A Robust Real-time Facial Alignment System with Facial Landmarks Detection and Rectification for Multimedia Applications

Kuang Pen Chou[1], Mukesh Prasad[2], Jie Yang[2], Sheng-Yao Su[1], Xian Tao[3], Amit Saxena[4], Wen-Chieh Lin[1], Chin-Teng Lin[2]

[1]Department of Computer Science, National Chiao Tung University, Taiwan
[2]Centre for Artificial Intelligence, School of Computer Science, FEIT, University of Technology Sydney, Australia
[3]Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, China
[4]Department of Computer Science and IT, Guru Ghasidas University, India

Abstract— Face detection often plays the first step in various visual applications. Large variants of facial deformations due to head movements and facial expression make it difficult to identify appropriate face region. In this paper, a robust real-time face alignment system, including facial landmarks detection and face rectification, is proposed. A facial landmarks detection model based on regression tree is utilized in the proposed system. In face rectification framework, 2-D geometrical analysis based on pitch, yaw and roll movements is designed to solve the misalignment problem in face detection. The experiments on the two datasets verify the performance significantly improved by the proposed method in the facial recognition task and outperform than those obtained by other alignment methods. Furthermore, the proposed method can achieve robust recognition results even if the amount of training images is not large.

Index Terms—Face alignment, Facial feature localization, Head pose estimation, Face recognition.

I. INTRODUCTION

In an automatic visual system, face detection plays a vital role as the first step to locate the face region in the target image for potential applications. A face recognition system usually consists of face detection and face identification. A face in images is susceptible to change in lighting variation, pose variation, and facial expression variation. For the above reasons, it is hard to make a universal model to describe facial characteristics in all uncontrolled scenes. According to the difference of intrinsic properties, there are mainly two strategies to solve the issue of building face model: 1) rule-based methods and 2) learning-based methods. The rule-based methods construct face model in a straightforward way by transforming human knowledge into explicit rules [1], [2], [3]. Through adding hand-crafted features of the facial structure (such as facial contours and relationships between facial features) as detailed as possible, the performance of well-built model can be improved. However, these models are sensitive due to lack of the generalized ability to describe facial properties. The learning-based methods construct model by the combination of distribution functions or discriminant functions under a probability framework. One of the impressive methods is Viola and Jones' face detector [4], a novel face detection framework based on cascade architecture with Haar-like features to describe face model by a series of weak learners using decision tree learning. This technique brings a significant improvement in detection precision by following strategies:

1. The concept of *Integral Image* was introduced to allow the detector to calculate feature values quickly.

2. Adaptive boosting (AdaBoost) learning algorithm [5] was utilized as weak learners to select a small number of features, from tens of thousands of potential features to form a stronger classifier via a linear combination.
3. Use cascade architecture to obtain the final classifier, which combines several stronger classifiers to create a complex one. Meanwhile, in cascading, only samples that pass through the classifiers in the preceding stage are considered in the current layer to save computation time effectively.

In their proposed architecture, the current layer only considers detection-windows passed by the previous layer that means learners in the deeper layer focus on relative harder-to classifier windows. Under a reasonable division of labor, vast amount of non-face patches are judgement at early stage to achieve real-time detection. Numerous approaches have been developed based on Viola-Jones face detector, they mainly focus on extracting a different type of features to replace the original Haar-like features and designing variants of cascade architecture to enhanced performance against the effects caused by occlusion and pose variation.

The goal of a face recognizer is to identify whether a target of interest in an image exists in the registered dataset or not. The technologies of face recognition have been applied in various applications, such as public security, criminal identification, multimedia data management, etc. Furthermore, various technologies have been developed and brought out a significant advantage in the field of facial and pattern recognition system. In recent years, Sparse Representation Classifier (SRC) [6] has attracted the attention of many researchers and engineers from the computer vision community due to its impressive performance against varying expression, occlusion and noise issues. The key idea of SRC is to seek a sparse representation of the test samples as a linear combination of the whole set of training samples by solving an L1-minimization problem. Once L1-minimization computation is finished, SRC selects the subsets of training samples, which, most compactly express the test sample and rejects all other possible but less compactly representation. Furthermore, SRC does not have the training process for its classification; so, it is not necessary to train the SRC model again when a new face subject is added into a training set. However, the variations of human faces under different environment and conditions are still a challenging task in real-world applications. All of these factors associated with uncontrolled environments degrade the recognition rate of a facial recognition system. To alleviate the effects caused by uncontrolled environments, numerous well-know methods of feature extraction have been introduced, including Local Binary Pattern (LBP) [7], Scale Invariant Feature Transform (SIFT) [8], Histogram of Oriented Gradients (HoG) [9], and Haar-like features, etc. However, these approaches employ features to form local feature descriptors that are often sensitive to changes of illumination and lighting conditions and occlusion. Moreover, these human-made features are often designed for the specific object and lead the results in poor performance on other unseen targets.

Recently, convolution neural networks (CNNs), a variant of artificial neural network, attract the attention of researchers and brings an impressive performance on challenge datasets. The visual representation learned by CNNs have shown clear advantages over hand-crafted features for many recognition tasks [10], [11]. Chen et al. [12] presented an algorithm for unconstrained face verification based on deep convolutional features and evaluated it on the Labeled Face in the Wild (LFW) dataset. Jiang [13] applied the Faster RCNN to face detection. Wu et al. [14] suggested a Light CNN framework to learn a compact embedding on the large-scale face data with massive noisy labels. Tao [15] proposes a comprehensive deep learning framework to jointly learn face representation using multimodal information, which is composed of a set of elaborately designed CNNs and a three-layer stacked auto-encoder (SAE). However, the CNN-based methods often take considerable computation cost and requires a large number of data for model training.

In general, the above detectors can identify the presence of a face in an image well. However, the detection results are often rough due to variants of image appearance that might not be appropriate to use directly for subsequent applications as shown in Fig. 1. Therefore, rectification algorithms are required to fine-tune the detection to cover a suitable region according to the head orientation or align the facial landmarks to extract face patches. Huang et al. [16] have shown that the rectification process can enhance the performance on recognition task by removing undesired intra-class variability or covering the meaningful area as part of a face. There are two categories of rectification methods: 1) landmark-based methods and 2) unsupervised joint alignment methods. Landmark-based methods locate facial landmarks by holistic or part localizers to estimate the orientation of face [17], [18]. Once the face pose is determined, the face image is placed on a canonical pose according to the locations of landmarks. Instead of prior knowledge of facial landmarks, unsupervised joint alignment approaches align face with image similarity [16]. Unsupervised joint alignment methods do not need the prior knowledge of face structure. One successful representative is known as congealing, which works directly on the pixel values in each image by minimizing the entropy of each column of pixels through a set of collections.

To our best knowledge, most of the works focus on either face detection or localization of facial landmarks. This paper introduces a robust real-time face rectification to cover more informative patches in the image based on landmarks information with geometrical analysis to improve the performance of face recognition system. The main contributions of this paper are: 1) An ensemble of regression tree based on gradient boosting method is utilized for achieving high speed and accurate location in facial landmarks detection. 2) The proposed rectification algorithm with geometrical analysis is designed according to the head orientation, which takes light computation cost as pre-processing for other potential applications.

The rest of the paper is organized as follows: Sec. II presents a brief of related works of face alignment. Sec. III provides an overview of the proposed scheme and the details of the proposed algorithm. Sec. IV shows the experimental results on challenging datasets which, involve large variants, and finally, the conclusions are covered in Sec. V.

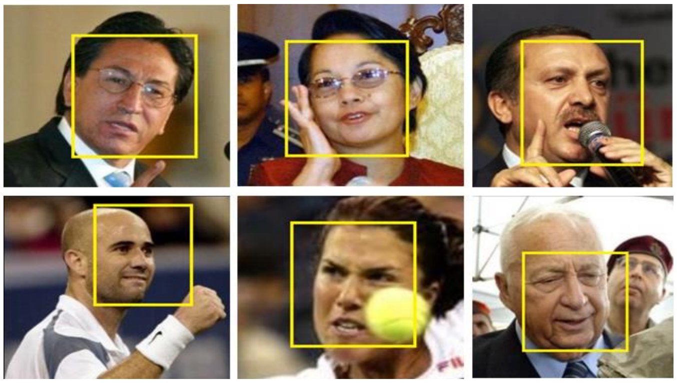


Figure 1: Unaligned face detection results.

II. RELATED WORK

Several approaches have been introduced to solve the problem of face alignment. As stated earlier, pose estimation can bridge the gap between human and visual systems. Two related model-based methods called Active Shape Models (ASM) [19] and Active Appearance Models (AAM) [20] are popular which utilizes the matching of statistical model of specific object shape and appearance to a new image. Both of them need to manually label facial landmarks as salient points on training images to obtain the face models. These landmarks are used to represent the face model and find the variation of facial shape by concatenating these landmarks as a vector and computing on the training set. Since positions of these landmarks are corrected, dimensionality techniques such as Principal Component Analysis (PCA) are further applied to obtain the primary modes of shape variation. By looking at the largest principal components, the direction of the search can be determined by corresponding variants in different directions. Once the location of facial features is known, the orientation of the image can be estimated by iteratively searching to fit the shape parameters to a pose setting.

The difference between ASM and AAM is that ASM uses only shape constraints (information about the image structure near the landmarks) and it does not consider the texture information. AAM model is more complex than ASM model, which learns the correlation between the shape model obtained from ASM and the shape free texture model to generate a combined appearance model. In general, coarsely alignment is much easier than a precisely alignment. As stated, even roughly alignment can improve the performance for further applications likes face recognition. In [16], Huang et al. introduced a congealing method, which jointly aligns a collection of images from a particular class by maximizing the likelihood of each image against to all the others. Those aligned images form as an alignment machine to align new input image efficiently. The proposed system automatically aligns images without prior knowledge of the object, which takes a poorly aligned image as input and return a well-aligned version of the image. In [17], a two-layered eyes localization method is proposed to align face images. At first, the input face image is transformed by a mathematical operator named Fast Radial Symmetry Transform to find potential locations of pupils with high radial symmetry. Then the eye classifier is applied to eliminate false candidates and identify the eyes in the image. After the localization process, the orientation of the face can be determined according to the angle of eyes and the image can be transformed into frontal pose by affine transform.

Deep learning methods have been achieving good performance on face alignment in recent years. Xiao [21] proposed a novel Recurrent Attentive Refinement (RAR) network for facial landmark detection under unconstrained conditions, suffering from challenges like facial occlusions and/or pose variations. Zhang et al. [22] presented a deep cascaded multitask framework that exploits the inherent correlation between detection and alignment to boost up their performance. Dapogny [23] presented an end-to-end deep convolutional cascade architecture for face alignment which uses fully-convolutional stages to keep full spatial resolution. In [24] a joint multi-task learning algorithm is proposed for both face alignment and segmentation which allows CNN model to simultaneously share visual knowledge between different tasks. However, a deep neural network usually requires a large number of training samples.

III. OVERVIEW OF PROPOSED FACE ALIGNMENT SYSTEM

The architecture for the proposed face alignment system is shown as Fig.2. Localize the landmark of input face obtained from the face detector [25] via a regression tree-based method.

According to the position of landmarks, the facial orientation can be easily estimated from two-dimensional geometric analysis. Then, the input face can be normalized to the frontal pose by compensating the face misalignment on each axis.

A. Localization for face feature point

The proposed system localizes facial landmarks of the input face obtained from the face detector via an ensemble of regression tree-based localizer as shown in Fig. 3(b). It should be noticed that we do not use all of 68 landmarks, only 7 feature points are used in our rectification framework, as demonstrated in Fig. 3(a), where E_r, E_l, C_r, C_l, NB, N and UL are outer corners of right and left eyes, side points of right and left face sides, tips of nose bridge, nose and upper lip, respectively. The orientation of a head is estimated based on these points using 2D geometrical analysis. The position vector of landmarks used in this work can be listed as follows:

$$P = [(x_1, y_1),\ (x_2, y_2), \dots,\ (x_7, y_7)]^T (1)$$

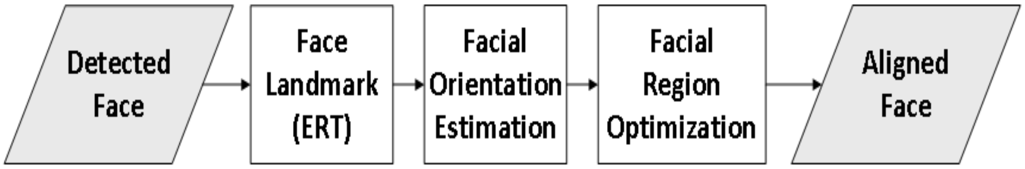
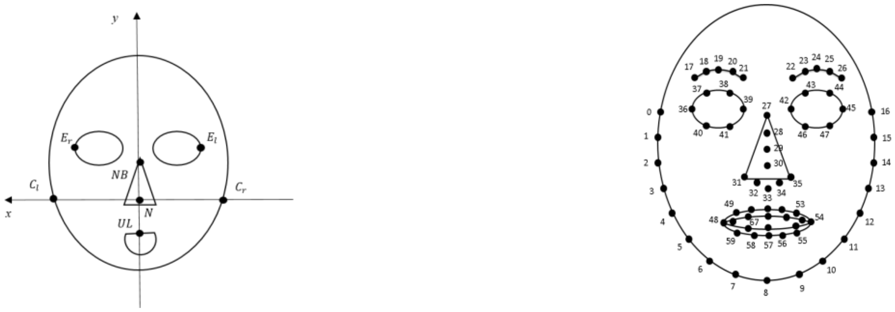


Figure 2: The flowchart of face alignment system



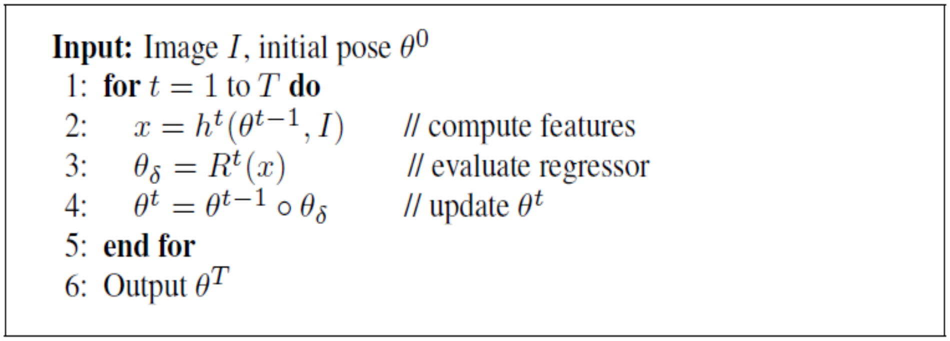
(a)

(b)

Figure 3: (a) an overview of the facial landmarks (b) 68 face feature points position

The detailed flow of regression tree is shown in Algorithm 1. In Algorithm 1, $R = (R^1, \dots, R^T)$ is a set of face pose regressions of a cascade structure. In order to train the cascade attitude regressor *R*, we input the image *I* and its initial pose angle θ^0 cascade to the pose regression to perform the program operation. After an iteration of $t = 1, \dots, T$, θ^T is the output. Each level of posture regression R^T is trained to minimize the amount of deviation between the correct posture and the posture being operated.

Algorithm 1 Regression tree of face pose



```
Input: Image $I$, initial pose $\theta^0$
for $t = 1$ to $T$ do
    $x = h^t(\theta^{t-1}, I)$        // compute features
    $\theta_\delta = R^t(x)$          // evaluate regressor
    $\theta^t = \theta^{t-1} \circ \theta_\delta$    // update $\theta^t$
end for
Output $\theta^T$
```

B. *Face direction evaluation*

Based on the facial landmarks in section III.A, we can predict the face pose and provide face direction compensation, which include weight, height, rotate angles and rotate centers. Head pose can be used to infer the relationship between face direction and coordinated system. Head can be limited in a 3-DOF (Depth of Field) space, and then describe the rotate feature in three directions. Fig.4 shows the three directions: Pitch (rotate with x-axis), Yaw (rotate with y-axis) and Roll (z-axis).

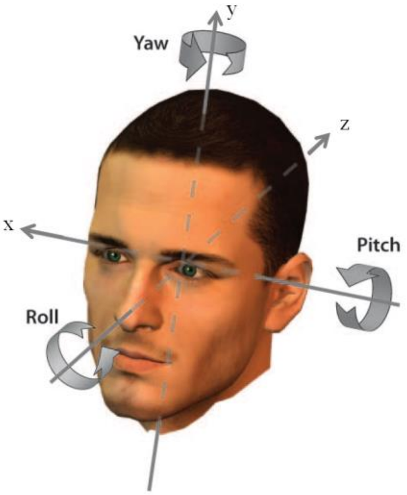


Figure 4: The three degrees of freedom of a human head can be described by the geocentric rotation as pitch, yaw and Roll [26]

C. *Compensation of Pitch*

Fig.5 indicates an example of a 3-D image projection on x-y 2-D plane, where *N* is nose apex, *NB* is the nose bridge location, *UL* is the central point of upper lip. Fig.6 shows the head geometric perspective that rotates with x-axis, where *n* is the 2-D image projection vector for *N*, *nb* is the 2-D image projection vector for *NB*, similar with *ul* for *UL, f* is the focus distance of camera,γ is the pitch angle of rotate with x-axis. Blue line means rotate with x-axis in anticlockwise and its projection vector, red line is rotate with x-axis in clockwise and its projection vector, black line is head in the condition of $\gamma = 0$and its projection vector.

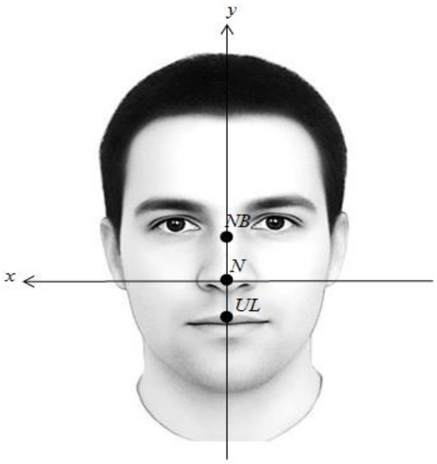


Figure 5: 3-D image projection on x-y 2-D plane with pitch

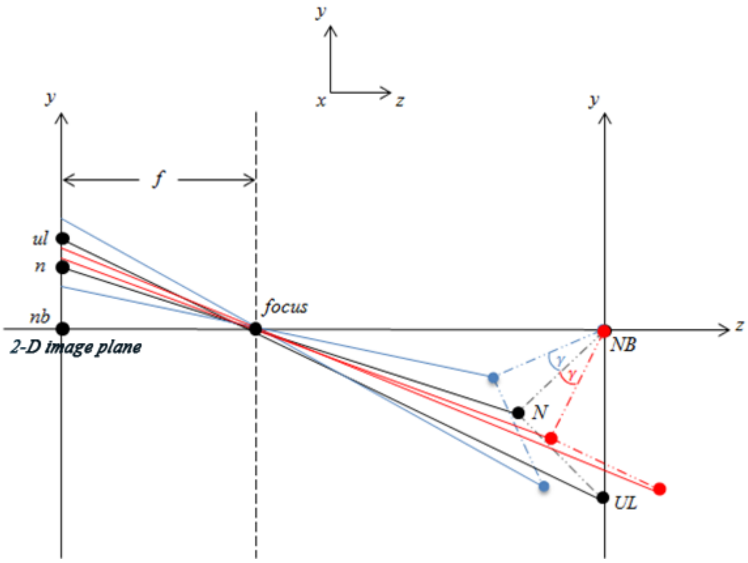


Figure 6: Head geometric perspective with x-axis rotation

The nose length and the distance between nose tip and upper lip are defined as $y_u = |nb - n|$ and $y_d = |n - ul|$. To investigate the influence caused by changes of pitch angle, y_u and y_d are used to formulate proportional relationship $\hat{p}$ as follows:

$$\hat{p} = \frac{y_u}{y_u + y_d} \tag{2}$$

According to the experiment results, the proposed system only compensates the face images in the case $\gamma > 0$ ∘, that the ratio of $\hat{p}$ have significantly various in the image plane when the head looks up. The compensation formula is shown as follows:

$$\Delta \mathrm{y} = \begin{cases} y_u\,, & \text{if } \gamma > 0 \\ 0\,, & \text{if } \gamma \leq 0 \end{cases} \tag{3}$$

where Δy is the compensation displacement to the central point of face along the y-axis. It should be noted that the total facial feature length $y_u + y_d$ is only slightly difference between the case of $\gamma = 0^{\circ}$ and $\gamma = -15^{\circ}$ degrees. For this reason, our research only aligns the face imagesin the case of $\gamma > 0^{\circ}$ that the facial feature length is shortened by the effect of pitch rotation. The compensation formulae are listed as follows:

$$Center_{pitch}(x, y') = Center_{origin}(x, y + \Delta y) \tag{4}$$

$$Height_{pitch} = Height_{origin} - \Delta y \tag{5}$$

D. Compensation of Yaw

Similar to the pitch rotation in 3.C, Fig.7 shows a 3-D image projection on x-y 2-D plane with yaw angle, Fig.8 shows the head geometric perspective that rotate with y-axis, where C_r and $C_l$are left and right-side points, E_r and $E_l$are outer corner of eyes, *N* is nose apex.c_r,c_l,e_r,e_l and *n*are the 2-D image projection vectors for them, respectively.

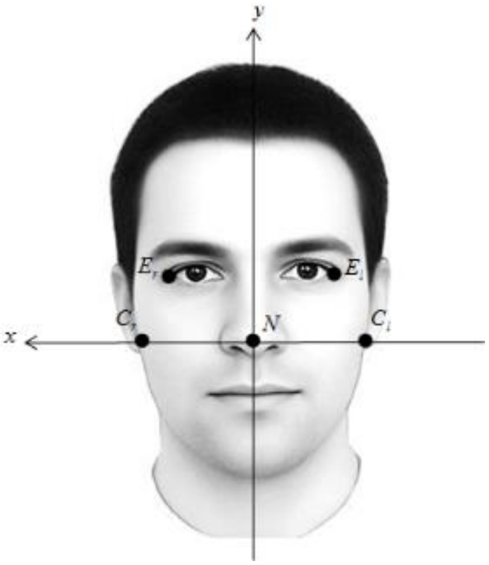


Figure 7: 3-D image projection on x-y 2-D plane with yaw

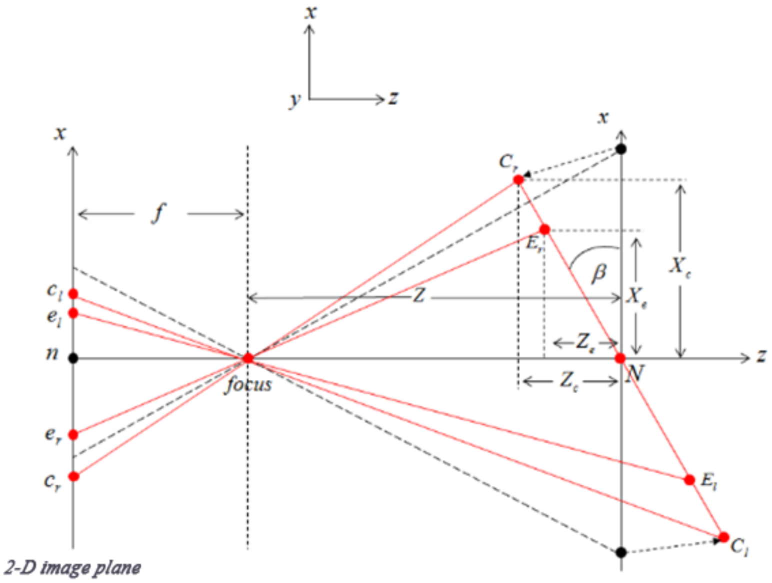


Figure 8: Head geometric perspective with y-axis rotation

The projection terms $c_r$and $c_l$can be described as follows:

$$c_r = \frac{fX_c}{Z-Z_c} = \frac{f\overline{C_rE_r}cos\beta}{Z-Z_c} \tag{6}$$

$$c_l = \frac{fX_c}{Z+Z_c} = \frac{f\overline{C_lE_l}cos\beta}{Z+Z_c} \tag{7}$$

The yaw angle β for c_r and c_l are given by:

$$\beta = arc\cos(A_rc_r) \tag{8}$$

$$\beta = arc\cos(A_lc_l) \tag{9}$$

where$A_r = \frac{Z-Z_c}{f\overline{C_rE_r}}, A_l = \frac{Z+Z_c}{f\overline{C_lE_l}}$

Clearly, c_r and $c_l$have inverse relationship with the yaw angle β. Take advantage of this proportional relation as the rotation score with y-axis,the position for face image can be designed as follows:

$$\Delta x = \begin{cases} \left(1-\frac{c_r}{c_l}\right)(c_l - e_l), & \text{if } \beta < 0 \\ \left(1-\frac{c_l}{c_r}\right)(c_r - e_r), & \text{if } \beta > 0 \end{cases} \tag{10}$$

where Δx is the compensate move distance with x-axis.

The face image after compensation can be shown as below:

$$Center_{yaw}(x', y) = \mathrm{Center}_{origin}(x + \Delta x, y) \tag{11}$$

where Center_{origin} is the face's original center position. Center_{yaw} is the center position after compensated with *Yaw* rotation.

E. Compensation of Roll

Fig.9 presents a 3-D image projection on x-y 2-D plane, rotate with z-axis in anticlockwise. Fig. 10(a) shows the geometric perspective of roll rotation, where e_l and e_r are projection points of E_l and E_r, respectively. In this case, the coordinates of facial landmarks change on x-y plane with a fixed z value. Moreover, geometric relationships between feature points keep while a head rotation against z-axis as shown in Fig. 10 (b). Therefore, outer corners of eyes are used as facial feature points to compensate an unaligned image for roll motion. The roll angle α can be written as follows:

$$\alpha = \mathrm{arc}\tan(\frac{Y_e}{X_e}) \tag{12}$$

where $X_e = e_r^x - e_l^x,\ \ Y_e = e_r^y - e_l^y$.

Once the roll angle is determined, an unaligned face image can be directly posted onto frontal face direction by performing the affine transform as follows:

$$M = \begin{bmatrix} cos\alpha & -sin\alpha \\ sin\alpha & cos\alpha \end{bmatrix} \tag{13}$$

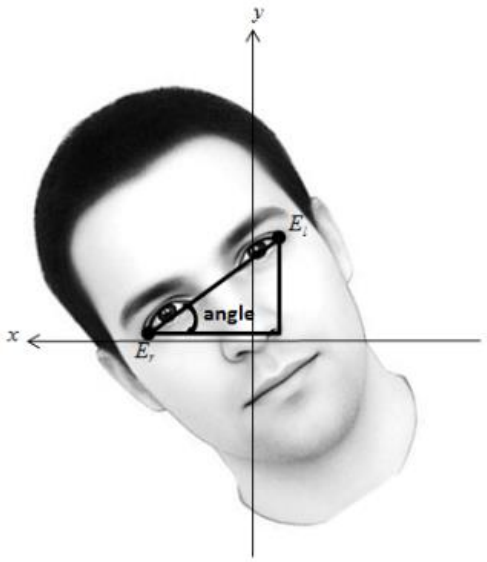
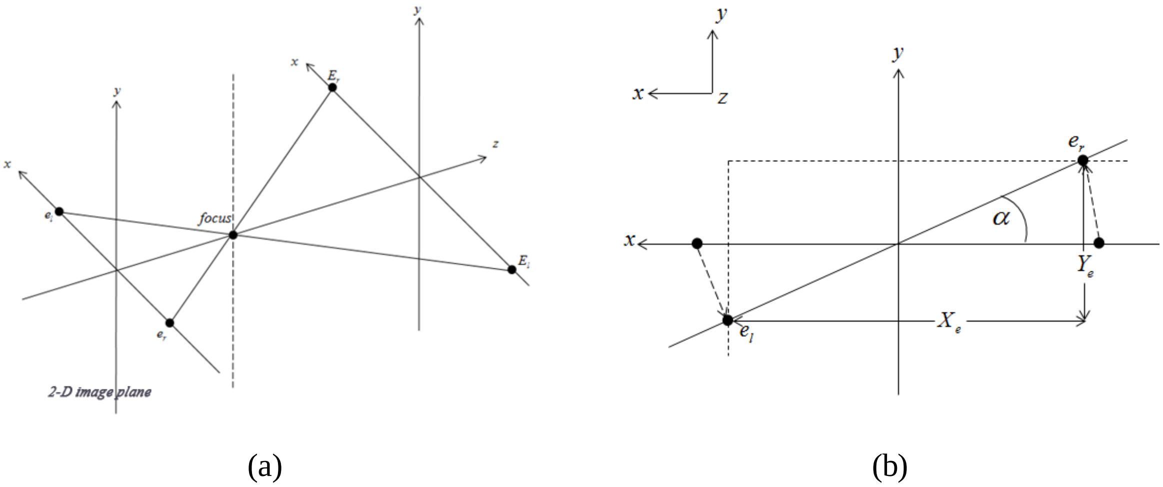


Figure 9: 3-D image projection on x-y 2-D plane with roll



(a) (b)

Figure 10: Head geometric perspective with z-axis rotation

IV. EXPERIMENTS AND DISCUSSION

We conduct a series of discussion about the rationality of the proposed method and compare it with other methods on the uncontrolled dataset. In order to investigate the rationality, the pixel-based statistic is used to demonstrate the robustness of our method. We use FEI face dataset [27] to show the benefit from rectification. For the uncontrolled setting, we use LFWdataset [28] to test the performance of the proposed method with other aligned methods. For all experiments, we construct a face recognition framework, and all methods follow the samesetting for a fair comparison. In the recognition framework, the Viola-Jones face detector and the Sparse Representation Classifier (SRC) recognizer [6] are used to identify face location and classification respectively.

A. Databases overview

Four open databases are used to evaluate the experiment. First one is FEI face database, which consists of 200 individuals, each has 14 images with different angles. This database is used to evaluate face features. The second database is LFW face database, which has 13,233 images in total. LFW include different features like pose and emotion change, sex and cover, etc. This database is used to evaluate the efficiency of face alignment. The third one is CAS-PEAL-R1 database [29], there are 1,040 individuals with different poses, emotion and light. This database can evaluate the influence of subject numbers. The last one is AR database [30], there are 126 individuals (70 males and 56 females) with a little light and emotion change, but take picture in two sessions. It is used to compare different face recognition systems.

B. The facial landmarks result of the Proposed Algorithm

We apply the regression tree to locate the feature points of the face image in the web camera. The result of the feature points of the positioning is shown in Fig. 11. As can be seen from Fig.11, the proposed algorithm has good feature point positioning accuracy.

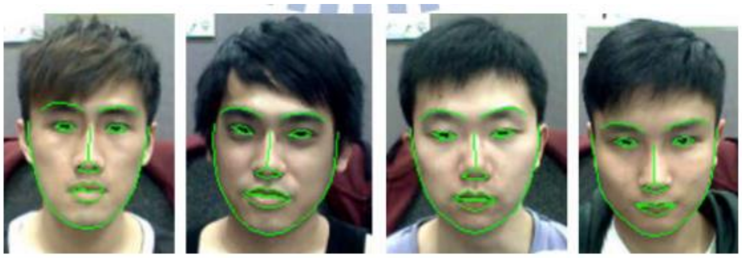


Figure 11: Face labeled result by Webcam

C. The Rationality of the Proposed Algorithm

In the above section, we introduced a geometrical analysis method for three DOF of head motion. For pitch direction, the ratio of length between features is used as criteria for compensation rather than the feature length directly. We use CAS-PEAL-R1 dataset to test the robustness of the proposed method. CAS-PEAL-R1 contains 30900 images of 1040 subjects. We consider the pose subset, which contains seven yaw changes with tree pitch angles for each subject. For each pitch angle, five images with different yaw changes are used for analysis. The statistical results of feature length change are shown in Table I. From the results in table I, we can clearly see that the length y_u is greater than y_d and the sum of y_u and y_d are almost equal in both cases γ = 0◦and γ = −15◦. In the case of γ = +15◦, y_d is smaller than y_u and the sum of yu and yd decreases, that means the projection length of face in the image plane decrease when the pitch angle increases.According to this property, we need to reduce the height of the face region in the image plane.

Table I: Face feature length in different pitch angles

Parameter γ	$\boldsymbol{y_u}$	$\boldsymbol{y_d}$	$\boldsymbol{y_u + y_d}$	$\hat{\boldsymbol{p}}$
+15^o	36.33	40.33	76.66	0.4727
0^o	45.59	38.67	84.26	0.5407
-15^o	53.37	32.25	85.61	0.6234

Table II: The standard deviation of the face feature measurements in different pitch angles

Parameter γ	$\sigma\,(\boldsymbol{y_u})$	$\sigma\,(\boldsymbol{y_d})$	$\sigma\,(\boldsymbol{y_u + y_d})$	$\sigma\,(\hat{\boldsymbol{p}})$
+15^o	15.1%	10.96%	10.14%	8.84%
0^o	12.21%	12.11%	9.06%	7.47%
-15^o	12.18%	15.92%	9.63%	7.6%

To further investigate the robustness of the proposed compensation scheme, we considered the distributions between subjects as shown in Table II. The results show that the proposed parameter $\hat{p}$ is more stable than other features in all pitch cases. Fig. 12 shows the compensation results by the proposed method. For yaw direction, we assume that face is symmetrical along y-axis. Under this assumption, we move predicted face bounding box to remove the undesired background and acquire potential useful face information. Fig. 13 shows the compensation results for yaw motion.

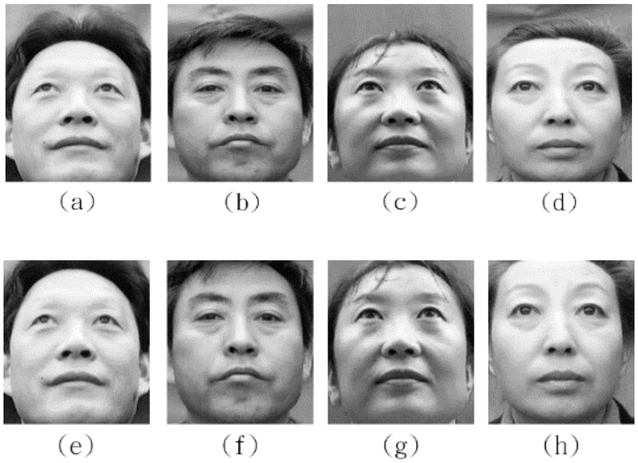


Figure 12: Pitch correction to CAS-PEAL-R1 database. (a)-(d) are uncorrected images, (e)-(h) are corrected images.

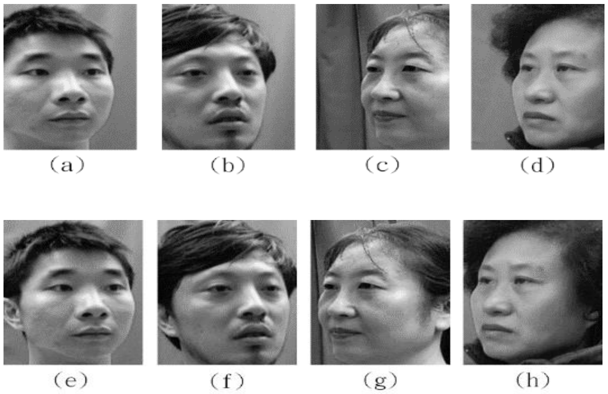


Figure 13: Yaw correction to CAS-PEAL-R1 database. (a)-(d) are uncorrected images, (e)-(h) are corrected images.

For roll motion, due to the phenomena that the angle between eyes and nose tip keep while the head rotates. It is straightforward to use this property to compensate the misaligned caused by rotation through the affine transform. Fig. 14 shows the compensation results for roll rotation on LFW dataset.

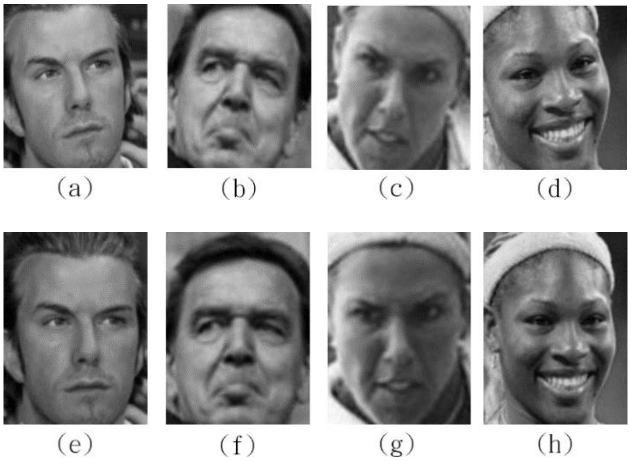
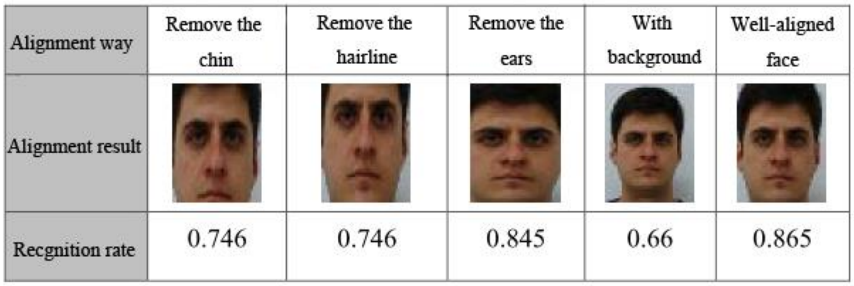


Figure 14: Roll correction to LFW database. (a)-(d) are uncorrected images, (e)-(h) are corrected images

D. *Influence of face alignment to revolution ratio*

This section demonstrate that a good alignment method can significantly enhance the performance of face recognition system. 14 images for each individual in FEI are selected firstly, but only 9 images without extreme illumination and yaw rotation are considered in the evaluation process. Fig. 15 shows various part in a face image and its revolution ratio. The accuracy of SRC without alignment is severely degraded if the input face image includes much background region or some facial feature points are lost. It proves that SRC can achieve higher performance via precise alignment technique. Fig. 16 shows the efficiency of face alignment for face recognition. In detail, the recognition rate by face alignment method (red curve) is 86.8%, which is higher than non-face alignment method by 11.2%.



Alignment way	Remove the chin	Remove the hairline	Remove the ears	With background	Well-aligned face
Alignment result					
Recgnition rate	0.746	0.746	0.845	0.66	0.865

Figure 15: Relationship between revolution ratio and recognition rate

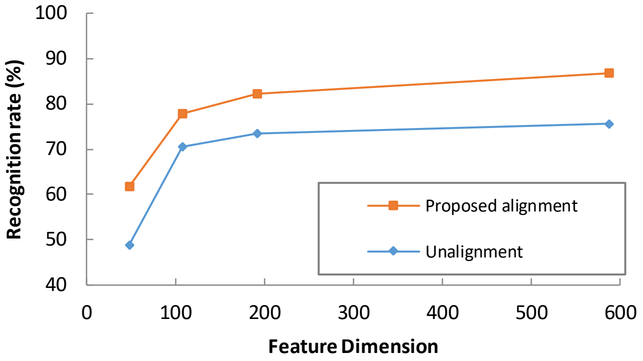


Figure 16: Face alignment comparison by FEI database

E. Alignment algorithms comparison

The datasets for face alignment consist of 34 individuals (each of them has 30 images) from LFW database. We use Viola and Jones's face detection system [4], after getting the aligned face position, the image can be trained by SRC and do the latter face recognition process.

The eye alignment method introduced by Li and Wang et. al [31] is used to compare the performance. Firstly, utilize Viola and Jones's detect method on LFW database, then align face by eye alignment method, finally train the images by SRC, it is the same as the proposed method.

In the above experiment, we use Bilinear Interpolation method in OpenCV dataset to decrease image's dimension from 250*250 to 48, 108, 192 and 588. Fig. 17 shows the influence of feature dimension to recognition accuracy, Fig. 18 shows the influence of subject amount to recognition rate. It is obvious that the proposed alignment method can eliminate useless background information and align face images more accurate than other methods.

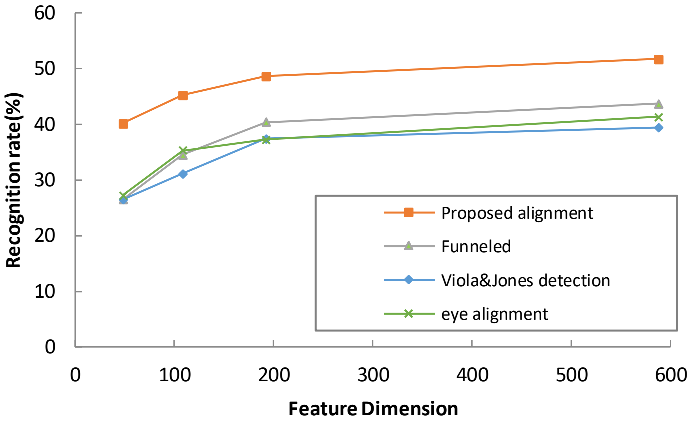


Figure 17: Face recognition comparison by LFW database

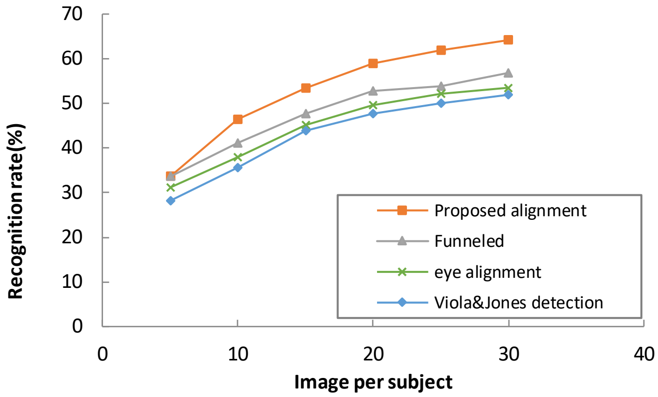


Figure 18: Relationship between subject amount and recognition rate

F. *Recognition efficiency comparison*

Nearest Neighbor classifier [32] and Nearest Subspace [33] are used to compare the efficiency of recognition task. Fig. 19 indicates result for different methods. The result of combined face alignment with SRC (red line) can reach at 88.27%, nearest neighbor classifier is 73.71%, nearest subspace is 81.71%, which means our proposed system is more efficient that other methods.

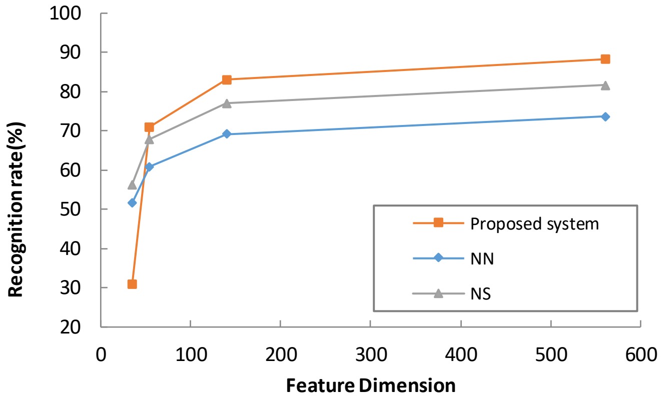


Figure 19: Face recognition comparison by AR database

G. *Evaluation to the number of subjects*

Fig. 20 indicates the recognition rate for different number of subjects. This validation is implemented using the subset of CAS-PEAL-R1 dataset. For each individual, the seven samples with normal pitch angle are employed to the training set. Furthermore, some variations such as scaling, rotation, and horizontal translation can be randomly applied to the frontal pose sample (yaw angle is 0^o). It shows that despite having the lack of training samples, the proposed system provides satisfactory robustness and accuracy.

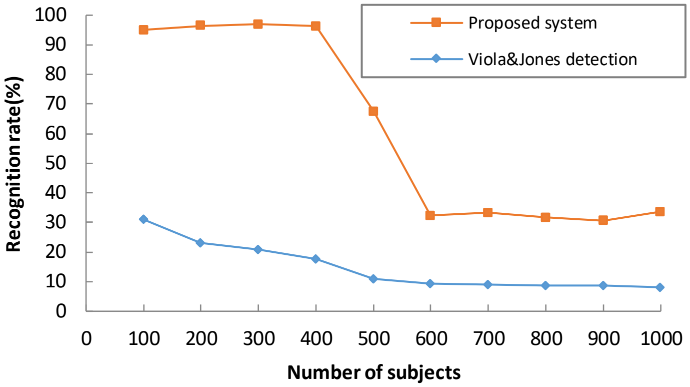


Figure 20: Recognition rate for various subject numbers by CAS-PEAL-R1 database

V. CONCLUSION

This paper introduces an efficient alignment system for face detection. At first, facial feature localizer based on regression tree is used to identify the facial landmarks for a given face region. While key features are acquired, origin face region is adjusted to compensate misalignment effects caused by pose variation. According to different conditions, the face region is modified to ignore unsuitable region and include potential valuable region. The experiments demonstrate that the proposed method can significantly enhance the performance for the recognition task and outperform other alignment methods. Although the performance of the proposed method is highly associated with the ability of facial localizer, the accuracy of localizer can be improved by applying the proposed method iteratively to acquire a proper face area. Moreover, the proposed algorithm can be extended for 3D face model, more complete geometrical information of facial characteristics can be used to further improving the performance of the pose correction process.

ACKNOWLEDGEMENT

This work was supported in part by the Australian Research Council (ARC) under Grant DP180100670 and Grant DP180100656, in part by the U.S. Army Research Laboratory under Agreement W911NF-10-2-0022, and in part by the Taiwan Ministry of Science and Technology under Grant MOST 106-2218-E-009-027-MY3.

REFERENCES

[1] R. Brunelli and T. Poggio, “Face recognition: features versus templates”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042–1052, Oct 1993.

[2] C. Kotropoulos and I. Pitas, “Rule-based face detection in frontal views,” in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, April 1997, pp. 2537–2540 vol.4.

[3] A. Lanitis, C. Taylor, and T. Cootes, “Automatic face identification system using flexible appearance models,” Image and Vision Computing, vol. 13, no. 5, pp. 393 – 401, 1995, 5th British Machine Vision Conference.

[4] P. Viola and M. J. Jones, “Robust real-time face detection”, International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, May 2004.

[5] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," in *Proceedings of the 2nd European Conference on Computational Learning Theory*, 1995.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, Feb 2009.

[7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, 2006.

[8] D. G. Lowe, "Distinctive Image Features from Scale-invariant Key points," *IJCV*, vol. 60, no. 2, pp. 91-110, 2004

[9] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, ser. NIPS’12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[11] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[12] Chen, Jun-Cheng, Vishal M. Patel, and Rama Chellappa. "Unconstrained face verification using deep cnn features." In2016 IEEE winter conference on applications of computer vision (WACV), pp. 1-9. IEEE, 2016.

[13] Jiang, Huaizu, and Erik Learned-Miller. "Face detection with the faster R-CNN." In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 650-657. IEEE, 2017.

[14] Wu, Xiang, Ran He, Zhenan Sun, and Tieniu Tan. "A light cnn for deep face representation with noisy labels." IEEE Transactions on Information Forensics and Security 13, no. 11 (2018): 2884-2896.

[15] Ding, Changxing, and Dacheng Tao. "Robust face recognition via multimodal deep face representation." IEEE Transactions on Multimedia 17, no. 11 (2015): 2049-2058.

[16] G. B. Huang, V. Jain, and E. G. Learned-Miller, “Unsupervised joint alignment of complex images,” 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8, 2007.

[17] H. Li, P. Wang, and C. Shen, “Robust face recognition via accurate face alignment and sparse representation,” in 2010 International Conference on Digital Image Computing: Techniques and Applications, Dec 2010, pp. 262–269.

[18] H. Lu and F. Yang, Active Shape Model and Its Application to Face Alignment. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 1–31.

[19] T. Cootes, C. Taylor, D. Cooper, and J. Graham, “Active shape models their training and application,” Computer Vision and Image Understanding, vol. 61, p. 389, 01 1995.

[20] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681–685, June 2001.

[21] Xiao, Shengtao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. "Robust facial landmark detection via recurrent attentive-refinement networks." InEuropean conference on computer vision, pp. 57-72. Springer, Cham, 2016

[22] Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23, no. 10 (2016): 1499-1503.

[23] Dapogny, A., Bailly, K., & Cord, M. (2019). DeCaFA: Deep Convolutional Cascade for Face Alignment In The Wild. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6893-6901).

[24] Zhao Y, Tang F, Dong W, et al. Joint face alignment and segmentation via deep multi-task learning. Multimedia Tools and Applications, 2019, 78(10): 13131-13148.

[25] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 2014, pp. 1867–1874.

[26] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 607–626, April 2009

[27] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," Image and Vision Computing, vol. 28, no. 6, pp. 902 – 913, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609002613

[28] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[29] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 38, no. 1, pp. 149–161, Jan 2008.

[30] A M. Martinez and R. Benavente, "The AR face database," CVC Technical Report 24, 1998.

[31] H. Li, P. Wang, and C. Shen, "Robust Face Recognition via Accurate Face Alignment and Sparse Representation," 2010 International Conference on Digital Image Computing: Techniques and Applications, pp. 262–269, 2010.

[32] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, 1967.

[33] J. Ho, M.-H. Y. M.-H. Yang, J. L. J. Lim, K.-C. L. K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," Proceedings. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 11-18, 2003.