Facial Pose Estimation by Deep Learning from Label Distributions

Zhaoxiang Liu
CloudMinds
robin.liu@cloudminds.com

Zezhou Chen
CloudMinds
chenzezhou007@aliyun.com

Jinqiang Bai
Beihang University
baijinqiang@buaa.edu.cn

Shaohua Li
CloudMinds
shaohua.li@cloudminds.com

Shiguo Lian
CloudMinds
sg_lian@163.com

Abstract

Facial pose estimation has gained a lot of attentions in many practical applications, such as human-robot interaction, gaze estimation and driver monitoring. Meanwhile, end-to-end deep learning-based facial pose estimation is becoming more and more popular. However, facial pose estimation suffers from a key challenge: the lack of sufficient training data for many poses, especially for large poses. Inspired by the observation that the faces under close poses look similar, we reformulate the facial pose estimation as a label distribution learning problem, considering each face image as an example associated with a Gaussian label distribution rather than a single label, and construct a convolutional neural network which is trained with a multi-loss function on AFLW dataset and 300W-LP dataset to predict the facial poses directly from color image. Extensive experiments are conducted on several popular benchmarks, including AFLW2000, BIWI, AFLW and AFW, where our approach shows a significant advantage over other state-of-the-art methods.

1. Introduction

Facial pose estimation has received more and more attentions in the past few years [17, 28, 55, 41, 48, 47, 27, 8, 51, 56, 2, 53, 34, 4, 22], it plays an important role in many practical applications such as driver monitoring [17, 28], human-robot or human-computer interaction [55, 41, 48, 47, 50, 25, 7], gaze estimation [27, 8, 51, 56], human behavior analysis [2], face alignment [53, 6] and face recognition [5]. All of these unconstrained scenarios require a facial pose estimator which is resistant to environmental variations (*e.g.* occlusion, pose, illumination and resolution variations).

Though some good results have been made by using commercial depth cameras [12], one limitation that could not be neglected lies in that depth camera does not work well under uncontrolled environment where sunlight or ambient light is strong, and it often needs more space and more power compared to monocular RGB camera. These impede its feasibility in real-world applications [40, 34].

Traditionally, facial pose can be computed by estimating some facial key-points from target face and solving 2D to 3D correspondence with a mean 3D head model. Though facial key-point estimation has been recently improved greatly by deep learning [4], facial pose estimation is inherently a two-step process which is error-prone. The accuracy of the pose estimate depends upon the quality of key-points as well as the 3D head model. If the localized key-points are inaccurate or inadequate,the estimate of pose becomes poor or the pose estimation may even become infeasible. Additionally, generic 3D head models can also bring in errors for any given individual, and deforming the head model to adapt to each individual demands significant amounts of data and computation.

Recently, it has become more popular to estimate facial pose end to end using deep learning due to its robustness to environmental variations. The deep learning-based methods have large advantages compared to traditional landmark-to-pose methods, for they always output a pose prediction which does not rely on landmark detection and 3D head model. However, the deep learning-based facial pose estimation has not been thoroughly investigated overall. In some of these cases, facial pose estimation is just one branch of multi-tasks for face analysis, which is used to improve the performances of these other tasks (*e.g.* face detection, key-points localization and gender recognition). The facial pose branch was not designed dedicatedly in terms of accuracy. Some other deep learning-based methods have dedicatedly addressed the facial pose estimation as a pose regression from image [34, 22, 37, 36] using convolution neural networks(CNN), while the work in [40] has concluded that the combination of binned classification and regres-

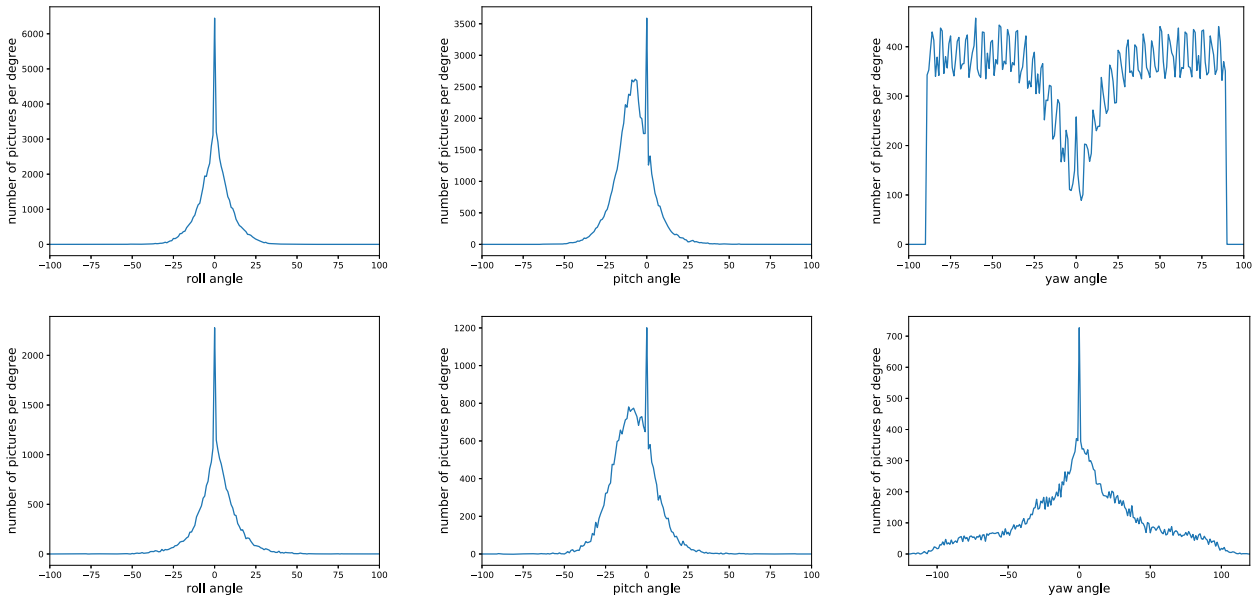

Figure 1. The sample distributions of two popular datasets. 300W-LP [59] (first row), AFLW [21] (second row). We can see that most faces lie in the area of small poses.

sion works better than regression solely. However, all these deep learning-based methods ignore an important fact that the distribution of training samples is quite imbalanced and there are not sufficient training samples for the large poses. To varying degrees, the most popular datasets for facial pose estimation, such as AFLW [21] and 300W-LP [59], exist this problem (as shown in Fig.1) which can degrade the accuracy of pose estimate, especially for large pose. We argue that it is unreasonable to use soft-max cross-entropy loss for facial pose estimation when training samples are considerably imbalanced, and the accuracy of facial pose estimate still has potential to be improved furtherly. For these losses ignore the similarity between adjacent poses, not taking the relationship between adjacent poses into consideration, other appropriate constraint should be introduced into the loss function.

To this end, we reformulate the facial pose estimation as label distribution learning problem and introduce a more intuitive similarity constraint: Gaussian label distribution loss into the training for facial pose estimation to improve the accuracy. The main contributions of our work can be summarized as follows:

- We reveal the fact that the lack of sufficient straining samples exists in the popular facial pose datasets. And we explain why it is not optimal to use soft-max cross-entropy loss for facial pose estimation under this situation.
- We introduce a novel Gaussian label distribution loss into the training for facial pose estimation, the Gaussian label distribution loss which constrains the similarities between neighbouring poses and can effectively mitigate the insufficiency of training samples, and dramatically boost the accuracy of facial pose estimate.
- We demonstrate the effectiveness of our method in facial pose estimation by various comparative experiments. Trained on publicly available datasets, such as AFLW [21]dataset and 300W-LP [59] dataset, our method achieves the-state-of-art results on AFLW2000 [59], BIWI [10], AFLW [21] and AFW [35] benchmarks.

2. Related Works

So far a variety of efforts on facial pose estimation have been dedicated. All these methods can be easily divided depending on whether they use 2D camera or depth camera. Since our work is concerned with deep learning-based method using RGB image from a monocular camera, any other methods using the depth camera will not be considered here. A more detailed description of depth camera-based methods can be found in a recent survey [30] and other previous works [27, 12, 3, 11, 54].

Some early classic studies [32, 43, 31] can be categorized as appearance template methods which match a view of a person's face to a set of exemplars with corresponding pose labels in order to find the most similar view. For example, the method in [31] adopts support vector machine (SVM) to model the appearance of human faces across multiple views and performs pose estimation by using nearest-neighbor matching. However, the appearance template

methods suffer from some limitations. They can only estimate discrete pose without the use of some interpolation method, and they also suffer from the accuracy concerns when the facial region is not localized accurately and efficiency concerns when the exemplar set is very large.

The face detector arrays [19, 57] whose idea is to train multiple face detectors for different facial poses once became popular as the success of frontal face detection [49, 33, 39]. The method in [57] uses a sequence of five multi-view face detectors to estimate facial pose. It is evident that many face detectors are required for each corresponding discrete facial pose, and it is difficult to implement a real time facial pose estimator with a large detector array.

Facial pose estimation can also be formulated as a manifold embedding problem that the high dimensional face image can be embedded into a low dimensional manifold in which facial pose is estimated. Any dimensionality reduction technique can be considered as a part of manifold embedding category. The methods in [42, 52] project a face image into a PCA or KPCA subspace and in which compare the result to a set of embedded templates. The method in [38] uses Isometric Feature Mapping (Isomap) to embed a face image into a nonlinear manifold which represents the pose-varying faces. These approaches ignore the pose labels that are available during training and operate in an unsupervised fashion. This results in that the built manifolds not only describe the pose variations but also identity variations [1]. The method [46] utilizes the feature correspondence of identity-invariant geometric features to learn a similarity kernel that only reflects the pose variation ignoring other sources of variation. This method shows a good reliability on benchmark dataset. However, further research is still needed to achieve state-of-the-art performance.

Facial pose estimation can be naturally formulated as a nonlinear regression problem which learns a nonlinear mapping from images to poses. The methods in [23, 26, 29] adopt support vector regressor(SVR) to estimate the facial pose after a series of preprocessing, including face region cropping, Sobel filtering, PCA [23], priori knowledge-based linear projection [26], or localized gradient orientation histogram [29]. The methods in [45, 44] utilize multilayer perception(MLP) to regress the facial pose. These methods have one disadvantage that they are prone to error from poor face localization. Recently thank to the great success of deep learning techniques, it has become popular to estimate facial pose using CNN which is robust to shift, scale and distortion. The method in [34] presents an in-depth study of CNN trained on AFLW dataset using L2 regression loss and tested on the Prima, AFLW and AFLW datasets. The method in [22] proposes a GoogLeNet-based architecture trained on AFLW dataset which can predict the key-points and facial pose jointly and reports the pose results on AFLW dataset and AFW [35]dataset. L2 Euclidean loss function is adopted to train the pose predictor which is used to improve key-point localization. The method in [53] also trains a pose estimator using 300W dataset to assist face alignment. Both the method in [37] and the method in [36] build a multitask learning framework for face analysis, including face detection, face alignment, face recognition, pose estimation, age prediction, gender recognition and smile detection. Both methods utilize AFLW dataset to train pose regressors and pose results are also reported on AFLW dataset and AFW dataset. The method in [40] makes an extensive study of combination of classification loss and regression loss on benchmark datasets, including 300w-LP dataset, AFLW dataset, BIWI [10] dataset and AFW dataset, and concludes that the combination of binned classification and regression works better than regression solely. However, all these deep learning-based methods pay no attention to the lack of sufficient training data for many poses. Consequently, the performance of facial pose estimator is limited. This reason motivates us to seek a better solution in this paper.

The label distribution learning (LDL) is a novel machine learning paradigm recently proposed for facial age estimation [16, 24]. The LDL is based on the observation that age is ambiguous and faces with adjacent ages are strongly correlated. The main idea of LDL is to utilize adjacent ages when learning a particular age. And a label distribution covers a number of class labels, representing the degree that each label describes the instance. Hence, the LDL is able to deal with insufficient and incomplete training data. Some other problems which share the same characteristic as facial age estimation, such as facial attractiveness computation [9], crowd counting [58] and pre-release movie rating prediction [14] have achieved outstanding performances by using LDL.

Facial pose appears similar to facial age, *i.e*. the faces under close poses look similar (as shown in Fig.2), the changing of facial pose can be regarded as a relative slow and smooth process and faces under adjacent poses are highly correlated. Thus, the LDL paradigm is an ideal match for the task of facial pose estimation. We notice that similar learning paradigms[15, 13] have been proposed to mitigate label ambiguity in head pose estimation. However, they only focused on 2D head pose estimation and were not extensively investigated on such precisely annotated benchmarks as AFLW2000 and BIWI.

3. Method

3.1. Gaussian Label Distribution Learning

We argue that the lack of sufficient training samples can degrade the accuracy of pose estimator. The reason is that the soft-max cross-entropy loss function used in training encodes the distance between all poses equally and does not

(a) $yaw = 24.8°$ (b) $yaw = 29.8°$ (c) $yaw = 34.8°$ (d) $yaw = 39.8°$

Figure 2. The faces of one subject under different poses.

take the relationship between adjacent poses into consideration. So it cannot effectively handle the insufficiency problem of training samples. Inspired by the previous work on age estimation [16, 24] and facial attractiveness ranking [9], we reformulate the facial pose estimation as a label distribution learning problem.

It is apparent that the faces under close poses look quite similar (as shown in Fig.2). Consequently, additional knowledge about the faces with different poses can be introduced to reinforce the learning problem. It is straightforward to utilize faces under neighboring poses while learning a particular pose. To achieve this, we assign a label distribution to each face image rather than a single label of real pose. This can make a face image contribute to not only the learning of its real pose, but also the learning of its neighbouring poses. We employ three Gaussian label distributions to describe a face example in the yaw, pitch and roll domain respectively to reinforce the whole learning process.

Here we take the yaw as an example to illustrate the Gaussian label distribution. Given a face image x_i and a complete set of yaw labels $\boldsymbol{y} = \{y_1, y_2, \ldots, y_M\}$, if its yaw label is $y_\alpha, \alpha = 1, 2, \ldots, M$, then the corresponding yaw label distribution is represented as a multi-dimension vector $\boldsymbol{D}_i^y = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \ldots, d_{x_i}^{y_M}\}$, with the l-th dimension as follows:

$$\begin{aligned} d_{x_i}^{y_l} &= \frac{exp\left(\overline{d_{x_i}^{y_l}}\right)}{\sum_{u=1}^{M} exp\left(\overline{d_{x_i}^{y_u}}\right)}, \\ \overline{d_{x_i}^{y_l}} &= exp(\frac{-(l-\alpha)^2}{2{\sigma^2}_y})/\sigma_y, l = 1, 2, \ldots, M \end{aligned} \tag{1}$$

where l denotes the l-th binned yaw, α is the binned ground-truth yaw, σ_y is the label standard deviation, and M is the dimension of the yaw label vector which also implicitly represents the maximum yaw. Consequently, $d_{x_i}^{y_l}$ represents the degree that the label y_l describes the example x_i under the constraint $\sum_{l=1}^{M} d_{x_i}^{y_l} = 1$, meaning that the label set $\boldsymbol{y}$ fully describes the example. Fig.3 demonstrates an example of Gaussian label distribution for yaw.

Following the same definition, another two label distributions: $\boldsymbol{D}_i^p = \{d_{x_i}^{p_1}, d_{x_i}^{p_2}, \ldots, d_{x_i}^{p_N}\}$ and $\boldsymbol{D}_i^r = \{d_{x_i}^{r_1}, d_{x_i}^{r_2}, \ldots, d_{x_i}^{r_K}\}$ can be obtained for x_i with a set of pitch labels $\boldsymbol{p} = \{p_1, p_2, \ldots, p_N\}$ and a set of roll labels

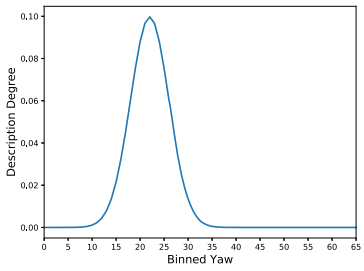

Figure 3. Gaussian label distribution with $\sigma_y = 4$ for the ground-truth $yaw = -30°$.

$\boldsymbol{r} = \{r_1, r_2, \ldots, r_k\}$ respectively as follows:

$$\begin{aligned} d_{x_i}^{p_j} &= \frac{exp\left(\overline{d_{x_i}^{p_j}}\right)}{\sum_{v=1}^{N} exp\left(\overline{d_{x_i}^{p_v}}\right)}, \\ \overline{d_{x_i}^{p_j}} &= exp(\frac{-(j-\beta)^2}{2{\sigma^2}_p})/\sigma_p, j = 1, 2, \ldots, N \end{aligned} \tag{2}$$

$$\begin{aligned} d_{x_i}^{r_k} &= \frac{exp\left(\overline{d_{x_i}^{r_k}}\right)}{\sum_{w=1}^{K} exp\left(\overline{d_{x_i}^{r_w}}\right)}, \\ \overline{d_{x_i}^{r_k}} &= exp(\frac{-(k-\gamma)^2}{2{\sigma^2}_r})/\sigma_r, k = 1, 2, \ldots, K \end{aligned} \tag{3}$$

where β and γ denote binned ground-truth pitch and roll of the face respectively. Consequently, the training set can be represented as $\{(x_i, (\boldsymbol{D}_i^y, \boldsymbol{D}_i^p, \boldsymbol{D}_i^r)), 1 \leq i \leq n\}$ and the goal of the learning becomes to train a set of network parameters $\boldsymbol{\theta}$ to generate a triplet of probability distribution $(\boldsymbol{F}^y(x_i; \boldsymbol{\theta}), \boldsymbol{F}^p(x_i; \boldsymbol{\theta}), \boldsymbol{F}^r(x_i; \boldsymbol{\theta}))$ for the three label sets, which is similar to $(\boldsymbol{D}_i^y, \boldsymbol{D}_i^p, \boldsymbol{D}_i^r)$. Wherein,

$$\begin{aligned} &\boldsymbol{F}^y(x_i; \boldsymbol{\theta}) = \{f(y_1|x_i; \boldsymbol{\theta}), f(y_2|x_i; \boldsymbol{\theta}), \ldots, f(y_M|x_i; \boldsymbol{\theta})\}, \\ &\textstyle\sum_{l=1}^{M} f(y_l|x_i; \boldsymbol{\theta}) = 1; \\ &\boldsymbol{F}^p(x_i; \boldsymbol{\theta}) = \{f(p_1|x_i; \boldsymbol{\theta}), f(p_2|x_i; \boldsymbol{\theta}), \ldots, f(p_N|x_i; \boldsymbol{\theta})\}, \\ &\textstyle\sum_{j=1}^{N} f(p_j|x_i; \boldsymbol{\theta}) = 1; \\ &\boldsymbol{F}^r(x_i; \boldsymbol{\theta}) = \{f(r_1|x_i; \boldsymbol{\theta}), f(r_2|x_i; \boldsymbol{\theta}), \ldots, f(r_K|x_i; \boldsymbol{\theta})\}, \\ &\textstyle\sum_{k=1}^{K} f(r_k|x_i; \boldsymbol{\theta}) = 1. \end{aligned} \tag{4}$$

The Euclidean distance and Kullback-Leibler (KL) divergence are adopted to construct the loss function measuring the similarity between the ground-truth distribution $(\boldsymbol{D}_i^y, \boldsymbol{D}_i^p, \boldsymbol{D}_i^r)$ and predicted distribution $(\boldsymbol{F}^y(x_i; \boldsymbol{\theta}), \boldsymbol{F}^p(x_i; \boldsymbol{\theta}), \boldsymbol{F}^r(x_i; \boldsymbol{\theta}))$. The objective of

the label distribution learning is to minimize either of the following overall loss functions:

$$L_{Eu} = \sum_{i=1}^{n} \left\| \boldsymbol{D}_i^y - \boldsymbol{F}^y\left(x_i;\boldsymbol{\theta}\right)\right\|_2 + \sum_{i=1}^{n} \left\| \boldsymbol{D}_i^p - \boldsymbol{F}^p\left(x_i;\boldsymbol{\theta}\right)\right\|_2 + \sum_{i=1}^{n} \left\| \boldsymbol{D}_i^r - \boldsymbol{F}^r\left(x_i;\boldsymbol{\theta}\right)\right\|_2,$$
$$L_{KL} = \sum_{i=1}^{n}\sum_{l=1}^{M} d_{x_i}^{p_l} \ln \frac{d_{x_i}^{p_l}}{f(y_l|x_i;\boldsymbol{\theta})} + \sum_{i=1}^{n}\sum_{j=1}^{N} d_{x_i}^{p_l} \ln \frac{d_{x_i}^{p_j}}{f(p_j|x_i;\boldsymbol{\theta})} + \sum_{i=1}^{n}\sum_{k=1}^{K} d_{x_i}^{p_l} \ln \frac{d_{x_i}^{r_k}}{f(r_k|x_i;\boldsymbol{\theta})} \tag{5}$$

And we define $L_{GLD} = L_{Eu} + L_{KL}$ as our Gaussian label distribution loss.

3.2. Network Architecture

We modify the framework presented in Hopenet [40] to construct our network architecture for facial pose estimation. The framework presented in Hopenet [40] originally consists of three separate losses for yaw, pitch and roll respectively and got state-of-the-art result. Each loss is a linear combination of a soft-max cross-entropy loss and a mean squared error(MSE) loss. To achieve better accuracy, we replace the soft-max cross-entropy loss with our Gaussian label distribution loss. Consequently, our learning architecture can be constructed as shown in Fig.4.

Our framework consists of a ResNet50 [18]-based backbone network and three branches for yaw, pitch and roll respectively. Each branch is comprised of a fully-connected layer with the number of neurons equal to the total number of corresponding labels and a soft-max layer followed by the combined loss layer. The soft-max operation ensures to satisfy the aforementioned constraints: $\sum_{l=1}^{M} d_{x_i}^{y_l} = 1$, $\sum_{j=1}^{N} d_{x_i}^{p_j} = 1$ and $\sum_{k=1}^{K} d_{x_i}^{r_k} = 1$.

Then the total loss is defined as $L_{total} = L_{GLD} + \alpha * L_{MSE}$. Wherein, L_{MSE} is the mean squared error loss, and α is a weight used to adjust the two loss components.

4. Experiments

4.1. Training Details

We choose the 300W-LP [59] and the AFLW [21] to train our network respectively. These two datasets have enough examples with enough different identities and different lighting conditions. The 300W-LP [59]dataset is a collection of popular in-the-wild 2D landmark datasets which have been grouped and re-annotated. The AFLW [21]dataset, which is commonly used to train and test landmark detection methods, also includes pose annotations.

We divide the facial pose into 66 bins within $\pm 99^\circ$ for yaw, pitch and roll respectively, *i.e.*, $M = N = K = 66$. And we set $\sigma_y = \sigma_p = \sigma_r = 4$. All the data is normalized before training by using the ImageNet mean and standard deviation for each color channel. And a pretrained ResNet50[18] on ImageNet is adopted to initialize our network. The proposed multi-loss network is trained with $\alpha = 0$, $\alpha = 0.01$, $\alpha = 0.1$, $\alpha = 1$ and $\alpha = 2$ on both the 300W-LP dataset and AFLW dataset. All the ten networks are trained using Adam optimization [51] with a learning rate of 10^{-6} and $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$.

4.2. Results on AFLW2000 and BIWI Benchmark

The AFLW2000 [59] dataset contains the first 2000 identities of the in-the-wild AFLW [21]dataset with accurate pose annotations. It is an ideal candidate to test our method. The BIWI [59] dataset is collected indoor by recording RGB-D video of different subjects across different facial poses using Kinect v2 device. It is commonly used as benchmark for depth-based pose estimation. Here we will only use the color frames instead of the depth information.

Firstly, we compare our results to the state-of-the-art method Hopenet [40] which is trained using a combination of L2 Euclidean loss and soft-max cross-entropy loss. Then, we compare to the pose estimated from 3DDFA [60] whose primary task is to align facial landmarks, and pose estimated from landmarks using two different landmark detectors: FAN [4] and Dlib[20], and ground-truth landmarks on both datasets. Additionally, we also list the results of KEPLER [22] on BIWI dataset reported in [40]. Table 1 shows the performance evaluations on AFLW2000 and BIWI Benchmark.

We can see that our best model($\alpha = 0.01$) outperforms all other baseline methods by a large margin on AFLW2000 benchmark, reducing the yaw error of the best-performing baselines 3DDFA[60] by 43.9%, reducing the yaw error of Hopenet[40] by 53.2%, reducing the pitch error, the roll error, and the mean average error (MAE) of the best-performing baseline Hopenet[40] by 22.8%, 32.2%, 36.2% respectively.

On BIWI benchmark, our method also performs better than all other baseline methods. Our best model($\alpha = 0$)trained on 300W-LP dataset reduces the error of the corresponding best-performing baseline Hopenet[40] trained on 300W-LP datatset by 14.3%, 15%, 3.7% and 12.3% for yaw, pitch, roll and MAE respectively. Our best model($\alpha = 0.1$)trained on AFLW dataset also outperforms Hopenet[40] trained on AFLW datatset, reducing the error by 20.8%, 19.4%, 0.9% and 13.8% for yaw, pitch, roll and MAE respectively.

4.3. Results on AFLW and AFW Benchmark

In this section, we present the evaluation results on AFLW [21] and AFW [35] benchmark, using the model trained on AFLW dataset. The AFW [35]benchmark which is commonly used to test landmark detection methods con-

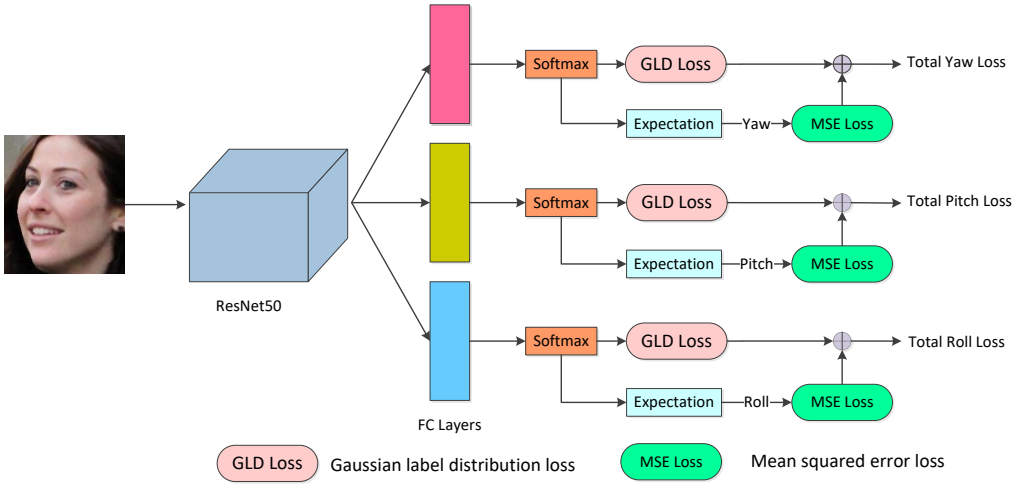

Figure 4. Proposed network architecture for facial pose estimation.

Benchmark	Method	Yaw	Pitch	Roll	MAE
AFLW2000	Hopenet[40]*	6.470	6.559	5.436	6.155
	FAN[4]	6.358	12.277	8.714	9.116
	3DDFA[60]	5.400	8.530	8.250	7.393
	Dlib[20]	23.153	13.633	10.545	15.777
	Ground-truth landmarks	5.924	11.756	8.271	8.651
	Ours($\alpha = 0$)*	3.1791	5.3372	3.7983	4.1049
	Ours($\alpha = 0.01$)*	3.0288	5.0634	3.6842	**3.9255**
	Ours($\alpha = 0.1$)*	3.1446	5.2047	3.6901	4.0131
	Ours($\alpha = 1$)*	3.1064	5.3446	3.6957	4.0489
	Ours($\alpha = 2$)*	3.3236	5.3570	3.8392	4.1733
BIWI	Hopenet[40]*	4.810	6.606	3.269	4.895
	Hopenet[40]+	5.785	11.726	8.194	8.568
	FAN[4]	8.532	7.483	7.631	7.882
	3DDFA[60]	36.175	12.252	8.776	19.068
	Dlib[20]	16.756	13.802	6.190	12.249
	KEPLER[22]+	8.084	17.277	16.196	13.852
	Ours($\alpha = 0$)*	4.1233	5.6142	3.1469	**4.2948**
	Ours($\alpha = 0.01$)*	4.2367	5.8446	3.4675	4.5163
	Ours($\alpha = 0.1$)*	4.0967	6.0498	3.2933	4.4799
	Ours($\alpha = 1$)*	3.9236	5.8832	3.4014	4.4027
	Ours($\alpha = 2$)*	4.6890	6.1271	3.3669	4.7276
	Ours($\alpha = 0$)+	4.5674	10.0874	8.0633	7.5737
	Ours($\alpha = 0.01$)+	4.5652	8.9595	8.7420	7.4223
	Ours($\alpha = 0.1$)+	4.5839	9.4471	8.1225	**7.3845**
	Ours($\alpha = 1$)+	4.3564	9.2310	8.8810	7.4895
	Ours($\alpha = 2$)+	4.3587	9.9015	8.6058	7.6220

*: trained on 300W-LP dataset.
+: trained on AFLW dataset.

Table 1. Evaluations on AFLW2000 and BIWI benchmarks.

Method	Yaw	Pitch	Roll	MAE
Hopenet[40]	6.26	5.89	3.82	5.324
KEPLER[22]	6.45	5.85	8.75	7.017
Patacchiola,Cangelosi[34]	11.04	7.15	4.4	7.530
Ours($\alpha = 0$)	6.83	5.26	3.92	5.34
Ours($\alpha = 0.01$)	6.00	5.31	3.75	**5.02**
Ours($\alpha = 0.1$)	5.93	5.30	4.03	5.085
Ours($\alpha = 1$)	5.90	5.51	3.87	5.094
Ours($\alpha = 2$)	5.90	5.62	3.77	5.097

Table 2. Evaluation on AFLW benchmark.

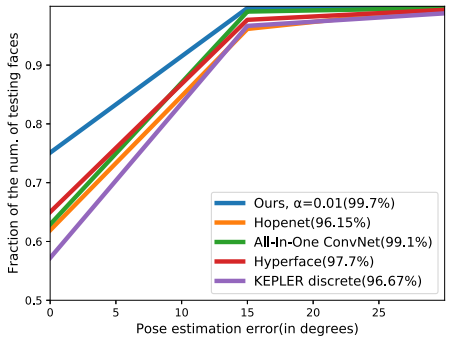

Figure 5. Evaluation on AFW benchmark.

tains rough pose annotations. Here, we compare our results to some deep learning-based methods, including Hopenet [40], KEPLER [22], the method proposed by Patacchiola and Cangelosi[34], Hyperface [36] and All-In-One [37]. Table 2 and Fig.5 respectively show the results on AFLW and AFW benchmark.

We can see that our method outperforms all other baseline methods on AFLW benchmark. Our best model($\alpha = 0.01$) reduces the error of the best-performing baseline Hopenet[40] by 4.2%, 9.85%, 0.53% and 5.71% for yaw, pitch, roll and MAE respectively. On AFW benchmark, our method also performs better than all other baseline methods. Our best model($\alpha = 0.01$) achieves a saturated accuracy of over 99%.

It is noteworthy that, on BIWI and AFLW benchmarks, the improvement of accuracy for roll is much less than for yaw and pitch. We argue that two reasons result in this situation. One reason is that, the distribution of training sets in roll domain is extremely imbalanced compared to that in yaw and pitch domains(as shown in Fig.1), and the most of

training examples lie in the area of small roll, which limits the learning ability of our method in roll domain, especially in the area of large roll. The other reason is that the test sets also have the similar characteristic as the first reason mentioned. In test sets, 67.65% examples of BIWI and 65.57% examples of AFLW lie in $\pm 10^{\circ}$ for roll, while 33.54% of BIWI and 26.23% of AFLW for yaw, and 22.97% of BIWI and 47.13% of AFLW for pitch. That is, the BIWI and AFLW benchmarks have relatively few examples with large roll. Both reasons restrict the improvement our method can make for roll.

5. Conclusion

This paper presents a novel computational model for facial pose estimation, which is reformulated as label distribution learning problem rather than the conventional single-label supervised learning. This makes a face image contribute to not only the learning of its real pose, but also the learning of its adjacent poses, mitigating the degradation of pose predictor caused by the lack of sufficient training data. Experiments on several popular benchmarks show our method is state-of-the-art.

References

[1] V. N. Balasubramanian, J. Ye, and S. Panchanathan. Biased manifold embedding: A framework for person-independent head pose estimation. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007. 3

[2] R. H. Baxter, M. J. Leach, S. S. Mukherjee, and N. M. Robertson. An adaptive motion model for person tracking with instantaneous head-pose features. *IEEE Signal Processing Letters*, 22(5):578–582, 2015. 1

[3] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister. Real-time face pose estimation from single range images. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 1, 5, 6

[5] K. Cao, Y. Rong, C. Li, X. Tang, and C. Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2018. 1

[6] F.-J. Chang, A. Tuan Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Faceposenet: Making a case for landmark-free face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1599–1608, 2017. 1

[7] Z. Chen, Z. Liu, H. Hu, J. Bai, S. Lian, F. Shi, and K. Wang. A realistic face-to-face conversation system based on deep neural networks. *arXiv preprint arXiv:1908.07750*, 2019. 1

[8] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):43, 2017. 1

[9] Y.-Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, and S. Z. Li. Label distribution-based facial attractiveness computation by deep residual learning. *IEEE Transactions on Multimedia*, 20(8):2196–2208, 2018. 3, 4

[10] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013. 2, 3

[11] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *CVPR 2011*, pages 617–624. IEEE, 2011. 2

[12] G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint Pattern Recognition Symposium*, pages 101–110. Springer, 2011. 1, 2

[13] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017. 3

[14] X. Geng and P. Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 3

[15] X. Geng and Y. Xia. Head pose estimation based on multivariate label distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1837–1842, 2014. 3

[16] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013. 3, 4

[17] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):1239–1258, 2009. 1

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines (svm). In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, volume 1, pages 154–156. IEEE, 1998. 3

[20] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 5, 6

[21] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011. 2, 5

[22] A. Kumar, A. Alavi, and R. Chellappa. Kepler: keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 258–265. IEEE, 2017. 1, 3, 5, 6

[23] Y. Li, S. Gong, J. Sherrah, and H. Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, 22(5):413–427, 2004. 3

[24] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015. 3, 4

[25] Z. Liu, H. Hu, Z. Wang, K. Wang, J. Bai, and S. Lian. Video synthesis of human upper body with realistic face. *arXiv preprint arXiv:1908.06607*, 2019. 1

[26] H. Moon and M. L. Miller. Estimating facial pose from a sparse representation, Apr. 28 2009. US Patent 7,526,123. 3

[27] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015. 1, 2

[28] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 709–714. IEEE, 2007. 1

[29] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 709–714. IEEE, 2007. 3

[30] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2009. 2

[31] J. Ng and S. Gong. Composite support vector machines for detection of faces across views and pose estimation. *Image and Vision Computing*, 20(5-6):359–368, 2002. 2

[32] S. Niyogi and W. T. Freeman. Example-based head tracking. In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 374–378. IEEE, 1996. 2

[33] E. Osuna, R. Freund, F. Girosi, et al. Training support vector machines: an application to face detection. In *cvpr*, volume 97, page 99, 1997. 3

[34] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017. 1, 3, 6

[35] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. Citeseer, 2012. 2, 3, 5

[36] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019. 1, 3, 6

[37] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017. 1, 3, 6

[38] B. Raytchev, I. Yoda, and K. Sakaue. Head pose estimation by nonlinear manifold learning. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 462–466. IEEE, 2004. 3

[39] H. A. Rowley. Neural network-based face detection. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1999. 3

[40] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 1, 3, 5, 6

[41] E. Seemann, K. Nickel, and R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 626–631. IEEE, 2004. 1

[42] J. Sherrah, S. Gong, and E.-J. Ong. Understanding pose discrimination in similarity space. In *BMVC*, pages 1–10, 1999. 3

[43] J. Sherrah, S. Gong, and E.-J. Ong. Face distributions in similarity space under varying head pose. *Image and Vision Computing*, 19(12):807–819, 2001. 2

[44] R. Stiefelhagen. Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data. In *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, volume 1, pages 21–24, 2004. 3

[45] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing. In *International Multimedia Conference: Proceedings of the seventh ACM international conference on Multimedia(Part 1)*, volume 30, pages 3–10. Citeseer, 1999. 3

[46] K. Sundararajan and D. L. Woodard. Head pose estimation in the wild using approximate view manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 50–58, 2015. 3

[47] Y.-J. Tu, C.-C. Kao, and H.-Y. Lin. Human computer interaction using face and gesture recognition. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–8. IEEE, 2013. 1

[48] Y.-J. Tu, C.-C. Kao, H.-Y. Lin, and C.-C. Chang. Face and gesture based human computer interaction. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(9):219–228, 2015. 1

[49] P. Viola, M. Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1:511–518, 2001. 3

[50] Z. Wang, Z. Liu, Z. Chen, H. Hu, and S. Lian. A neural virtual anchor synthesizer based on seq2seq and gan models. *arXiv preprint arXiv:1908.07262*, 2019. 1

[51] U. Weidenbacher, G. Layher, P. Bayerl, and H. Neumann. Detection of head pose and gaze direction for human-computer interaction. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 9–19. Springer, 2006. 1

[52] J. Wu and M. M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138–1158, 2008. 3

[53] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. *arXiv preprint arXiv:1507.03148*, 2015. 1, 3

[54] Y. Yu, K. A. F. Mora, and J.-M. Odobez. Robust and accurate 3d head pose estimation through 3dmm and online head model reconstruction. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 711–718. Ieee, 2017. 2

[55] D. Zanatto, M. Patacchiola, J. Goslin, and A. Cangelosi. Priming anthropomorphism: Can the credibility of human-like robots be transferred to non-humanlike robots? In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 543–544. IEEE, 2016. 1

[56] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 1

[57] Z. Zhang, Y. Hu, M. Liu, and T. Huang. Head pose estimation in seminar room using multi view face detectors. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 299–304. Springer, 2006. 3

[58] Z. Zhang, M. Wang, and X. Geng. Crowd counting in public video surveillance by label distribution learning. *Neurocomputing*, 166:151–163, 2015. 3

[59] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 2, 5

[60] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 5, 6