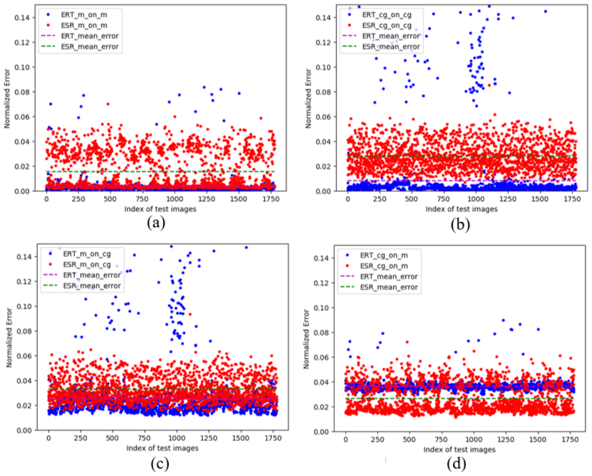
The section presents experimental evaluation on the proposed CG rendering pipe-line. Firstly, the cross validation between CG dataset and the Multi-PIE dataset is carried out to verify the effectiveness of the CG dataset on real images. Thereafter, own to the benefits of the proposed CG rendering pipe-line the paper implemented a system for facial landmark localization across large rotation by integrating off-the-shelves algorithms.

A. Cross Validation between Multi-PIE and CG rendered datasets

This subsection presents cross validation between Multi-PIE and CG rendered datasets to prove the effectiveness of the CG rendered dataset. The cross validation is evaluated using the performance of frontal facial landmark localization and face orientation classification. The Multi-PIE[10] database contains 337 subjects, captured under 15 views and 19 illumination conditions in four recording sessions. Labels are provided for a total of 6152 images. The labels have 39 or 68 feature points depending on the pose which were determined manually. Out of the total labeled images, 3,800 frontal faces are used for evaluation of landmark localization and 6,000 images are used for evaluation of face orientation classification.

The experiment is carried out using two landmark localization algorithms; (ERT and ESR) and two datasets (Multi-PIE and our CG rendered datasets). The dataset is further divided into training and test data. Therefore, to do thoroughly cross validation, eight possible configurations between the two algorithms and the two types of datasets are implemented. The 3800 labeled frontal face images from the Multi-PIE database are divided into a training dataset of 2000 images and a test dataset of 1800 images. The same amount of training and test datasets are generated using the proposed CG rendering pipe-line to make an equal comparison. The evaluation metric in all datasets is the average error between the estimated shape and the ground-truth shape. The error is normalized by the distance between two pupils as most previous works do. Fig. 9 shows the normalized errors of the eight comparison configurations and Fig.10 shows some result samples of the eight configurations.

In Fig. 9 the blue dots indicate the normalized error of the ERT algorithm while the red dots indicate that of the ESR algorithm. The naming format of the legends in the subfigures is “XXX_x_on_x”, where the “XXX” (either “ERT” or “ESR”) stands for the algorithm and the first “x” (either “cg” or “m”) stands for the dataset used for training and the second “x” (either “cg” or “m”) stands for the dataset used for test. For example, “ERT_cg_on_m” stands for the ERT algorithm is trained with CG rendered dataset and tested on Multi-PIE dataset. (a) and (b) of Fig.9 show the results of the algorithms trained and tested with the same type of datasets, while (c) and (d) of Fig. 9 show the results of the algorithms trained and tested with the different types of datasets. The errors of Fig. 9 (a) and (b) are slightly smaller than that of Fig. 9 (c) and (d) while Fig. 9 (c) and (d) have similar normalized error level, which means the CG rendered data is effective for training the model. When trained with Multi-PIE it has better performance on Multi-PIE test data than on CG rendered data and vice versa. When trained with one type data (CG rendered data, e.g.) and tested on the other type data (Multi-PIE data, e.g.) the performance degrades slightly. It seems the performance depends on the whether the training and test data are of the same type. However, as (c) and (d) shows, the generalizations of both Multi-PIE and CG rendered datasets are limited, regarding only 2000 images are utilized to train the model. If more variations and more images are added into the dataset the performance of the cross validation between Multi-PIE and CG rendered datasets will be the same as that of the model trained and tested using only one type dataset. It is in common sense that creating a very large labelled dataset manually needs high cost. In contrast, thanks to the efficiency of the CG pipe-line it is possible to create large scale labelled dataset with good generalization in low cost, therefore to get a trained model of good generalization.

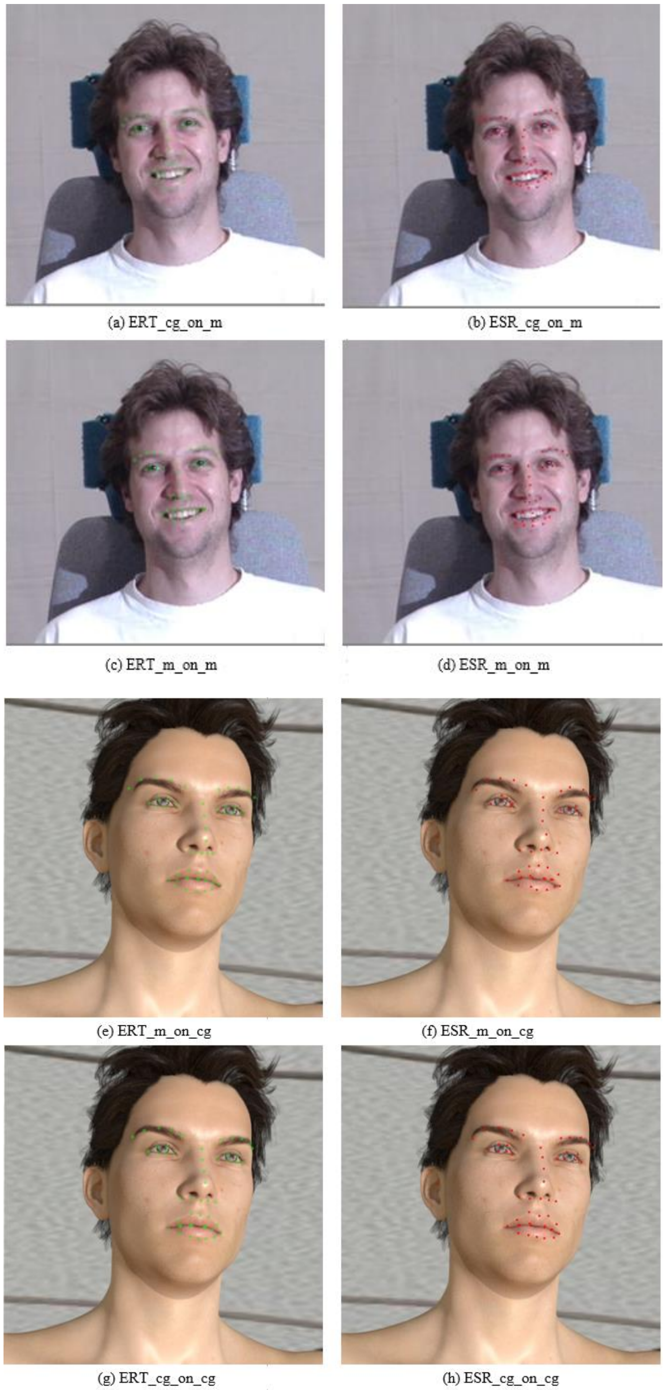


(a) (b) (c) (d)

Fig. 9. Normalized error distribution of cross validation. Two landmark localization algorithms, ERT and ESR, are utilized to do cross validation between Multi-PIE and our CG render datasets. The blue dots indicate the normalized error of the ERT algorithm while the red dots indicate the normalized error of the ESR algorithm. The naming format of the legends in the subfigures is “XXX_x_on_x”, where the “XXX” (either “ERT” or “ESR”) stands for the algorithm, the first “x” (either “cg” or “m”) stands for the dataset used for training and the second “x” (either “cg” or “m”) stands for the dataset used for test. For example, “ERT_cg_on_m” stands for the

ERT algorithm is trained with CG rendered dataset and tested on Multi-PIE dataset

Sample images of the result of cross validation with Multi-PIE dataset using ERT and ESR landmark localization algorithms are presented in Fig. 10. The naming format of the subfigures is the same as the that of the legends in subfigures of Fig. 9. The result shows that the algorithms trained either with CG rendered dataset or with Multi-PIE dataset have similar performance. Hence, it is effective to train the model with CG rendered data and test on real images.


(a) ERT_cg_on_m

(b) ESR_cg_on_m

(c) ERT_m_on_m

(d) ESR_m_on_m

(e) ERT_m_on_cg

(f) ESR_m_on_cg

(g) ERT_cg_on_cg

(h) ESR_cg_on_cg

Fig. 10. Samples of the result of cross validation with Multi-PIE dataset using ERT and ESR landmark localization algorithms. Refer to Fig.9 for interpreting the naming format of the subfigures. The result show that the algorithms trained either with CG rendered dataset or with Multi-PIE dataset have similar performance. Hence, the effectiveness of the CG rendered data is obvious.

B. Landmark Localization Across Large Rotation

Landmark localization for frontal face has been extensively researched and many well performed algorithms have been proposed recently. This experiment shows how to localize landmarks across large yaw rotation, say [-90°, +90°], by utilizing the benefit of the proposed CG data rendering pipe-line.

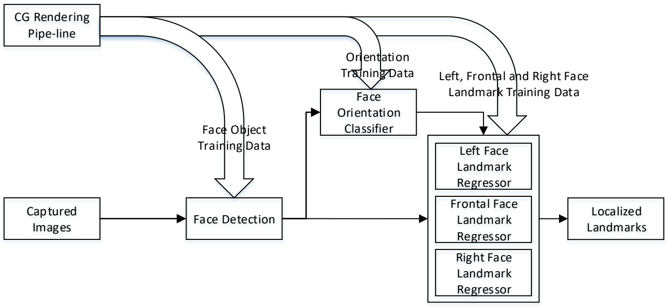

CG Rendering Pipe-line
Orientation Training Data
Left, Frontal and Right Face Landmark Training Data
Face Orientation Classifier
Face Object Training Data
Left Face Landmark Regressor
Frontal Face Landmark Regressor
Right Face Landmark Regressor
Captured Images
Face Detection
Localized Landmarks

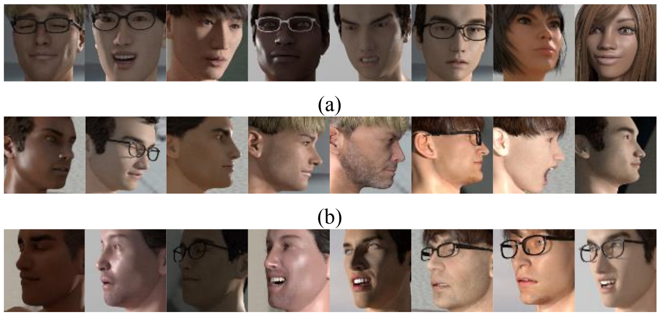
Fig. 11. System of facial landmark localization across large rotation. The system integrates three functional blocks: face detector, orientation classification and landmark regression. The CG rendering pipe-line produces large rotation training datasets for the three functional blocks respectively.

The system of facial landmark localization across large rotation consists of three functional blocks: face detection, orientation classification and landmark regressors as shown in Fig. 11. Firstly, the face detector scans the captured image and outputs the face region, then a face orientation classifier is utilized to determine its orientation, thereafter depends on the orientation class one of the three face landmark regressors are selected to localize the landmarks. The three functional blocks integrate simply. But lack of labelled data barriers the training of face detection, orientation classification and landmark regression across large rotation angle. Hence, the implementation of the simple integration is impossible unless enough data is available. Thanks to the benefits of the proposed CG rendering pipe-line the three functional blocks could be trained and integrated in low cost and fast.

The head orientation is divided into three classes according to the yaw angle. When the yaw is between [-90°, -40°] it belongs to left face class, when the yaw is between [-40°, +40°] it belongs to frontal face class, when the yaw is between [+40°, +90°] it belongs to right face class. Each orientation class has its own landmark regressor and there are a total of three landmark regressors. The regressor is implemented utilizing the ERT algorithm. Each regressor is trained with 30,000 CG rendered photorealistic images sampled randomly within the corresponding rotation region with variations of illuminations, expressions, appearances and backgrounds.

The paper utilizes the random forests [19, 20] as face orientation classification algorithm. The face orientation is classified into three categories: frontal, left and right. A total of 800 trees constitute the random forests, and each tree has a maximum depth of 12. The training dataset is created using the proposed CG rendering pipe-line. 30,000 frontal face images, 30,000 left face images and 30,000 right face images are generated as training dataset. The face orientation category label is determined by the yaw rotation. Randomness among [-7°, +7°] is added to the pitch and roll rotations respectively. The illumination, expression, appearance, hair style, glasses and gender attributes are randomly assigned to the 3D face model to make good generalization. The face region is cropped and resized to 50 * 50 pixels. The pixel comparison feature is utilized to feed the random forests. Some samples of CG rendered training data are shown on Fig. 12.

2,000 frontal face images, 2,000 left face images and 2,000 right face images are chosen randomly from the Multi-PIE database as test dataset. There are variations of illuminations, expressions and appearances in the test dataset. Table. 2 shows the positive classification accuracy. Therefore, the experiment proves the CG rendered facial data is also effective to train a face orientation classifier.


(a)

(b)

(c)

Fig. 12. Samples of CG rendered training data for face orientation classification. (a) frontal faces, (b) left faces and (c) right faces.

Table 2 Result of head orientation determination classification

Categories	Face Orientation Accuracy
front	100%
left	97%
right	96%

As for face detection we adopt the NPD Face Detector [21] algorithm due to its fast speed and can work in large face rotation. A total of 90,000 CG rendered face images across the left, frontal and right orientations are feed into the training algorithm which yields a fast and accurate face detector.

To verify the effectiveness of the landmark localization system across large rotation, which is benefited by the proposed CG rendering pipe-line, two experiments have been carried out: one on Multi-PIE dataset and the other on real persons. A total of 2,094 images across large rotation are selected from the Multi-PIE database, out of which 498 are left faces, 1,000 are frontal faces and 596 are right faces. Faces in all of these image can be detected by the NPD face detector. Fig.13 illustrates the landmark localization errors across large rotation angle. In Fig. 13 the blue dots are the individuals' error. The purple line, whose value is 0.022, is the overall mean error. The green line, whose value is 0.029, is the mean error for frontal face landmark localization. The red and golden lines, who's value are 0.014 and 0.013, are mean errors for left and right landmark localization respectively. Fig. 14 shows result samples of landmark localization across large rotation on both Multi-PIE dataset and real persons. It shows the proposed system trained with CG rendered images could do face detection, face orientation classification and landmark localization correctly across [-90°, +90°] yaw rotation.

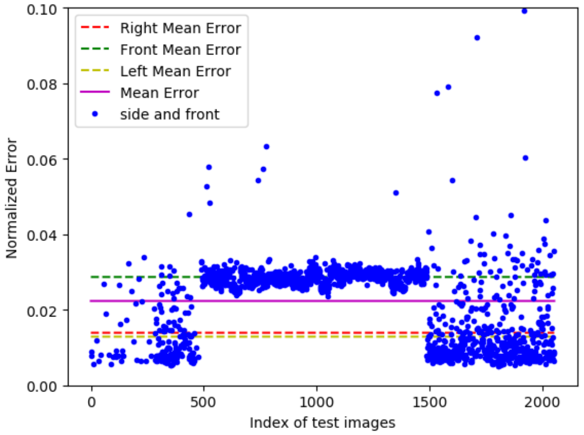
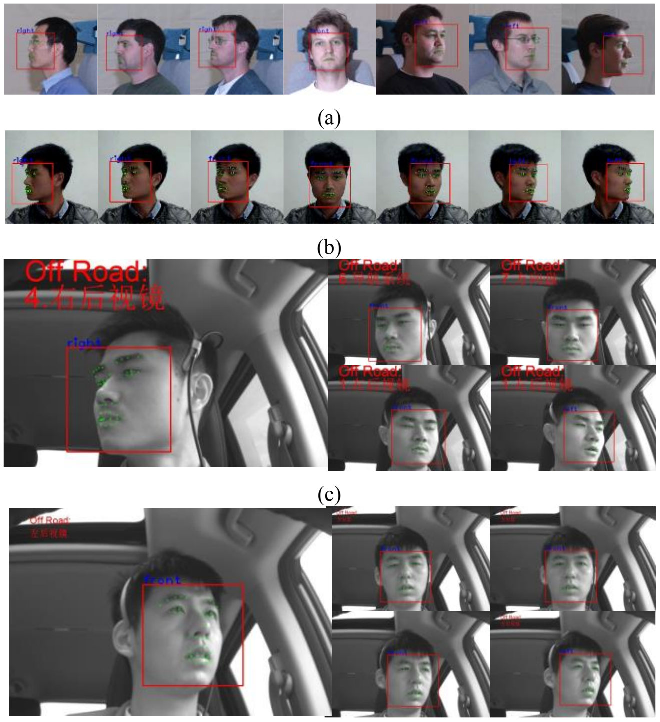


Fig. 13. landmark localization errors across large rotation angle on Multi-PIE dataset. The x-axis is the index of the test images, where the first 498 belong to left face, the middle 1,000 belong to the frontal face and the last 596 belong to right face. The y-axis is the normalized error. The blue dots are the individuals' error. The purple line, who's value is 0.022, is the over all mean error. The green line, whose value is 0.029, is the mean error for frontal face landmark localization. The red and golden lines, whose values are 0.014 and 0.013, are mean errors for left and right landmark localization respectively.



(a)

(b)

(c)

(d)

Fig. 14. Result samples of landmark localization across large rotation. (a) Result samples of Multi-PIE database, (b) Result samples of real persons, (c) and (d) are result samples of experiment in a car.

V. Conclusion

Facial landmark localization or facial alignment is a crucial initial step for Driver Inattention Monitoring. Though the state-of-the-art discriminative regression methods make significant progress for frontal facial landmark localization, its performance degrades dramatically when partial landmarks become invisible because of occlusions or large pose variation, which occurs often during driving. One major reason is lack of standard training dataset across large pose variations. There is a demand or creating data sets of face images together with ground truth labels. But creating data sets with ground truth labels is of high cost and needs too much manual labor. The paper proposes a flexible CG rendering pipe-line for creating facial image datasets associated with automatic ground truth labelling. The proposed pipe-line could generate photorealistic facial images with variations of poses, expressions, appearances and illumination as demanded. The benefits of CG (Computer Graphics) techniques such as 3D face modelling and morphing, photorealistic rendering and ground truth generation are utilized. Thanks to the efficiency of the CG pipe-line it could produce a huge amount of data fast and in low cost compared to traditional dataset creation methods which need high cost hardware and longtime manual ground truth labelling. Based on the established CG rendering pipe-line, the paper also proposes a capture setup of the CG environment for creating the dataset for facial landmark localization and its effectiveness is verified by cross validation with Multi-PIE dataset. Own to the benefits of the proposed CG rendering pipe-line the paper implemented a system for facial landmark localization across large rotation by integrating off-the-shelves algorithms. CG rendered data are feed into the training process of the functional blocks of the system. The experiments of the implemented system on Multi-PIE and real persons show that it could localize facial landmarks across [-90°, +90°] in yaw rotation accurately and in real time.

Since the proposed CG rendering pipe-line is of high efficiency for creating a huge amount of labelled data and there are unlimited setup configurations when generating data, the challenge becomes how to make an optimal setup configuration for producing high quality dataset for a particular application environment. In the future we will thoroughly investigate t varying factors, such as variations of the eye and mouth activities, expressions, appearances, rotations and illuminations, to give complete guidelines for designing new dataset and for assessing the capability of existing datasets.

REFERENCES

[1] Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1867-1874).

[2] Xiong, X., De la Torre, F. Supervised descent method and its applications to face alignment. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Portland, OR, June 2013, pp. 532-539

[3] Ren, S., Cao, X., Wei, Y., et al. Face Alignment at 3000 FPS via Regressing Local Binary Features. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Columbus, OH, June 2014, pp. 1685-1692。

[4] Dong Y, Wang Y, Yue J, et al. Robust facial landmark localization using multi partial features[C], Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015: 98-102.

[5] Joo H, Liu H, Tan L, et al. Panoptic studio: A massively multiview system for social motion capture[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3334-3342.

[6] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[J]. arXiv preprint arXiv:1611.08050, 2016.

[7] Koestinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011, November). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on (pp. 2144-2151).

[8] Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments (Vol. 1, No. 2, p. 3). Technical Report 07-49, University of Massachusetts, Amherst.

[9] Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012, October). Interactive facial feature localization. In European Conference on Computer Vision (pp. 679-692). Springer, Berlin, Heidelberg.

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

[11] Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., & Zhao, D. (2008). The CAS-PEAL large-scale Chinese face database and baseline evaluations. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38(1), 149-161.

[12] Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behavior research methods, 47(4), 1122-1135.

[13] Milborrow, S., Morkel, J., & Nicolls, F. (2010). The MUCT landmarked face database. Pattern Recognition Association of South Africa, 201(0).

[14] Kasinski, A., Florek, A., & Schmidt, A. (2008). The PUT face database. Image Processing and Communications, 13(3-4), 59-64.

[15] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009, September). A 3D face model for pose and illumination invariant face recognition. In Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on (pp. 296-301). IEEE.

[16] Blender. https://www.blender.org/

[17] Wang, Y., Yue, J., Dong, Y., & Hu, Z. (2016). Robust discriminative regression for facial landmark localization under occlusion. Neurocomputing, 214, 881-893.

[18] Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. International Journal of Computer Vision, 107(2), 177-190.

[19] Criminisi, A., & Shotton, J. (2013). Decision forests for computer vision and medical image analysis. 273-293.

[20] Dong, Y., Zhang, Y., Yue, J., & Hu, Z. (2016). Comparison of random forest, random ferns and support vector machine for eye state classification. Multimedia Tools & Applications, 75(19), 1-21.

[21] Liao, S., Jain, A. K., & Li, S. Z. (2016). A fast and accurate unconstrained face detector. IEEE transactions on pattern analysis and machine intelligence, 38(2), 211-223.

[22] Cycles. https://www.cycles-renderer.org/

[23] Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. IEEE Transactions on pattern analysis and machine intelligence, 25(9), 1063-1074.

[24] Blanz, V., & Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques (pp. 187-194). ACM Press/Addison-Wesley Publishing Co..