2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) | 978-1-6654-1867-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAICA52286.2021.9497992

A Fast and Efficient Face Pose Estimation Algorithm for Mobile Device

Peng Yan, Binbin Wu, Xiaoqian Nie, Bingqi Wang
Shanghai Radio Equipment Research Institute
Shanghai, China
nuaawbb@163.com

Xianliang Wang, Ziwei Liu
Beijing Ivsign Technology Co., Ltd
Beijing, China
wangxianliang@hisign.com.cn

***Abstract*—This paper proposes a method for face pose estimation from a single image. Face pose estimation is mainly used to estimate the 3D Euler Angle of the input face region image and is widely applied in human-computer interaction, video surveillance and other fields. For the purpose of improving the speed and accuracy of face pose estimation, multifarious operators or modules are utilized to build an efficient convolution neural network. Moreover, a variety of methods for data enhancement are used while training the CNN. Experiments show that our proposed method has higher speed and accuracy compared with previous face pose estimation methods.**

Keywords—Face Pose Estimation, Convolution Neural Network

I. INTRODUCTION

The main purpose of face pose estimation is to obtain the angle information of face orientation. Generally it can be expressed in terms of quaternions, rotation vectors, rotation matrices, or Euler angles. And the pattern of Euler angles is more widely used as a result of its more readable. In this paper, the face posture information is represented by three Euler angles (i.e. Pitch, Yaw and Roll).

Face pose estimation can be applied to face recognition (such as face alignment, data processing), human-computer interaction (such as replacement of mouse or gamepad, fatigue driving detection), video surveillance (such as human behavior analysis and understanding) and other specific scenarios.

Essentially face pose estimation is a method of learning or looking for mapping rules from 2D space to 3D space. Based on this, a large category of face pose estimation algorithms is obtained, that is, algorithms that estimate 3D pose information through 2D labeled information[1, 2]. For example, the face points are calculated firstly and then a reference frame (key points of average face) is selected to use, and transformation matrix of key points and reference frame is calculated, then face pose estimation could be implemented by means of the iterative optimization algorithm. In sum, this kind of method for solving the problem of face pose estimation is based on the starting point of algorithm or mathematics and principles summary. The other category is to train a regressor in a data-driven way, which makes a direct prediction on the input face image blocks[3]. In conclusion, it is big data that the entry point of this kind of method to solve the problem of face pose estimation. In this kind of method, the design of regressors can be divided into two parts, namely, the construct of backbone network and the design of regressors. Among them, the backbone network design can be light or heavy. A larger network, such as VGG16[4], ResNet50[5], ResNet101[5] or InceptionV3[6], could be selected when it is deployed on the PC platform. However, a lightweight network such as MobileNet[7], ShuffleNet[8] or SqueezeNet[9] will be more appropriate if the model need to be deployed to an embedded device. For the design of regressors, the input data will generate a high-dimensional nonlinear feature after passing through the backbone network. There are two commonly used designs, namely, direct regression and multi-branch regression.

In the current main research, the key points and transformation matrix can be utilized to estimate the face angle with the help of Dlib[10],FAN[1],Landmarks[3] and other methods. Hopenet[3] method adopts a concise and robust way to determine the position. By training a multi-loss convolutional neural network, it directly uses RGB combined with classification and regression loss to predict Euler Angle. In the FSA-Caps[11] method, the image features are divided into several groups, and the features obtained after encoding are used as aggregation materials. In this method, more effective local features can be obtained through fine-grained mapping of pixel features.

There is significant untapped opportunity in terms of accuracy and speed of the existing face pose estimation method, which cannot fully meet the needs of mobile terminals. To solve this problem, this paper proposes an algorithm that based on neural network constructed by a variety of operators or modules and a direct regression to the face pose angle. Our method has good performance in speed, accuracy and other aspects.

II. RELATED WORK

A. Central Difference Convolution[12]

Central Difference Convolution, i.e. CDC) can capture a third of the detailed patterns by aggregating intensity and gradient information. The network constructed by CDC has stronger modeling ability than the one constructed by vanilla convolution.

The CDC greatly improves the ability of the network to represent detailed information and the robustness to environmental changes on the premise of not increasing parameters and computation basically.

The vanilla convolution is defined as:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \quad (1)$$

And the CDC convolution is defined as:

$$y(p_0) = \theta \cdot \sum_{p_n \in R} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)) + (1-\theta) \cdot \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \quad (2)$$

B. SE Block

In the structure presented in the article Squeeze-and-Excitation Networks[9] (or SENet, for short), Squeeze and Excitation are the two key operations to explicitly model the inter dependencies between the feature channels. The author proposes a novel feature re-calibration strategy to avoid the introduction of a new spatial dimension for feature channel fusion. In essence, it automatically calculates the importance degree of each feature channel through learning, and then strengthens the useful features according to this importance degree and inhibits the features that are less relevant to the current task.

C. Channel Shuffle Operation[8, 13]

Channel Shuffle Operation could solve the problem caused by the group convolution.

Group convolution is to group different feature maps of the input layer, and then do the convolution in each group with different convolution kernels, which will reduce the calculation cost of convolution.

D. Dilated Convolution[14]

In order to increase the receptive field and reduce the calculation cost in the neural network, down sampling (pooling or convolution with the stride=2) is always required. Although this operation can increase the receptive field, the spatial resolution is reduced at the same time. For the purpose of maintaining the resolution while enlarging the receptive field, dilated convolution can be utilized to construct the network.

There is a parameter for dilated convolution named dilation rate. When different dilation rates are set, the receiver field will be different. As a result, multi-scale information will be obtained.

III. Method

A. Pipelines

The network proposed in this paper is based on MobilenetV2,SE Block,CDC,dilated convolution and Channel Shuffle, and so on. And Global Depthwise Convolution[15] is used to replace the layer of global average pooling to enhance the ability of extracting the local features of the face region. The stage in Table I is constructed from the basic unit (BU for short) of Figure 1, the down sampling unit (DU for short) of Figure 2 and the SE Block (SE for short). GDConv8x8 in table I is a Global Depthwise Convolution with a 8×8 kernel size.

TABLE I OVERALL PIPELINE OF OUR NET.

Layer	Output size	KSize	Stride	Unit	Repeat	Output channels
Image	64×64			-		3
CDC1	32×32	3×3	2	-	1	16
Stage2	16×16		2	DU×1 SE×1	1	24
	16×16		1	BU×1 SE×1	2	
Stage3	8×8		2	DU×1 SE×1	1	48
	8×8		1	BU×1 SE×1	2	
Conv4	8×8	1×1	1	-	1	128
GDConv8x8	1×1	8×8	1	-	1	128
conv1x1	1×1	1×1	1	-	1	3

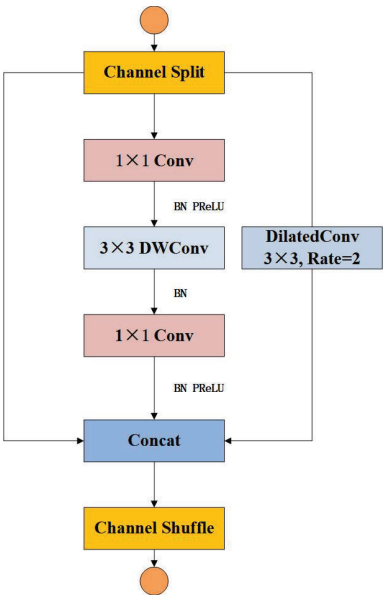

Fig. 1.The basic unit

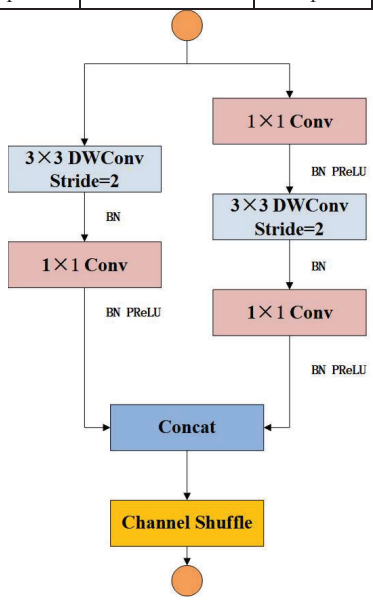

Fig. 2. The unit for spatial down sampling

B. Loss Function

The mean absolute error (MAE) is the mean of the distance between the predicted value f(x) of the model and the true value y of the sample. And the formula is shown as below:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - f(x_i)\right| \quad (3)$$

C. Data Enhancement

In order to make full use of the limited data and improve the generalization ability of the model, data enhancement operations are carried out, mainly including geometric transformation, adding noise, adding blur and so on.

Among them, geometric transformation mainly includes random horizontal flip, random vertical flip, random clipping, scaling and affine transformation.

Random noise is a random overlay of noise on the basis of the original picture. Noise mainly includes Gaussian noise, Coarse Dropout and Mix-up.

Random blur is to reduce the difference between the values of each pixel to achieve image blur and achieve pixel smoothing.

IV. Experiments

300W-LP dataset was used for training, while AFLW2000 dataset and BIWI dataset were used for testing.

Table II shows the experimental results that angle errors tested on the AFLW2000 data. And Table III shows the experimental results that angle errors tested on the BIWI data. Moreover, an ablation study was conducted for understanding the influence of individual components. Table IV presents the ablation results for the network built by different operators or modules and the one merged by all these operators and modules.

TABLE II. ANGLE ERRORS (TRAINED ON 300W-LP, TESTED ON AFLW2000)

	MB	Yaw	Pitch	Roll	MAE
Dlib(68 points)[10]	-	23.1	13.6	10.5	15.8
FAN(12 points)[1]	183	6.36	12.3	8.71	9.12
Landmarks[3]	-	5.92	11.86	8.27	8.65
Hopenet(a=2)[3]	95.9	6.47	6.56	5.44	6.16
SSR-Net-MD[16]	1.1	5.14	7.09	5.89	6.01
FSA-Caps (1×1)[11]	1.1	4.83	6.21	4.77	5.27
FSA-Caps-Fusion[11]	5.1	4.51	6.07	4.68	5.09
Ours	0.67	4.31	5.83	4.26	4.80

TABLE III. ANGLE ERRORS (TRAINED ON 300W-LP, TESTED ON BIWI)

	MB	Yaw	Pitch	Roll	MAE
Dlib(68 points)[10]	-	16.8	13.8	6.19	12.2
FAN(12 points)[1]	183	8.53	7.48	7.63	7.89
Hopenet(a=2)[3]	95.9	5.17	6.98	3.39	5.18
SSR-Net-MD[16]	1.1	4.49	6.31	3.61	4.65
FSA-Caps (1×1)[11]	1.1	4.77	6.26	3.32	4.78
FSA-Caps-Fusion[11]	5.1	4.28	4.95	2.76	4.00
Ours	0.67	4.22	4.83	2.61	3.89

TABLE IV. THE ABLATION STUDY FOR THE NETWORK BUILT BY DIFFERENT OPERATORS OR MODULES AND THE ONE MERGED BY ALL THESE OPERATORS AND MODULES.(TRAINED ON 300W-LP, TESTED ON AFLW2000 AND BIWI)

Testing set	operators or modules added	Model size (MB)	MAE	MAE (late fusion)
AFLW2000	-	0.35	6.22	
	CDC	0.35	5.85	4.80
	SE Block	0.35	5.78	
	Channel Shuffle	0.38	6.13	
	Dilated convolution	0.55	5.36	
	GDConv8x8	0.35	5.96	
BIWI	-	0.35	5.26	
	CDC	0.35	4.88	3.89
	SE Block	0.35	4.77	
	Channel Shuffle	0.38	5.04	
	Dilated convolution	0.55	4.73	
	GDConv8x8	0.35	4.91	

Table V shows the inference time of each model with CPU mode and pb format under the same configuration.

TABLE V. MODEL INFERENCE TIME

	Time/ms
SSR-Net-MD	6.5
FSA-Caps (1×1)	7.8
FSA-Caps-Fusion	10.1
Ours	2.4

In addition, the model with ONNX format takes less inference time that even less than 0.3ms.

V. Conclusion

In this paper, a neural network based on a variety of operators and modules, including depth separable convolution, SE Block, CDC, dilated convolution and Channel Shuffle, etc., is proposed to obtain real-time and efficient face pose estimation model. The experimental results show that our model has advantages of small size, fast speed and small error, and the algorithm is effective.

Acknowledgement

This work was supported by the foundation National key R&D plan of the Ministry of science and technology (Project Name: "Grid function expansion technology and equipment for community risk prevention", Project No. 2018YFC0809704).

References

[1] BULAT A, TZIMIROPOULOS G. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230, 000 3D Facial Landmarks) [M]. 2017: 1021-30.

[2] KUMAR A, ALAVI A, CHELLAPPA R. KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors [M]. 2017: 258-65.

[3] RUIZ N, CHONG E, REHG J M. Fine-Grained Head Pose Estimation Without Keypoints [M]. 2018: 2074-83.

[4] SIMONYAN K, ZISSERMAN A J C S. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. 2014

[5] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [M]. 2016: 770-8.

[6] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision [M]. 2016: 2818-26.

[7] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [J]. CoRR, 2017, abs/1704.04861

[8] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices [J]. 2017

[9] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-Excitation Networks [J]. IEEE Transactions on Pattern Analysis and Machine

Intelligence, 2020, 42(8): 2011-23.

[10] KAZEMI V, SULLIVAN J. One Millisecond Face Alignment with an Ensemble of Regression Trees [M]. 2014: 1867-74.

[11] YANG T, CHEN Y, LIN Y, et al. FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image; proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), F 15-20 June 2019, 2019 [C].

[12] YU Z, ZHAO C, WANG Z, et al. Searching Central Difference Convolutional Networks for Face Anti-Spoofing [M]. 2020: 5294-304.

[13] MA N, ZHANG X, ZHENG H-T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design [J]. 2018

[14] YU F, KOLTUN V. Multi-Scale Context Aggregation by Dilated Convolutions [J]. 2015

[15] SHENG C, YANG L, XIANG G, et al. MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices [J]. 2018

[16] YANG T-Y, HUANG Y-H, LIN Y-Y, et al. SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation [M]. 2018: 1078-84.