FDN: Feature Decoupling Network for Head Pose Estimation

Hao Zhang,[1] **Mengmeng Wang,**[1] **Yong Liu,**[1]* **Yi Yuan**[2]
[1]Institute of Cyber-Systems and Control, Zhejiang University, China
[2]NetEase Fuxi AI Lab
{zjucsezh, mengmengwang}@zju.edu.cn, yongliu@iipc.zju.edu.cn, yuanyi@corp.netease.com

Abstract

Head pose estimation from RGB images without depth information is a challenging task due to the loss of spatial information as well as large head pose variations in the wild. The performance of existing landmark-free methods remains unsatisfactory as the quality of estimated pose is inferior. In this paper, we propose a novel three-branch network architecture, termed as Feature Decoupling Network (FDN), a more powerful architecture for landmark-free head pose estimation from a single RGB image. In FDN, we first propose a feature decoupling (FD) module to explicitly learn the discriminative features for each pose angle by adaptively recalibrating its channel-wise responses. Besides, we introduce a cross-category center (CCC) loss to constrain the distribution of the latent variable subspaces and thus we can obtain more compact and distinct subspaces. Extensive experiments on both in-the-wild and controlled environment datasets demonstrate that the proposed method outperforms other state-of-the-art methods based on a single RGB image and behaves on par with approaches based on multimodal input resources.

Introduction

Facial analysis is one of the most studied topics in the past decades and plentiful methods have been proposed for various 2D facial problems, such as face detection, face recognition and face alignment. In recent years, 3D facial analysis has received more and more attention. Being an important basic of the nonverbal communication of humans, 3D head pose estimation can be utilized for human-computer interaction, human attention modelling, group behavior analysis, etc.

Since head pose estimation is a 3D problem, early methods (Martin, Van De Camp, and Stiefelhagen 2014; Meyer et al. 2015) using depth images can achieve high estimation performance. However, their applications are primarily limited by device, i.e., RGB-D cameras. Recently, with the breakthrough of deep convolutional neural networks (CNN), accurate facial landmark detection methods (Zhu et al. 2016; Bulat and Tzimiropoulos 2017) promote the performance of landmark-based head pose estimation from RGB images to a large extent and get rid of the limitation of devices. Nevertheless, such two-stage methods incur extra error and computation caused by landmark detection process which can not be neglected.

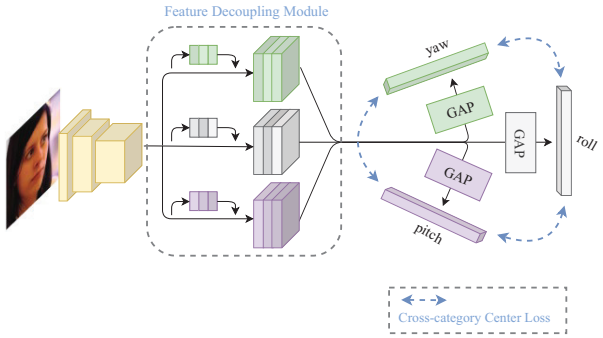

Figure 1: Overview of our FDN. The feature decoupling (FD) module is proposed to achieve the decoupled features of yaw, pitch, and roll, respectively, from the identical feature map input. GAP denotes global average pooling.

To tackle these problems, landmark-free methods (Ranjan, Patel, and Chellappa 2017; Ranjan et al. 2017) choose to regress three pose angles directly using features extracted from CNN. However, all these landmark-free methods regard head pose estimation as a straightforward regression or regression with classification problem under the assumption that the identical feature is applicable for the predictions of all three angles. Though existing landmark-free methods have achieved convincing performance in this manner, using identical feature is inappropriate as it still suffers from three main disadvantages. First, the feature subspaces of different angles are hybrid and non-separable, which leads to degradation of performance as features of yaw, pitch and roll have complementary but not identical information. Second, the exclusive features of each angle are not highlighted and are underutilized, which restricts the model from learning discriminative features for each angle. Third, it is inconsistent with human behavior as we will not focus on the same areas twice when asked to judge different pose angles (e.g., yaw angle vs. pitch angle) respectively from the same image.

*Corresponding author

To the best of our knowledge, none of the previous methods tries to deal with each angle on its characteristics, but to follow an image feature extraction and angle regression pipeline. In contrast, we argue that the feature subspaces of different angles are non-identical, and the customized features for each angle is beneficial to make the final predictions. Afterwards, we propose a novel three-branch network architecture in this paper, termed as Feature Decoupling Network (FDN), a more powerful architecture for landmark-free head pose estimation from a single RGB image, as shown in Figure 1. In our FDN, we add a feature decoupling (FD) module for the identical feature extracted from backbone CNN to learn discriminative features for each pose angle respectively. Furthermore, based on our proposed FDN, we propose a cross-category center (CCC) loss to implement decoupling of latent variable subspaces. With the novel three-branch architecture, our FDN has three main advantages. First, benefited from FD module and CCC loss, it realizes the decoupling of latent variable subspaces, which makes the model focuses on the prediction of each pose angle separately. Second, it learns to highlight exclusive features of each angle meanwhile suppress less useful ones which is beneficial for final predictions. Third, being consistent with human behavior, it is comprehensible.

In summary, the main contributions of our work can be summarized as follows:

- We propose a Feature Decoupling Network to achieve decoupling of latent variable subspaces to learn exclusive features of different angles in head pose estimation problem for the first time.
- We design a feature decoupling module to explicitly highlight discriminative features meanwhile suppress less useful ones by recalibrating channel-wise responses of each pose angle.
- We introduce a cross-category center loss to constrain the distribution of latent variable subspaces so that we can obtain more compact and distinct subspaces for different angle categories.
- Our proposed method outperforms other state-of-the-art methods based on a single RGB image and behaves on par with methods based on multimodal input resources.

Related Work

In this section, we provide a brief survey of head pose estimation and softmax-based loss functions.

Head Pose Estimation

Human head pose estimation has been a widely studied task in computer vision during the past few years with various methods and different modal databases proposed. Traditionally, RGB based methods (Murphy-Chutorian and Trivedi 2008; Huang, Shao, and Wechsler 1998) often use rotation-specific facial features to estimate head pose which is fragile due to various illumination conditions, expressions and occlusions. Some approaches based on depth images improve the robustness of estimation by registering a morphable face model to the depth images and combining several optimization methods (Meyer et al. 2015) or the ensembling of discriminative random regression forests (Fanelli et al. 2011).

In recent years, with the progress in RGB based facial analysis using convolutional neural networks (CNN), there are two main categories in existing methods for head pose estimation.

One is landmark-based which benefits from accurate facial landmark detectors (Bulat and Tzimiropoulos 2017; Zhu et al. 2016). Given 2D facial landmarks, we can obtain the head pose using algorithms such as CASSC (Gao et al. 2003). But such two-stage methods are indirect for head pose estimation. Zhu *et al.* (Zhu et al. 2016) directly fit a 3D face model using CNN and head pose is produced in the 3D fitting process.

The other is landmark-free, as landmark-based methods are affected by the accuracy of facial landmark detectors to a large extent and suffer from severe performance degradation in landmark invisible conditions as well as incur unnecessary computation. Closely related to other facial analysis tasks, head pose estimation is often performed in a multi-task learning framework. KEPLER (Kumar, Alavi, and Chellappa 2017) uses H-CNN to capture structured global and local features for accurate keypoint detection and provide head pose as a by-product. Hyperface (Ranjan, Patel, and Chellappa 2017) fuses the intermediate layers to perform face detection, landmarks localization, pose estimation and gender recognition simultaneously under a multi-task learning framework and demonstrates the synergy among these tasks. However, jointly learning with other tasks lacks the targeted research on head pose estimation problem. Methods such as (Ahn, Park, and Kweon 2014; Chang et al. 2017) exploit CNN architecture for head pose estimation in a regression manner. Drouard *et al.* (Drouard et al. 2017) propose to mix linear regressions with partially-latent output based on CNN. In (Ruiz, Chong, and Rehg 2018; Wang et al. 2019; Wang, Chen, and Zhou 2019), head pose estimation networks are trained under the joint supervision of classification loss and regression loss. Recently, FSA-Net (Yang et al. 2019) which employs the soft stagewise regression scheme and adopts a fine-grained structure aggregation outperforms the state-of-the-art methods on head pose estimation. In addition to single RGB images, Gu *et al.* (Gu et al. 2017) propose to use a recurrent neural network (RNN) for dynamic facial analysis in videos which is also landmark-free.

Softmax-based Loss Functions

In classification problems, a widely adopted pipeline which accepts class labels as supervision is softmax function followed by cross-entropy-loss to supervise the training process of the network. Based on softmax, advanced loss functions are proposed in face recognition. Schroff *et al.* (Schroff, Kalenichenko, and Philbin 2015) demonstrate the effectiveness of metric learning in face recognition by introduce the triplet loss. Wen *et al.* (Wen et al. 2016) propose the center loss which assigns embedding centers for each class and pulls the deep features closer to corresponding centers. Recently, some approaches take a deeper look into

the feature spaces distribution. Zhao *et al.* (Zhao, Xu, and Cheng 2019) employs an exclusive regularization to enlarge the angle between different classes. Duan *et al.* (Duan, Lu, and Zhou 2019) propose an equidistributed constraint in loss function to uniform the distribution of deep face features in feature space. Different from the above methods, our proposed loss function is cross-category as features distribute in different subspaces. Apart from face recognition, DEPICT (Ghasedi Dizaji et al. 2017) introduces Kullback-Leibler (KL) divergence to decrease the distance between prediction and target distribution in clustering problems based on softmax.

Method

In this section, we first formulate our problem. Then we elaborate the proposed Feature Decoupling Network for head pose estimation.

Problem Formulation

Given a set of training images $X = \{x^{(i)} \in \mathbb{R}^{3\times H\times W}\}_{i=1}^{N}$ and labels $Y = \{y^{(i)} \in \mathbb{R}^{3}\}_{i=1}^{N}$, where N is number of training images, $x^{(i)}$ is a RGB face image with size $H \times W$ and $y^{(i)}$ is a pose vector whose elements correspond to the angles of yaw, pitch, and roll respectively. Our goal is to model a mapping function ϕ by minimizing the mean absolute error (MAE) between the predictions and labels,

$$J(X) = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{y}^{(i)} - y^{(i)}\right\|_1 \tag{1}$$

where $\hat{y}^{(i)} = \phi(x^{(i)})$ is the output pose predictions by the model.

The Proposed Feature Decoupling Network

Overview of FDN Let $\mathbf{z}$ and $\mathbf{a} = [\mathrm{a}_{yaw}, \mathrm{a}_{pitch}, \mathrm{a}_{roll}]^{\top}$ denote the latent variable subspace and the pose angles respectively. Then, the pose prediction process can be viewed from the perspective of conditional probability, i.e., $p(\mathrm{a}_{yaw}, \mathrm{a}_{pitch}, \mathrm{a}_{roll} \mid \mathbf{z})$. Given $\mathbf{z}$, a_{yaw}, a_{pitch}, and a_{roll} are blocked from each other. Hence, previous works predict all three pose angles based on the identical feature provided by $\mathbf{z}$. In contrast, we propose to customize latent variable subspace for each pose angle. Afterwards, the prediction process turns into $p(\mathrm{a}_{yaw}, \mathrm{a}_{pitch}, \mathrm{a}_{roll} \mid \mathbf{z}_{yaw}, \mathbf{z}_{pitch}, \mathbf{z}_{roll})$, where $\mathbf{z}_{yaw}$, $\mathbf{z}_{pitch}$, and $\mathbf{z}_{roll}$ are latent variable subspaces for yaw, pitch, and roll respectively. Given $\mathbf{z}_j$, $j \in \{yaw, pitch, roll\}$, $\mathbf{a}$ is blocked from $\mathbf{z}$, which means $\mathbf{a}$ is dependent on $\mathbf{z}_j$, not $\mathbf{z}$. Therefore, the latent variable subspaces can be decoupled by appropriate $\mathbf{z}_j$.

As shown in Figure 1, the proposed FDN takes a single RGB image as input and pass it to the backbone CNN. The acquired feature map is then fed into the proposed FD module which we will describe in the next subsection. The output of branches are the final latent variables for predictions of different angles. The whole network can be trained in an end-to-end fashion.

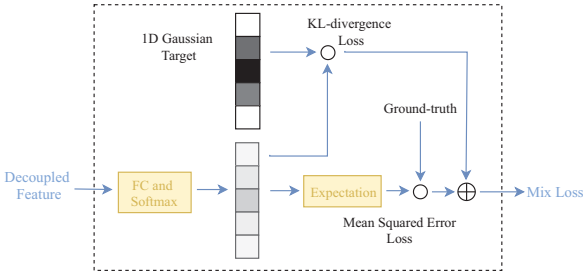

Figure 2: Our angle-dependent mix loss takes the decoupled feature as input and consists of the classification term and the regression term.

Feature Decoupling Module This module takes the acquired feature map as input and performs the decoupling of latent variable subspaces, as shown in Figure 1. Inspired by Squeeze-and-Excitation blocks (Hu, Shen, and Sun 2018) we implement our FD module with three channel attention blocks. It is a basic of CNN that features of different layers aim to encode different-level information. Pose information is encoded in high-layer features as low-layer features contain more detail information such as edges and texture. By adding FD module to the output feature map of backbone CNN, our FDN will not suffer from the problem of background clutter and semantic ambiguity which appears in low-level features. In FD module, we achieve the decoupling by adaptively recalibrating the channel-wise responses of each angle branch using parametric channel attention mechanism with a bottleneck layer consisting of two fully connected (FC) layers around the nonlinear function. Since channel dependencies are implicitly embedded in learned filters, our FD module explicitly captures discriminative features for each angle by performing angle-dependent feature re-weighting to select the more informative channel features while suppressing less useful ones and updating module parameters with angle-dependent losses.

Head pose estimation can be seen as a regression problem naturally. Previous works (Ruiz, Chong, and Rehg 2018; Wang, Chen, and Zhou 2019) show that the combined utilization of classification and regression supervision can further enhance the model performance. Thus we also follow this way to construct a mixed loss with two losses. As shown in Figure 2, the decoupled feature is then fed into an FC layer followed with a softmax function to obtain bins probability prediction, and the classification target is a 1D Gaussian with the mean centered at the ground-truth class and a small variance as consecutive discrete targets have explicit semantics in this case. We use Kullback-Leibler (KL) divergence between target and prediction distributions to compute classification loss. Then, the angle prediction is obtained by computing expectation of the bins output, followed with a Mean Squared Error (MSE) loss used as regression loss. The final mix loss for each angle branch is the following,

$$\mathcal{L}_{mix} = D_{\mathrm{KL}}\left(G(y_j)||-\log(q_j)\right) + \lambda MSE(y_j, \hat{y}_j) \tag{2}$$

where q_j is the output classification probability, y_j is the angle ground-truth, $G(\cdot)$ denotes 1D Gaussian of the class

target, $\hat{y}_j = q_j \cdot b$ is the final angle prediction, b is the indexes of bins, $j \in \{yaw, pitch, roll\}$ denotes each angle branch, and λ is a hyper parameter to trade off the classification loss term and the MSE loss term.

Cross-category Center Loss We further propose a cross-category center (CCC) loss to achieve the intra-class compactness and inter-category separability of latent variable subspaces at the same time. Similar to (Wen et al. 2016), the center loss part of CCC loss for each angle branch is defined as,

$$\mathcal{L}_c = \frac{1}{2}\sum_{i}^{m}\left\|\mathbf{z}_j^{(i)} - \mathbf{c}_j^{(y_i)}\right\|_2^2 \tag{3}$$

where $\mathbf{z}_j^{(i)} \in \mathbb{R}^d$ is the ith embedding deep feature, $\mathbf{c}_j^{(y_i)}$ is the embedding center of the y_ith class which is updated during training phase, m is the size of mini-batch and j denotes each angle branch.

The above part shortens the distance between latent variables of the same discrete angle ground-truth to ensure the intra-class compactness according to different angle categories. It is a fact that latent variables of different angles should distribute in decoupled subspaces. However, this is omitted in the center loss. To alleviate this shortcoming, we further define the decoupling loss part,

$$\mathcal{L}_d = \frac{1}{s(j,j') + s(j,j'') + 1} \tag{4}$$

where $j, j', j'' \in \{yaw, pitch, roll\}$ and $j \neq j' \neq j''$, $s(j,j') = \|\bar{\mathbf{c}}_j - \bar{\mathbf{c}}_{j'}\|_2$ and $s(j,j'') = \|\bar{\mathbf{c}}_j - \bar{\mathbf{c}}_{j''}\|_2$ denote the cross-category correlation distance, $\bar{\mathbf{c}}_j$ is the average of latent variable centers from each angle category and adding one in denominator is to prevent from results overflow.

The proposed cross-category center (CCC) loss consists of the above two parts and can be formulated as,

$$\mathcal{L}_{CCC} = \mathcal{L}_d + \alpha\mathcal{L}_c \tag{5}$$

where α is a hyper parameter to trade off the two parts.

Backbone Architecture and Optimization Taking feature extraction capability and model size into account, we construct our backbone network based on inverted residual blocks proposed in (Sandler et al. 2018). To be more specific, except for the initial convolution layer with 32 filters and the last one with 640 filters, our backbone network consists of 9 inverted residual blocks described in detail in Table 1 where we follow the notations in (Sandler et al. 2018).

Table 1: Detailed architecture of our backbone network.

Input	t	c	n	s
$112 \times 112 \times 32$	1	16	1	1
$112 \times 112 \times 16$	6	24	2	2
$56 \times 56 \times 24$	6	32	2	2
$28 \times 28 \times 32$	6	96	2	2
$14 \times 14 \times 96$	6	240	2	2

The total loss for each angle branch can be formulated as,

$$\mathcal{L}_j = \mathcal{L}_{mix} + \mathcal{L}_{CCC} \tag{6}$$

where $j \in \{yaw, pitch, roll\}$ denotes each angle branch. We update embedding centers with the following equation (Wen et al. 2016),

$$\Delta\mathbf{c}_{jk} = \frac{\sum_{i=1}^{m}\delta(y_i = k)\cdot(\mathbf{c}_{jk} - \mathbf{z}_j^{(i)})}{1 + \sum_{i=1}^{m}\delta(y_i = k)} \tag{7}$$

where jk denotes the kth center of the jth angle branch, $\delta(condition) = 1$ if the condition is true and $\delta(condition) = 0$ otherwise. Algorithm 1 details the training process of proposed method.

Algorithm 1 Feature Decoupling Network (FDN) Training

Require: Training data $\{x^{(i)}\}$, training labels $\{y^{(i)}\}$. The trade-off parameters λ, α, the learning rate μ_1 for update network parameters, the learning rate μ_2 for update embedding centers and the number of iteration T.
Ensure: The network parameters θ_C in backbone CNN, θ_{yaw}, θ_{pitch}, and θ_{roll} in FD module.
Initialize θ_c, θ_{yaw}, θ_{pitch} and θ_{roll}.
Initialize the centers $\mathbf{c}_{yaw}$, $\mathbf{c}_{pitch}$ and $\mathbf{c}_{roll}$ randomly.
for $t = 1, ..., T$ **do**
  Sample a mini-batch from the training set.
  for $j = yaw, pitch, roll$ **do**
    Compute the joint loss $\mathcal{L}_j = \mathcal{L}_{mix} + \mathcal{L}_{CCC}$.
    Update the θ_j via back-propagation.
    Update the centers $\mathbf{c}_j$ with Eq.7.
  end for
  Update the θ_c via back-propagation.
end for
return θ_C, θ_{yaw}, θ_{pitch} and θ_{roll}.

Experiments

In this section, our proposed FDN is systemically evaluated against several state-of-the-art methods. Both quantitative results and qualitative results are reported.

Implementation Details

All the images are cropped around the face to include the whole head. After being randomly cropped to 224×224, the images are normalized by ImageNet mean and standard deviation. The trade-off parameters λ, α are set to 2.5 and 0.01 respectively in all experiments. SGD optimizer is used to update centers with the learning rate 5×10^{-4} and Adam optimizer is used to update the network parameters with the learning rate 1×10^{-4}. Batch size is set to 16, and the network is trained for 100 epochs in total. All experiments are carried out based on Pytorch. Our method bins angles range from $-99°$ to $99°$, and we discard images with angles outside of this range following (Yang et al. 2019).

Datasets and Protocols

The 300W-LP dataset (Zhu et al. 2016) is a large synthetic dataset which was derived from 300W dataset (Sagonas et al. 2013). Zhu *et al.* re-annotated a collection of popular in-the-wild facial 2D landmark datasets by fitting the 3D dense

Figure 3: Head pose estimation results on the challenging AFLW2000 dataset. The blue, green and red lines point forward, downward and to the side respectively. The first row are the ground-truth. The second and the third row are the results of FSA-Net and our FDN, respectively. Best viewed in color.

Table 2: Comparisons with other state-of-the-art methods on the AFLW2000 and the BIWI dataset. All models are trained on the 300W-LP dataset.

Method	MB	AFLW2000				BIWI			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
Dlib (68 points) (Kazemi and Sullivan 2014)	-	23.1	13.6	10.5	15.8	16.8	13.8	6.19	12.2
FAN (12 points) (Bulat and Tzimiropoulos 2017)	183	6.36	12.3	8.71	9.12	8.53	7.48	7.63	7.89
Landmarks (Ruiz, Chong, and Rehg 2018)	-	5.92	11.86	8.27	8.65	-	-	-	-
3DDFA (Zhu et al. 2016)	-	5.40	8.53	8.25	7.39	36.2	12.3	8.78	19.1
Hopenet (α=2) (Ruiz, Chong, and Rehg 2018)	95.9	6.47	6.56	5.44	6.16	5.17	6.98	3.39	5.18
Hybrid Classification (Wang, Chen, and Zhou 2019)	96.7	4.82	6.23	5.14	5.40	-	-	-	-
FSA-Net (Yang et al. 2019)	5.1	4.50	6.08	4.64	5.07	**4.27**	4.96	2.76	4.00
Ours	5.8	**3.78**	**5.61**	**3.88**	**4.42**	4.52	**4.70**	**2.56**	**3.93**

face model to the image to construct the database which contains 61,225 samples across large poses and further expanded it to 122,450 samples with flipping. The AFLW2000 dataset (Zhu et al. 2016) contains the ground truth 3D faces and the corresponding 68 landmarks for the first 2,000 samples of the AFLW dataset (Koestinger et al. 2011). The faces in the dataset have large pose variations with various occlusions, expressions as well as illumination conditions. The BIWI dataset (Fanelli et al. 2013) provides pose annotations for roughly 15,000 frames derived from 24 videos of 20 subjects. Fanelli *et al.* used a Kinect v2 device to record RGB-D videos of different subjects in the controlled laboratory environment. Our experiments are conducted on these datasets following two widely used protocols described below.

In protocol 1, the 300W-LP dataset is only used for training and the trained models are evaluated on both of the AFLW2000 and the BIWI dataset. When testing on the AFLW2000 dataset, we retrieve the ground-truth landmarks to loosely crop the faces and when testing on the BIWI dataset, we employ dlib for face detection and do not using tracking. In protocol 2, we split videos in the BIWI dataset in a ratio of 7:3 for training and testing respectively following (Yang et al. 2019). Face bounding boxes in the BIWI dataset are detected by MTCNN face detector (Zhang et al. 2016).

Comparison with the State of the Art

Results with protocol 1. Table 2 shows the comparison with state-of-the-art methods, including landmark-based and landmark-free methods on two benchmark datasets. Compared with landmark-based methods (Kazemi and Sullivan 2014; Bulat and Tzimiropoulos 2017; Zhu et al. 2016), our FDN is landmark-free, thus the prediction error is not affected by landmark detection process and the model size can be very compact. In addition, benefited from the proposed feature decoupling strategy and the carefully designed model structure, our FDN is more accurate and lightweight than landmark-free methods (Ruiz, Chong, and Rehg 2018; Wang, Chen, and Zhou 2019). On the challenging AFLW2000 dataset, our FDN outperforms previous state-of-the-art methods such as FSA-Net (Yang et al. 2019) by 12.8% on MAE, which further exhibits the superiority of our FDN. It is the first time that a method proposes to learn customized features for each angle branch respectively and provides impressive performance improvements.

Table 3: Comparisons with other state-of-the-art methods on the BIWI dataset.

Method	Input modality RGB	Depth	Time	MB	Yaw	Pitch	Roll	MAE
DeepHeadPose (Mukherjee and Robertson 2015)	✓	-	-	-	5.67	5.18	-	-
SSR-Net-MD (Yang et al. 2018)	✓	-	-	1.1	4.24	4.35	4.19	4.26
VGG16 (Gu et al. 2017)	✓	-	-	500	3.91	4.03	3.03	3.66
FSA-Net (Yang et al. 2019)	✓	-	-	5.1	**2.89**	4.29	3.60	3.60
Ours	✓	-	-	5.8	3.00	3.98	2.88	3.29
DeepHeadPose (Mukherjee and Robertson 2015)	✓	✓	-	-	5.32	4.76	-	-
Martin (Martin, Van De Camp, and Stiefelhagen 2014)	✓	✓	-	-	3.6	**2.5**	**2.6**	**2.9**
VGG16+RNN (Gu et al. 2017)	✓	-	✓	>500	3.14	3.48	2.60	3.07

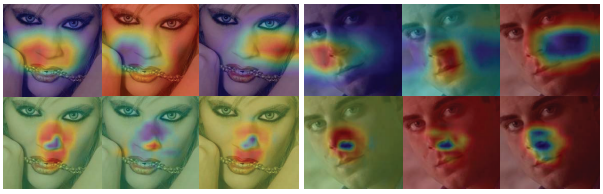

Figure 4: CAM visualization with two subjects from the AFLW2000 dataset. The first row are results of Hopenet and the second row are results of our FDN. Yaw, pitch, and roll, respectively in order from left to right. Best viewed in color.

Results with protocol 2. We also compare our approach with other methods that input with different modalities such as RGB, RGB-Depth, and RGB-Time while our FDN only uses a single RGB image on the BIWI dataset in Table 3. Our method outperforms the state-of-the-art RGB-based methods such as FSA-Net (Yang et al. 2019) by 8.6%. In addition, our method is more compact than VGG16 (Gu et al. 2017) and more concise than FSA-Net. DeepHeadPose (Mukherjee and Robertson 2015) estimates head pose in multi-modal RGB-D videos by combining classification and regression model. Martin (Martin, Van De Camp, and Stiefelhagen 2014) builds and registers a 3D head model to estimate head pose from depth images. VGG16+RNN (Gu et al. 2017) uses an end-to-end network containing a CNN and an RNN for head pose estimation from consecutive video frames. Compared to these multi-modal methods, our method only uses pixel intensity information and narrows the performance gap between RGB based and multi-modal inputs based methods.

Visualization

Figure 3 shows a few results of our FDN compared with the previous state-of-the-art method, i.e., FSA-Net. It can be seen that our FDN is more robust for various poses and lighting. In addition, Ruiz *et al.* also try to predict three angles separately in Hopenet but using hybrid features for various angles. We visualize the Class Activation Map (Zhou et al. 2016) of Hopenet and our proposed FDN, as shown in Figure 4. Compared with Hopenet, our FDN makes predictions

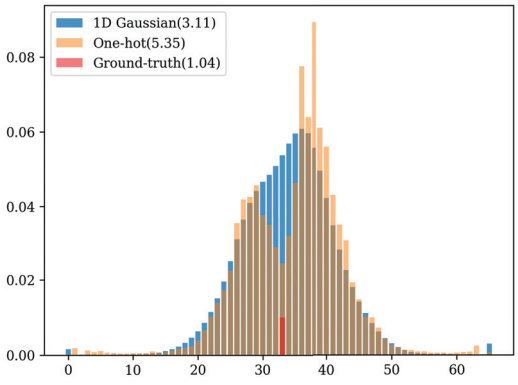

Figure 5: Comparison of probability distributions using various classification targets. The red one denotes the classification ground-truth. The numbers in parentheses indicate poses. Best viewed in color.

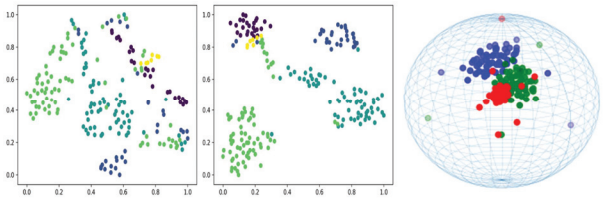

Figure 6: Visualization of the learned feature distributions on the AFLW2000 dataset. Left: comparison of features belonging to 5 different bins from Hopenet and our FDN respectively. Right: distribution of features belonging to various angle categories, i.e., yaw, pitch, and roll in hypersphere from our FDN. Best viewed in color.

based on more discriminative local regions meanwhile most global features are fully utilized for all three angles and the boundary between local and global regions output by our FDN is more stable. It proves that our FDN learns more distinct subspace for each angle. In Figure 5, we show that by using 1D Gaussian, our FDN can not only reduce evident

Table 4: Ablation study for FDN on the AFLW2000 and the BIWI datasets.

FD module	L_{CCC}	AFLW2000				BIWI			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
-	-	4.18	5.73	4.05	4.65	2.98	4.53	3.38	3.63
✓	-	3.97	5.71	4.00	4.56	3.13	3.94	3.28	3.45
✓	✓	3.78	5.61	3.88	4.42	3.00	3.98	2.88	3.29

Table 5: Evaluation of different backbone structures on the BIWI dataset.

Model Structure	MB	Yaw	Pitch	Roll	MAE
ResNet18	62.7	3.71	4.63	3.71	4.02
FDN	5.8	3.00	3.98	2.88	3.29

misclassification but also provide more accurate pose prediction. In Figure 6, we visualize feature distributions using the t-SNE (Maaten and Hinton 2008), which shows that our FDN achieves the decoupling of feature subspaces across angle categories meanwhile maintains the intra-class compactness for each bin.

Ablation Study

In this subsection, we conduct extensive ablation studies to further demonstrate the effectiveness of our Feature Decoupling Network (FDN).

Table 4 shows the experimental results on the AFLW2000 dataset and the BIWI dataset. In order to fairly compare, We adopt the same backbone network architecture as described in Table 1 across all three sets of experiments. The baseline model is in the 1st row which uses 3 FC layers to estimate three angles separately from the hybrid feature. Compared to the baseline model, adding our FD module to the hybrid feature and training the network with $\mathcal{L}_{mix}$ loss (2nd row) improves the performance by 1.9% and 4.9% on the AFLW2000 and the BIWI dataset respectively, which indicates that our FD module learns to customize discriminative features for each angle branch and is beneficial to estimate the angles. As shown in the 3rd row, adding our $\mathcal{L}_{CCC}$ loss to the model with FD module further enhance the performance by 3.0% on the AFLW2000 dataset and 4.6% on the BIWI dataset as the proposed $\mathcal{L}_{CCC}$ loss promotes the decoupling of latent variable subspaces with different angles on the basis of FD module. Overall, our FDN outperforms the baseline model by a large margin, i.e., 5.0% and 9.4% on the AFLW2000 and the BIWI dataset, respectively.

Furthermore, we evaluate the influence of different backbone structures in Table 5. We use ResNet18 with residual blocks for comparison. It shows that using inverted residual blocks is beneficial to the performance as well as the model size. However, the performance margin brought by our proposed method is still considerable. As shown in Table 4, our proposed FDN outperforms the baseline model by 9.4% on the BIWI dataset, which proves the effectiveness of our method.

Conclusion

In this paper, a novel three-branch network architecture, called Feature Decoupling Network (FDN), is proposed for landmark-free head pose estimation from a single RGB image. Different from previous works, We propose to decouple hybrid features and customize exclusive latent variable subspace of each angle by our proposed feature decoupling (FD) module and cross-category center (CCC) loss. In the FD module, we explicitly select discriminative features and suppress less useful ones of each angle by adaptively recalibrating its channel-wise responses. The CCC loss further improves the performance by encouraging more compact and distinct latent variable subspaces of different angle categories. Extensive experiments on the 300W-LP, the AFLW2000 and the BIWI datasets show the superiority of our FDN which results in the state-of-the-art performance on these datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant U1509210 and Key R&D Program Project of Zhejiang Province (2019C01004).

References

Ahn, B.; Park, J.; and Kweon, I. S. 2014. Real-time head orientation from a monocular camera using deep neural network. In *Asian conference on computer vision*, 82–96. Springer.

Bulat, A., and Tzimiropoulos, G. 2017. 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*, 1021–1030.

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

Drouard, V.; Horaud, R.; Deleforge, A.; Ba, S.; and Evangelidis, G. 2017. Robust head-pose estimation based on partially-latent mixture of linear regressions. *IEEE Transactions on Image Processing* 26(3):1428–1440.

Duan, Y.; Lu, J.; and Zhou, J. 2019. Uniformface: Learning deep equidistributed representation for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3415–3424.

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

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

Gao, X.-S.; Hou, X.-R.; Tang, J.; and Cheng, H.-F. 2003. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence* 25(8):930–943.

Ghasedi Dizaji, K.; Herandi, A.; Deng, C.; Cai, W.; and Huang, H. 2017. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, 5736–5745.

Gu, J.; Yang, X.; De Mello, S.; and Kautz, J. 2017. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1548–1557.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

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

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

Koestinger, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2011. 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)*, 2144–2151. IEEE.

Kumar, A.; Alavi, A.; and Chellappa, R. 2017. 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)*, 258–265. IEEE.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605.

Martin, M.; Van De Camp, F.; and Stiefelhagen, R. 2014. Real time head model creation and head pose estimation on consumer depth cameras. In *2014 2nd International Conference on 3D Vision*, volume 1, 641–648. IEEE.

Meyer, G. P.; Gupta, S.; Frosio, I.; Reddy, D.; and Kautz, J. 2015. Robust model-based 3d head pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3649–3657.

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

Murphy-Chutorian, E., and Trivedi, M. M. 2008. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 31(4):607–626.

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

Ranjan, R.; Patel, V. M.; and Chellappa, R. 2017. 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.

Ruiz, N.; Chong, E.; and Rehg, J. M. 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2074–2083.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 397–403.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Wang, Y.; Liang, W.; Shen, J.; Jia, Y.; and Yu, L.-F. 2019. A deep coarse-to-fine network for head pose estimation from synthetic data. *Pattern Recognition* 94:196–206.

Wang, H.; Chen, Z.; and Zhou, Y. 2019. Hybrid coarse-fine classification for head pose estimation. *arXiv preprint arXiv:1901.06778*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, 499–515. Springer.

Yang, T.-Y.; Huang, Y.-H.; Lin, Y.-Y.; Hsiu, P.-C.; and Chuang, Y.-Y. 2018. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, 7.

Yang, T.-Y.; Chen, Y.-T.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1087–1096.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23(10):1499–1503.

Zhao, K.; Xu, J.; and Cheng, M.-M. 2019. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1136–1144.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

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