RankPose: Learning Generalised Feature with Rank Supervision for Head Pose Estimation

Donggen Dai
daidonggen@gmail.com

Wangkit Wong
wwongaa@connect.ust.hk

Zhuojun Chen
georgechenzj@outlook.com

Ping An Technology
Shenzhen, China

Abstract

We address the challenging problem of RGB image-based head pose estimation. We first reformulate head pose representation learning to constrain it to a bounded space. Head pose represented as vector projection or vector angles shows helpful to improving performance. Further, a ranking loss combined with MSE regression loss is proposed. The ranking loss supervises a neural network with paired samples of the same person and penalises incorrect ordering of pose prediction. Analysis on this new loss function suggests it contributes to a better local feature extractor, where features are generalised to Abstract Landmarks which are pose-related features instead of pose-irrelevant information such as identity, age, and lighting. Extensive experiments show that our method significantly outperforms the current state-of-the-art schemes on public datasets: AFLW2000 and BIWI. Our model achieves significant improvements over previous SOTA MAE on AFLW2000 and BIWI from 4.50 [11] to 3.66 and from 4.0 [24] to 3.71 respectively. Source code is available at: https://github.com/seathiefwang/RankHeadPose.

1 Introduction

Head pose estimation has aroused much interest recently as it serves as auxiliary information improving the accuracy of many computer vision tasks such as face recognition, expression recognition and attention analysis. It aims at predicting mathematical representations for human head orientation (e.g. Euler angles, quaternion [11]). There exist accurate solutions [4, 18] that estimate head pose from depth images which are, however, not accessible in many real-world applications. Hence, building a model based on 2D images has emerged as a good trade-off between complexity and accuracy. There are two-stage methods [9] that predict facial landmarks first to approach head pose estimation. The accuracy of the second stage, i.e. head pose estimation, is heavily constrained by the first stage landmark prediction which compromises under conditions of poor lighting and large head pose. Considering that, DCNN methods based on dense face alignment and multi-task learning for estimating both facial landmarks and head pose are recently proposed [7, 20, 26]. But, such multi-objective

design has not focused on head pose, resulting in suboptimal pose estimation [21]. Instead, dedicated head pose estimation models combining L2 regression loss and specifically designed losses achieve superior performance over landmark-based schemes [21, 24].

In this paper, we conduct an in-depth study on 2D image head pose estimation. We design a novel deep learning approach to seek to improve current methods. Firstly, we reformulate the output function of a regression neural network in a normalisation manner to achieve a model closer to the head pose representation. In light of that Euler angles are bounded (within $[-\pi/2, \pi/2]$), we as well bound the model output space by L2 normalisation of deep head pose features F and the weight vectors W of the output layer. This results in a cosine transformation with which head pose is predicted as the projection of F on W (also known as cosine similarity between F and W). Further, we apply an inverse cosine function (Arccos) to arrive at an angle model specialised for Euler angle representation. By so doing, head pose is estimated as the angle between F and W. Our experimental results show this approach has substantial positive effect on head pose prediction accuracy.

More importantly, we introduce a new loss function for head pose estimation. We address the fact that a face image consists of two major components: global and local features. Pose-related features which a good model is expected to extract should be invariant with most of the global features such as lighting condition, biological appearance, facial expression, etc. Motivated by this, we employ a Siamese architecture and a novel loss function combining MSE regression loss and a pairwise ranking loss. The pairwise ranking loss takes two images of the same person as input and requires the network to rank angles (yaw, pitch, and roll) of them so as to predict larger values for the image with larger ground-truth label. Since for most training data, images of the same person are obtained under similar conditions, the neural network learns from pairs of inputs saturated with global similarities (pose-irrelevant futures such as identity and lighting) yet only differentiable by local features (e.g. nose shape, pupillary distance). In such cases, ordering of head pose angles highlights pose-related features. The ranking loss thus penalises incorrect ordering of head pose predictions of similar input images and the learnt model can be more generalisable to condition changes. We analytically show that local features which we consider as abstract landmarks are crucial to accurate prediction. We name our method RankPose.

The main contributions of this paper are summarised as follows.

- We propose a reformulation of head pose representation learning by adopting a transformation scheme. We show that transforming model output to angle space helps improve head pose prediction accuracy.

- We propose a novel CNN model exploring Siamese architecture and ranking loss to distinguish pose-related features from a mixture of pose-related and -irrelevant features.

- We improve the current state-of-the-art results on public benchmarks: AFLW2000 and BIWI significantly. Our model improves the MAE on AFLW2000 and BIWI from 4.50 [11] to 3.66 and from 4.0 [24] to 3.71 respectively.

2   Related Work

In this section, we briefly revisit previous works of head pose estimation. Previous methods used as comparison schemes in the Experimental section 4.3 are emphasised with bold text.

Methods using depth information: Martin et. al. [17] propose an iterative closest point algorithm to register depth image with a 3D head model. Initial pose is estimated using features of the head, and then the algorithm improves the current estimated pose over time. A similar method is proposed in [18], which combines particle swarm optimisation and the iterative closest point algorithm to eliminate the need for explicit initialisation. Random regression forests trained on synthetic data for head pose estimation are presented in [6]. Mukherjee et. al. [19] present a CNN-based model for head pose estimation running on RGB-D data, which is first trained to predict coarse gazing direction classes and then fine-tuned to regress head poses. Though these methods achieve accurate estimation, depths data are not accessible in many real-world applications.

Landmark-based methods: Early landmark-based works use the Perspective-n-Point (P-nP) algorithm [9] to estimate pose from a single RGB image. The PnP algorithm is 2-step where it first obtains 2D facial landmarks, and then establishes correspondence between the 2D landmarks and a generic 3D head model. Subsequently, face alignment is performed on the correspondence between the 2D landmarks and the 3D head pose is obtained from the rotation matrix. **Dlib** presented in [12] uses an ensemble of regression trees to localise facial landmarks with realtime prediction speed. Recently, methods that estimate 2D facial landmarks using modern deep learning networks [2, 8] have become dominant. **KEPLER** [14] predicts landmarks and pose simultaneously using an improved GoogLeNet architecture. Landmark localisation is improved by coarse pose supervision. **Two-Stage** [16] uses a deep regression network with two-stage re-initialisation for initial landmark localisation. However, the performance of sparse facial landmark localisation for face alignment often drops under large head pose and heavy occlusion. To address these issues, dense face alignment methods [7, 25] based on deep CNNs have been proposed. **3DDFA** [25] uses a cascaded CNN to fit a dense 3D face model to face images for face alignment. Head pose is obtained as a by-product of 3D face alignment. Posing facial feature analysis as a multi-task problem, Hyperface [20] learns a shared CNN to perform face detection, facial landmark localisation, head pose estimation, and gender recognition simultaneously. To improve landmark localisation accuracy for low-resolution images, **FAN** [3] simultaneously improves face resolution and detects the facial landmarks.

Landmark-free methods: Models designed for landmark localisation or dense face alignment may not have focused on head pose accuracy. Instead, landmark-free approaches have emerged as more promising schemes. Ruiz et. al. [21] propose **HopeNet**, a multi-loss CNN trained on synthetic data to predict intrinsic Euler angles from image intensities. The classification loss guides the networks to learn the ordering of neighbourhood of poses in a robust fashion. It leverages feature aggregation across different semantic levels using a fine-grained spatial structure and soft stagewise regression. **FSA-Net** [24] achieves state-of-the-art accuracy by adopting an attention mechanism. To avoid the ambiguity problem in the commonly used Euler angle representation, **QuatNet** [11] trains ResNet50 to predict the quaternion of head pose instead of Euler angles. QuatNet adopts a multi-regression loss function which addresses the non-stationary property in head pose estimation by ranking different intervals of the angles with a classification loss.

It is worth noting that our ranking supervision scheme is fundamentally different from QuatNet's in two aspects: 1) Our model utilises Siamese architecture to measure ranking errors between paired samples to focus on pose-related features while QuatNet puts ranking as more of a classification problem; 2) QuatNet aims at tackling the non-stationary property, whereas our RankPose makes effort to differentiate local changes.

3 Method

In this section, we introduce two key ingredients of the proposed RankPose in detail. The output space transformation is described in Section 3.1. Our loss function is described in Section 3.2. In Section 3.3, we provide an abstract landmark view to interpret our RankPose method.

3.1 Output Space Transformation

Most commonly, a CNN-based head pose regression model takes input image X and predicts head pose representation Y. Dissecting the network to a DCNN backbone (D) and the final output layer (FC), Y is predicted as dot product of a feature vector $F = D(X)$ produced by the backbone and the weight matrix W in FC:

$$Y_{dot} = W^T F. \tag{1}$$

In the above equation, Y_{dot} is unbounded ($Y_{dot}^d \in R$) but head pose representation should not. Euler angles, for instance, are bounded within $[-\pi/2, \pi/2]$ in this case. To rectify, we adopt a transformation scheme.

We first apply L2 normalisation to F and W, s.t. $\|F\| = 1$ and $\|W_j\| = 1$. Vector F and each column j of W are normalised to have unit length. The L2 normalised version of F and W are denoted as $\hat{F}$ and $\hat{W}$ respectively. In this way, Y is obtained as projections of $\hat{F}$ on $\hat{W}_j$ which is also the cosine similarity between $\hat{F}$ and $\hat{W}_j$ since $Y_j = \alpha(\|F\|\|W_j\|cos\theta_j) = \alpha cos\theta_j$, so it becomes:

$$Y_{cosine} = \alpha \hat{W}^T \hat{F}, \tag{2}$$

where α is a scalar that scales the output range from $[-1,1]$ to the head pose range. Here, $\alpha = \pi/2$ for Euler angles.

The above is the Cosine transformation which transforms the model output to the range of cosine function. Moreover, for Euler angle representation, substitution of α with the inverse cosine function ($Arccos$) changes the measurement into the geometric angle θ_j between $\hat{F}$ and $\hat{W}_j$ since $Y_j = Arccos(cos\theta_j) = |\theta_j|$ where, for $\theta_j \in [0, \pi]$, the Euler angles are linearly associated with the angles between $\hat{F}$ and $\hat{W}_j$. To obtain Euler angles in range $[-\pi/2, \pi/2]$, a $\pi/2$ shift to the function is further performed, and Y becomes:

$$Y_{arccos} = Arccos(\hat{W}^T \hat{F}) - \pi/2. \tag{3}$$

3.2 Pairwise Ranking Loss

As shown in Figure 1, the model consists of a CNN backbone. The backbone can be one of the modern deep convolutional neural networks such as VGG [22], GoogleNet [23] or ResNet [11]. The network forwards two images and always computes loss function of both at the same time, and in the backpropagation, the two images condition each other. This is known as the Siamese architecture [1]. In a training iteration, the network is trained with a mini-batch of N training samples. Each training sample $\{X_{i1}, X_{i2}\}$ is a pair of sampled training images of the same person, where i is the sample index. Subsequently, $\{Y_{i1}, Y_{i2}\}$ is the head pose prediction and $\{\tilde{Y}_{i1}, \tilde{Y}_{i2}\}$ is the ground-truth head pose. For Euler angles in our case, Y is a three dimensional vector representing yaw, pitch and roll respectively.
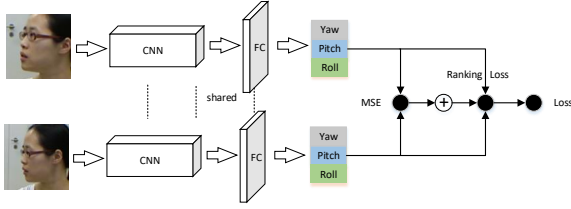
Figure 1: Illustration of our proposed model: Two images of the same person are passed through a CNN model, which predicts 3 angle values representing pitch, yaw, and roll respectively. The prediction is supervised with a loss function combining MSE loss and ranking loss.

In most regression models, mean squared error (MSE) is a close and effective loss function for modelling minimisation problem of a particular dataset. In our model, it is also necessary to make the predictions match the real head pose for any given image as much as possible. Therefore, one key objective is the following:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(\|\tilde{Y}_{i1} - Y_{i1}\|^2 + \|\tilde{Y}_{i2} - Y_{i2}\|^2 \right), \qquad (4)$$

where $\|.\|$ is the L2 norm. However, it is well-known that MSE is prone to overfit to outliers because it heavily weights them. As the model moves towards outliers, it learns pose-irrelevant features to memorise the dataset which makes it lack generalisation. To learn generalised pose-related features which are invariant to biological appearance and lighting condition etc., we impose the following pairwise ranking loss on the model.

$$L_{ranking} = \frac{1}{N} \sum_{i=1}^{N} \sum_{d} \left[\max \left\{ 0, -sgn(\tilde{Y}_{i2} - \tilde{Y}_{i1}) \cdot (Y_{i2} - Y_{i1}) \right\} \right]^d, \qquad (5)$$

where $sgn(z)$ is the sign function that returns 1 if z is greater than 0, returns -1 if z is less than 0, and returns 0 otherwise. Operation $\{\cdot\}$ is the element-wise product between vectors. $[v]^d$ is the dth element of vector v and $\sum_d [v]^d$ sums all elements of vector v. $\max \left\{ 0, -sgn(\tilde{Y}_{i2} - \tilde{Y}_{i1}) \cdot (Y_{i2} - Y_{i1}) \right\}$ is a 3-dimensional vector, whose elements represent the pairwise ranking loss of yaw, pitch, and roll respectively. Its dth element is,

$$\left[\max \left\{ 0, -sgn(\tilde{Y}_{i2} - \tilde{Y}_{i1}) \cdot (Y_{i2} - Y_{i1}) \right\} \right]^d = \begin{cases} \max \left\{ 0, Y_{i2}^d - Y_{i1}^d \right\} & \text{if } \tilde{Y}_{i2}^d \leq \tilde{Y}_{i1}^d \\ \max \left\{ 0, Y_{i1}^d - Y_{i2}^d \right\} & \text{if } \tilde{Y}_{i1}^d \leq \tilde{Y}_{i2}^d \end{cases}, \qquad (6)$$

which requires the prediction of yaw (or pitch or roll) angle, Y_{i1}^d of image X_{i1} to be larger than the prediction of Y_{i2}^d, if the yaw (or pitch or roll) ground-truth label $\tilde{Y}_{i1}^d$ of image X_{i1} is larger. We supervise the network with a novel loss function combining MSE loss and pairwise ranking loss, given by:

$$L = \beta L_{MSE} + (1 - \beta) L_{ranking}, \qquad (7)$$

where β trades off the importance between the MSE loss and the ranking loss.

3.3 Visualisation & Abstract Landmark Interpretation

Facial landmark-based methods rely purely on local conditioning features, i.e. the exact landmark locations. Their success suggests that local information is good representation
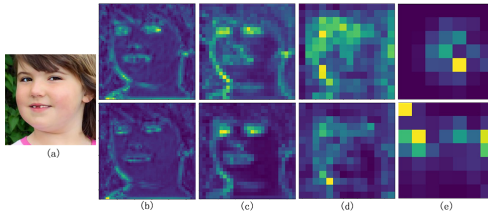
Figure 2: An input image and its corresponding activation maps from the first, second and third residual blocks. The upper row is produced by RankPose, the lower being of a model learnt with MSE loss.
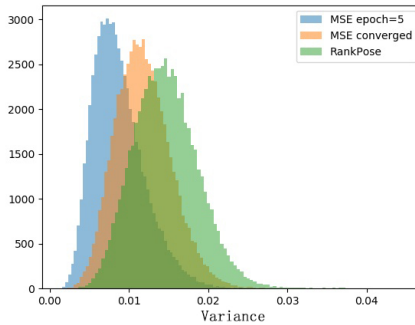

Figure 3: Distribution of variance of the activation maps produced by the first residual block

for head pose. Efficient as they remain in practice, landmark-based methods appear to underwhelmingly perform when a pose regression CNN is a feasible option. Needless to say, CNN can learn strong abstract features, but we observe that an accurate model tends to be more confined to local features such as the face silhouette and edges of eyes and mouth. We consider those features as the abstract landmarks which can be much more informative than the explicit landmark coordinates.

Figure 2 visualises the activation maps learnt by RankPose. Activation maps from the first, second, and third residual blocks of ResNet50 are visualised in the top row of Fig.(b), Fig.(c), Fig.(d) and Fig.(e) respectively. Correspondingly, the activation maps of a model trained with MSE loss without using transformation are visualised in the bottom row. Activation values are normalised to $[0, 1]$. Brighter colours specify greater values. It can be seen from the figure that RankPose produces sharper edges compared to the MSE model. RankPose learns stronger activation near abstract facial landmarks. We believe that facial landmarks are generalisable features for head pose estimation, which justifies that RankPose learns more generalised features.

Figure 3 shows the variance of the activation maps produced by the first residual block. Variance is computed as $\sum_{i=1}^{n} \frac{1}{n}(pix_i - \overline{pix})^2$, where pix_i is the value of the ith pixel of the activation map and $\overline{pix}$ is the mean value. The peak in blue colour is of an MSE model at an early training stage. Keeping training the model to convergence results in the peak in salmon colour. The peak in green where there is the highest variance distribution is of RankPose. Better models exhibit clearly different statistics over activation variances. Higher variance indicates a better local feature extractor. It suggests RankPose learns more discriminative features.

Table 1: MAE Comparison on AFLW2000.

Method	Yaw	Pitch	Roll	Avg. MAE
Dlib [12]	23.1	13.6	10.5	15.8
3DDFA [25]	5.40	8.53	8.25	7.39
FAN [3]	6.36	12.3	8.71	9.12
HopeNet [21]	6.47	6.56	5.44	6.16
FSA-Net [24]	4.50	6.08	4.64	5.07
QuatNet [11]	3.97	5.62	3.92	4.50
RankPose (ours)	**2.99**	**4.75**	**3.25**	**3.66**

4 Experiments

In this section, we present implementation, datasets, evaluation protocols, benchmark results and ablation studies.

4.1 Implementation

To make fair comparison with previous works, we build our model on top of a ResNet50 backbone pre-trained on ImageNet. The last FC layer is replaced by new output layer with weights of shape $[2048, 3]$. Unless otherwise stated, as shown in Section 3.1, each column of the weight matrix is normalised to have unit length. Random cropping, random scale, Gaussian noise, and Gaussian blur are adopted as standard augmentation pipeline. The model is trained by Adam optimisation for 80 epochs, with an initial learning rate of 0.001 and 128 batch size. Cosine decay [15] is applied to the learning rate during training. In all experiments, we sample randomly two images of the same identity to form a training sample $\{X_{i1}, X_{i2}\}$. β in the loss function (Equation 7) is set to 0.5.

4.2 Datasets & Evaluation Protocols

Training dataset: RankPose is trained on 300W-LP [25]. 300W-LP is synthesised from 300W dataset by a 3DMM model. The synthesis is done by predicting the depth of face image. Then, the profile views are generated by 3D rotation of the model.

Testing datasets: Two datasets are used for testing: AFLW2000 [25] and BIWI [6] datasets. The AFLW2000 dataset consists of the first 2000 images in AFLW [13], which collects face images from the internet. It is a test set considering different identities, ambience, and quality, as well as large head poses. BIWI is a dataset of 24 video sequences. They record 20 persons rotating across large pose degrees (75, 60 and 50 for yaw, pitch and roll respectively) about one metre in front of a Kinect V2 camera. 15000 frames are sampled from those videos.

Evaluation protocol: We follow the protocol of Hopenet [21] which is also a landmark-free head pose estimation method. It is trained on the 300W-LP dataset and tested on the two datasets: the AFLW2000 and the BIWI datasets. When testing on BIWI, MTCNN is used as face detector.

4.3 Benchmark Results

We compare RankPose with the following state-of-the-art methods: Dlib [12], 3DDFA [25], FAN [3], KEPLER [14], Two-Stage [16], HopeNet [21], FSA-Net [24], and QuatNet [11]. Brief description of these methods can be found in the related work section 2.

Table 1 compares RankPose with other state-of-the-art methods on the AFLW2000 dataset. Following previous works, the mean absolute error (MAE) is computed as evaluation metric. We report the predicted MAE of yaw, pitch, and roll angles of all methods in the second,
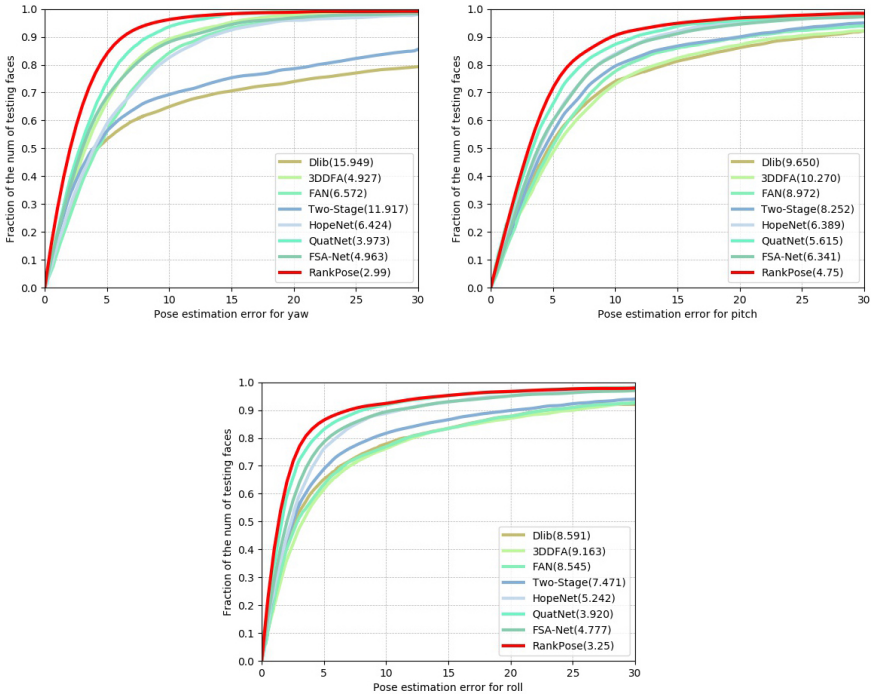
Figure 4: Error distribution in terms of yaw, pitch and roll.

third, and forth columns respectively. The average value of MAE of 3 angles is reported in the last column. As expected, deep learning-based approaches (3DDFA, FAN, HopeNet, FSA-Net, QuatNet and RankPose) perform better than the method based on the regression tree (Dlib). We can observe that landmark-free methods (HopeNet, FSA-Net, QuatNet, and RankPose) outperform landmark-based methods (Dlib, 3DDFA, and FAN). Our proposed RankPose outperforms all the comparison methods in all dimensions significantly. Remarkably, RankPose improves the current SOTA result (achieved by QuatNet) from 4.50 to 3.66, which is improvement of 18.66%.

Error distribution in terms of yaw, pitch and roll is plotted in figure 4. The cumulative error curves of RankPose are much skewer than other comparison methods. There are significant margins between the curves of RankPose and that of other schemes. Using our method, most test samples (98.7%) are with predicted yaw error less than $15°$; 97.25% are with pitch error less than $20°$; 97.15% are with roll error less than $20°$.

Table 2 compares RankPose with other state-of-the-art methods on the BIWI dataset. The characteristics of the training set and BIWI are quite different. Similar to the result on the AFLW2000 dataset, landmark-free methods perform better than landmark-based methods. Our RankPose outperforms other comparison schemes significantly showing that it generalises better across deferent domains. Specifically, RankPose improves the current SOTA result (achieved by FSA-Net) from 4.0 to 3.71, which is 7.25% improvement.

Table 2: MAE Comparison on BIWI.

Method	Yaw	Pitch	Roll	Avg. MAE
Dlib [□]	16.8	13.8	6.19	12.2
3DDFA [□]	36.2	12.3	8.78	19.1
FAN [□]	8.53	7.48	7.63	7.89
KEPLER [□]	8.80	17.3	16.2	13.9
Two-Stage [□]	9.49	11.34	6.00	8.94
HopeNet [□]	5.17	6.98	3.39	5.18
FSA-Net [□]	4.27	4.96	2.76	4.00
QuatNet [□]	4.01	5.49	2.94	4.15
RankPose (ours)	**3.59**	**4.77**	**2.76**	**3.71**

Table 3: Ablation studies of different loss functions & output transformations.

Dataset	AFLW2000				BIWI			
Method	Yaw	Pitch	Roll	Avg. MAE	Yaw	Pitch	Roll	Avg. MAE
MSE	3.52	6.56	4.49	4.85	4.70	10.18	3.36	6.08
MSE+Cosine	3.5	5.58	4.24	4.44	4.42	6.75	3.38	4.85
MSE+Arccos	3.51	5.52	4.18	4.40	4.77	5.60	3.08	4.48
Ranking loss+MSE	2.90	5.13	3.68	3.91	3.73	5.69	2.79	4.07
Ranking loss+MSE + Cosine	2.96	4.70	3.37	3.68	3.64	4.8	2.73	3.72
Ranking loss+MSE + Arccos	2.99	4.75	3.25	3.66	3.59	4.77	2.76	3.71

4.4 Ablation studies

We conduct ablation studies to understand the influence of individual components. Results are reported in Table 3. Two different loss functions: 'MSE' (Equ. 4) and 'ranking loss+MSE' (Equ. 7) are studied; two different transformation schemes: Cosine transformation (Equ. 2) and Arccos transformation (Equ. 3) are compared. We first compare a baseline model ('MSE') without any transformation with a model using Cosine transformation ('MSE+Cosine') and a model using Arccos transformation ('MSE+Arccos'). Models with transformations show consistently higher performance. Particularly, the model with Arccos transformation performs the best. We see the transitive property also applies to the models with ranking loss. We next compare models ('MSE'/'MSE+Cosine'/'MSE+Arccos') without ranking loss and models ('Ranking loss+MSE'/'Ranking loss+MSE + Cosine'/'Ranking loss+MSE + Arccos') with ranking loss. Nevertheless, it is most indubitable here that ranking loss+MSE always outperforms MSE alone by a large margin. The absence of Cosine/Arccos transformation does not affect it being state-of-the-art.

5 Conclusion

In this paper, we have discussed the challenges encountered by current head pose estimation methods. We propose to transform head pose representation learning to a bounded angle space to boost prediction accuracy. More importantly, we propose a model named RankPose to learn generalised abstract landmarks with ranking loss. With these improvements, we train a model which achieves state-of-the-art results on public benchmarks. Hereby, we deliver our work in a hope that deeper understanding of head pose estimation will arise in future discussions.

References

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[2] Adrian Bulat and Georgios 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.

[3] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark lo-calization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018.

[4] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, 17(11):2049–2058, 2015.

[5] Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time head pose estimation with random regression forests. In *CVPR 2011*, pages 617–624. IEEE, 2011.

[6] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International journal of computer vision*, 101(3):437–458, 2013.

[7] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face recon-struction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[8] Aryaman Gupta, Kalpit Thakkar, Vineet Gandhi, and PJ Narayanan. Nose, eyes and ears: Head pose estimation by locating facial keypoints. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1977–1981. IEEE, 2019.

[9] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quat-net: Quaternion-based head pose estimation with multiregression loss. *IEEE Transac-tions on Multimedia*, 21(4):1035–1046, 2018.

[12] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensem-ble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[13] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark local-ization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[14] Amit Kumar, Azadeh Alavi, and Rama 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.

[15] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[16] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017.

[17] Manuel Martin, Florian Van De Camp, and Rainer Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 641–648. IEEE, 2014.

[18] Gregory P Meyer, Shalini Gupta, Iuri Frosio, Dikpal Reddy, and Jan Kautz. Robust model-based 3d head pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3649–3657, 2015.

[19] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.

[20] Rajeev Ranjan, Vishal M Patel, and Rama 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, 2017.

[21] Nataniel Ruiz, Eunji Chong, and James 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.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[24] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. 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*, pages 1087–1096, 2019.

[25] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan 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.

[26] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017.