# Intensity-Depth Face Alignment Using Cascade Shape Regression

Yang Cao[1] and Bao-Liang Lu[1,2](✉)

[1] Department of Computer Science and Engineering,
Center for Brain-like Computing and Machine Intelligence,
Shanghai Jiao Tong University,
800 Dongchuan Road, Shanghai 200240, China
bllu@sjtu.edu.cn

[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University,
800 Dongchuan Road, Shanghai 200240, China

**Abstract.** With quick development of Kinect, depth image has become an important channel for assisting the color/infrared image in diverse computer vision tasks. Kinect can provide depth image as well as color and infrared images, which are suitable for multi-model vision tasks. This paper presents a framework for intensity-depth face alignment based on cascade shape regression. Information from intensity and depth images is combined during feature selection in cascade shape regression. Experimental results show that this combination improves face alignment accuracy notably.

**Keywords:** Face alignment · Depth image · Cascade shape regression

## 1 Introduction

Face alignment is to detect key points such as eye corner, mouth corner and nose tip on human face, and is an important step for many face related vision tasks like face tracking and face recognition. A variety of models are proposed to tackle face alignment problem. Notable ones are Active Shape Model (ASM) [8], Active Appearance Model (AAM) [7], Constrained Local Model (CLM) [9], Explicit Shape Regression [4] and Deep Convolutional Network [14]. Besides, numerous improvements for these models have been proposed.

Among those different models, the family of cascade regression models is the leading one. Generally, cascade regression models solve face alignment problem with many stages of regressions (in a cascade manner). Based on the basic structure, a lot of models are proposed. Cao et al. [4] use two-level boosted regression, shape indexed features and fast correlation-based feature selection. They achieve remarkable results. Asthana et al. [1] make the learning parallel and incremental, so that the model can automatically adapt to data. Burgos et al. [3] model occlusion explicitly to locate occluded regions and improve performance for occluded faces. Chen et al. [5] deal with face detection and face alignment jointly, which improves performance on both problems.

S. Arik et al. (Eds.): ICONIP 2015, Part IV, LNCS 9492, pp. 224–231, 2015.
DOI: 10.1007/978-3-319-26561-2_27

Most face alignment studies focus on intensity image. Unlike intensity image, each pixel in a depth image represents the distance between the point in the scene and the camera. Intuitively, information from intensity image and depth image is complementary. For example, eyelid and eyebrow are very distinct on intensity image because of their darkness; tip of nose is very distinct on depth image because of its shape. However, only a few researchers try to combine them in face alignment tasks [2,6].

Both [2] and [6] are based on CLM. Nevertheless, this approach has some limitations. The first one is that it is not optimal because intensity information and depth information do not always contribute equally to face alignment in the whole face region. Moreover, this approach is based on CLM, but recent studies show that cascade shape regression achieves higher performance. In this paper, we address those two problems in a novel way.

## 2 Framework

We adapt cascade shape regression model to depth image, and use a novel approach to combine intensity image and depth image. Figure 1 is an overview of our framework.

### 2.1 Problem Description

Suppose that $I$ is an image (intensity-depth image in our case), which contains human face, $R$ is a rectangle, which gives face region in $I$, and $Y = [x_1, y_1, \dots, x_N, y_N]^T$ is the ground truth of facial landmarks. The task of face alignment is computing an estimation $\hat{Y}$ from only $I$ and $R$, which minimizes

$$\|Y - \hat{Y}\|_2. \tag{1}$$

### 2.2 Framework Structure

Since computing $\hat{Y}$ in one-shot is difficult, almost all face alignment models work in a cascade manner. That is, let $\hat{Y}_1$ (which is usually the average landmarks

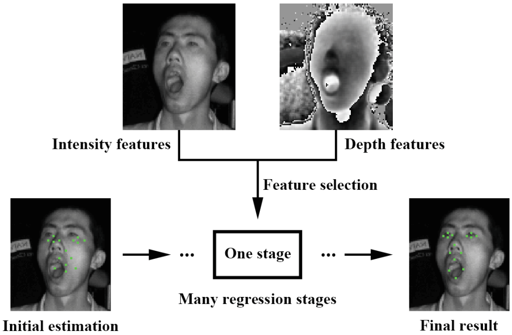

**Fig. 1.** The workflow of our framework.

positions translated into $R$) be the initial estimation, and compute a better estimation $\hat{Y}_2$ from $I$ and $\hat{Y}_1$. Repeat this for several times and we can get the final result $\hat{Y}$.

Our framework also works in this manner. We use the two-level structure proposed in [4]. In our framework, there are $T$ stages. Therefore we begin with $\hat{Y}_1$, and get $\hat{Y}_2, \hat{Y}_3, \ldots, \hat{Y}_t, \ldots$, iteratively, until $\hat{Y}_T$. $\hat{Y}_T$ will be our final result $\hat{Y}$.

### 2.3 Feature

Doing regression on $I$ and $\hat{Y}_t$ directly is hard. We must extract features from $I$ and $\hat{Y}_t$.

**Shape Indexed Features.** Shape indexed features mean that features are extracted relative to landmarks. For face alignment, shape indexed features are more robust against pose variation. In [4], a feature is the difference between intensity values of two pixels. The pixel position is generated by taking an offset to a certain facial landmark. Therefore, pixels positions are relative to facial landmarks, which makes them invariant in different poses. In [3], linear interpolation between two landmarks is used as the position of a pixel, which makes it more robust.

In our framework, we also use shape indexed features. For intensity image, we directly use the difference between intensity values of two pixels, whose positions are randomly generated, as a feature. We generate many such features and then do feature selection. For depth image, we use normalized depth features.

**Normalized Depth Features.** Shape indexed features can effectively deal with pose variation problem for intensity image. But for depth image, shape indexed features are not enough, since the depth value of each pixel will change dramatically during pose variation. To make information of depth image more robust against pose variation, we propose normalized depth features.

As in the intensity image case, we use the difference between depth values of two pixels as a feature. To make it invariant under pose variation, we compensate each value of pixel according to the pose.

If we use a plane to approximate the human face, the plane can be expressed by

$$X\boldsymbol{\beta} = z, \tag{2}$$

where $X = [1, x, y]^T$ is the location of the pixel, $\boldsymbol{\beta}$ is the parameter of the plane and $z$ is the depth of the pixel.

Suppose $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}]^T$, $\boldsymbol{x_k} = [1, x_k, y_k]^T$ is a landmark on the depth image, and $\boldsymbol{z}$ is the corresponding depth values. Then their relationship can be expressed by

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{z}, \tag{3}$$

which can be solved by linear regression. Using normal equation, the result is

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{z}. \tag{4}$$

Having got the parameter $\boldsymbol{\beta}$ of the face plane, we can compensate the depth value of a pixel by

$$z' = z - \alpha X\boldsymbol{\beta}, \tag{5}$$

where $z$ is the original depth value, $z'$ is the compensated depth value, and $X$ is the location of the pixel. We also use attenuation parameter $\alpha$ for the compensation, because the estimated face pose may not be very accurate.

The estimation of $\boldsymbol{\beta}$ is an iterative process. As the estimation of landmarks becomes more accurate during face alignment, the estimation of $\boldsymbol{\beta}$ will also become more accurate, which in turn contributes to the estimation of landmarks.

**Feature Selection and Fusion.** Using the feature extraction method described above, we can randomly (randomly select a landmark and randomly choose an offset) generate a huge feature pool for each image of training data, which contains not only intensity features but also depth features. We use correlation-based (Pearson Correlation) feature selection method proposed in [4],

$$j_{\text{opt}} = \arg\min_j \text{corr}(\boldsymbol{Y}\boldsymbol{v}, \boldsymbol{X}_j), \tag{6}$$

where $\boldsymbol{X}$ is a matrix to represent all randomly generated features of all the training data in current external-stage, $\boldsymbol{X}_j$ is a column vector which represents the $j$th feature of all the training data, $\boldsymbol{Y}$ is a matrix in which each row represents the disparity between current estimation of landmarks positions and true landmarks positions, and $\boldsymbol{v}$ is a vector drawn from unit Gaussian to project $\boldsymbol{Y}$ into a vector. Note that $\boldsymbol{v}$ is necessary here, so that we can compute the correlation.

In our model, each stage contains $K$ internal-stages [4]. In the training of each internal-stage (a fern [10]), we use the feature selection method described above to select $F$ features for that fern. Both intensity features and depth features are considered. Therefore, one fern can use intensity features and depth features at the same time, which effectively combine the best part of two sources of features. Figure 2 gives an example of such a fern. Further more, we use a parameter $\rho$ to control the ratio between intensity features and depth features, and use cross validation to select the best value for $\rho$.

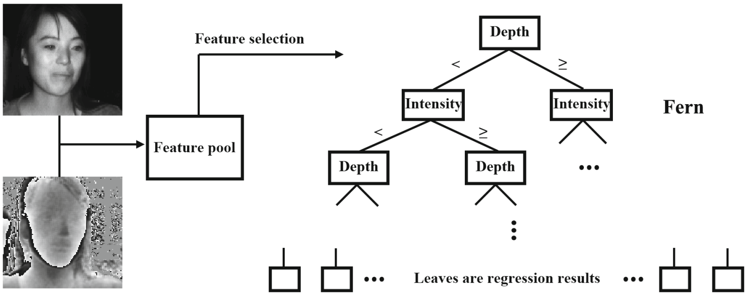

**Fig. 2.** Feature fusion. A fern can contain both intensity features and depth features at the same time.

### 2.4 Initial Estimation

An accurate initial estimation of facial landmarks positions can greatly contribute to face alignment. We use k-means on training data to get representative initial estimations of facial landmarks positions. Multiple initializations [4] are used in testing.

In a tracking scenario, a reasonable initial estimation is the face alignment result on previous frame. However, using that estimation directly will lead to model drift, since the model is trained with certain initial estimations obtained by k-means as described above. To solve drift problem, we use Procrustes method [8] to align initial estimations to face alignment result on previous frame, and then use those transformed initial estimations.

## 3 Experiments

We conduct experiments on two datasets to evaluate our model. Shape Root-Mean Square (RMS) error normalized w.r.t the inter-ocular distance of the face is used in the evaluation. In each experiment, we create three models: intensity, depth, intensity-depth. The only difference among those three models is feature. Depth model only uses depth features, intensity model only uses intensity features, and intensity-depth model is allowed to use both intensity features and depth features. All of the parameters are determined by cross validation on training data. To make the comparison fair, even though intensity-depth model is allowed to use both kinds of features, the numbers of total features which these three models are allowed to use are equal.

### 3.1 FRGC

FRGC (Face Recognition Grand Challenge) Version 2.0 database [11] is originally designed for face recognition, but some researchers [12,13] annotated those face images with 68 landmarks. It consists 4950 face images with color and depth information. We conduct this experiment in a similar manner (with minor difference) as [6], so that the results can be compared.

From Fig. 3, we can see that the performance of intensity model and intensity-depth model is on the same level, and both are much higher than that of the depth model. We also observe that the performance of our model is better than that of [6]. The average landmark errors of our intensity model, depth model and intensity-depth model are 0.0211, 0.0408 and 0.0209 respectively.

### 3.2 LIDF

The experiment on FRGC validates our model. However, the improvement by combining intensity feature and depth feature is very small. We believe that it is because of performance saturation (faces in FRGC are almost frontal and not

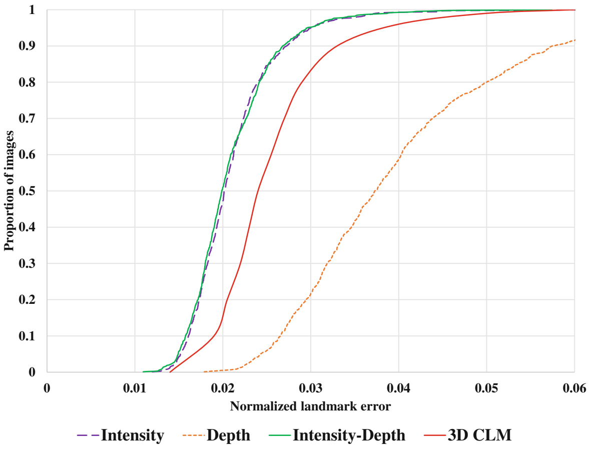

**Fig. 3.** Landmark error curves of our three models and 3D CLM [6] on FRGC.

challenging enough). To better evaluate the effect of combining intensity feature and depth feature, we create a more difficult dataset. Our labeled infrared-depth face (LIDF) dataset is acquired in a lab environment. It consists of 17 subjects (10 males and 7 females), each subject has 9 poses, and each pose has 6 expressions. Infrared image and depth image are taken simultaneously and are perfectly aligned (using Microsoft Kinect One). So we have 918 infrared-depth images in total. 15-points manually labeled landmarks are provided. This dataset is publicly available[1]. We use images from first 7 male subjects and first 4 female subjects as training set, and the rest images as testing set (Fig. 4).

From Fig. 5 we can see that the performance of intensity-depth model is notably higher than both intensity model and depth model. The average landmark errors of our intensity model, depth model and intensity-depth model are 0.0329, 0.0380 and 0.0305 respectively. Some alignment results obtained by our method are shown in Fig. 6.

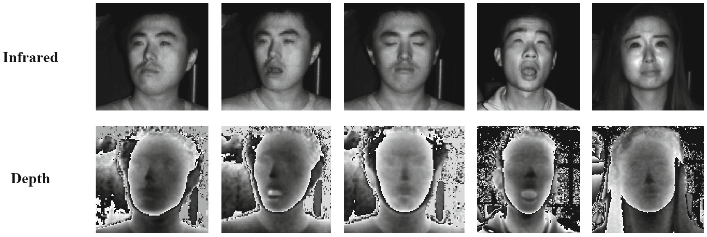

**Fig. 4.** Some images in LIDF.

[1] Available soon on http://bcmi.sjtu.edu.cn/resource.html.

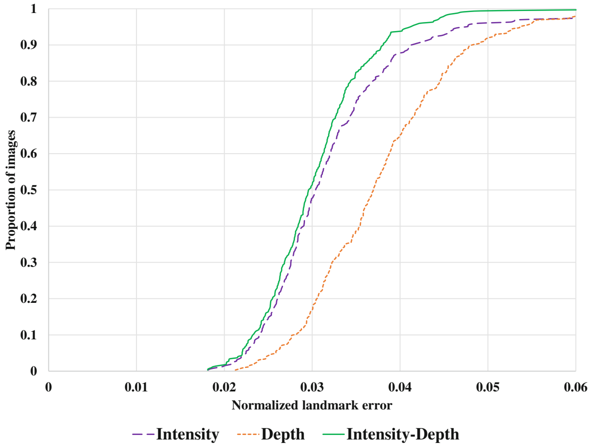

**Fig. 5.** Landmark error curves of our three models on LIDF.

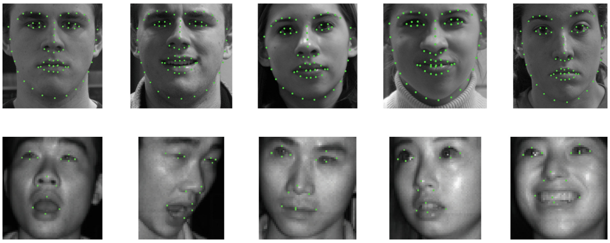

**Fig. 6.** Some alignment results of our intensity-depth model (first row: FRGC, second row: LIDF).

## 4 Conclusion

We proposed a face alignment model for intensity-depth image based on cascade regression model. Depth features were made to be robust against pose variation. Intensity and depth features were effectively combined during feature selection. Our intensity-depth model got 0.9 % error reduction on FRGC and 7.3 % error reduction on LIDF over intensity model, which indicated that higher performance could be achieved by combining intensity image and depth image.

**Acknowledgments.** This work was partially supported by the National Natural Science Foundation of China (Grant No. 61272248), the National Basic Research Program of China (Grant No. 2013CB329401), and the Science and Technology Commission of Shanghai Municipality (Grant No. 13511500200).

## References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1859–1866. IEEE (2014)
2. Baltrusaitis, T., Robinson, P., Morency, L.: 3D constrained local model for rigid and non-rigid facial tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2610–2617. IEEE (2012)
3. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision (ICCV), pp. 1513–1520. IEEE (2013)
4. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. (IJCV) **107**(2), 177–190 (2014)
5. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 109–122. Springer, Heidelberg (2014)
6. Cheng, S., Zafeiriou, S., Asthana, A., Pantic, M.: 3D facial geometric features for constrained local model. In: IEEE Conference on Image Processing (ICIP). IEEE (2014)
7. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **23**(6), 681–685 (2001)
8. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. (CVIU) **61**(1), 38–59 (1995)
9. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: British Machine Vision Conference (BMVC), vol. 2, p. 6 (2006)
10. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085. IEEE (2010)
11. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 947–954. IEEE (2005)
12. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 397–403. IEEE (2013)
13. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 896–903. IEEE (2013)
14. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3476–3483. IEEE (2013)