

# HYBRID COARSE-FINE CLASSIFICATION FOR HEAD POSE ESTIMATION

Haofan Wang<sup>1,2,\*</sup>, Zhenghua Chen<sup>2</sup>, Yi Zhou<sup>1</sup>

School of Information Science and Technology, Dalian Maritime University<sup>1</sup>, Horizon Robotics<sup>2</sup>

## ABSTRACT

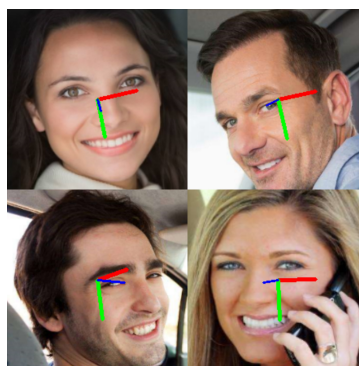
Head pose estimation, which computes the intrinsic Euler angles (yaw, pitch, roll) from a target human head, is crucial for gaze estimation, face alignment and 3D reconstruction. Traditional approaches to head pose estimation heavily relies on the accuracy of facial landmarks, and solve the correspondence problem between 2D facial landmarks and a mean 3D head model (ad-hoc fitting procedures), which seriously limited their performance, especially when the visibility of face is not in good condition. But existed landmark-free methods either treat head pose estimation as a sub-problem, or bring extra error during problem reduction. Therefore, in this paper, we present our efficient hybrid coarse-fine classification to deal with issues above. First of all, we extend previous work with stricter fine classification by increasing class number. Then, we introduce our hybrid coarse-fine classification scheme into the network. Integrate regression is adopted to get the final prediction. Our proposed approach to head pose estimation is evaluated on three challenging benchmarks, we achieve the state-of-the-art on AFLW2000 and BIWI, and our approach closes the gap with state-of-the-art on AFLW.

**Index Terms**— coarse-fine classification, head pose estimation, 3D facial understanding, deep learning, convolutional neural network.

## 1. INTRODUCTION

Facial expression recognition is one of the most successful applications of convolutional neural network in the past few years. Recently, more and more attentions have been posed on 3D facial understanding. Most of existed methods of 3D understanding require extracting 2D facial landmarks, and 3D pose of the head can be seen as a by-product, through establishing the corresponding relationship between 2D landmarks and a standardized 3D head model. While facial landmark detection has improved by large scale due to the adaptation of deep neural network, a two-step based head pose estimation still brings extra errors for precise head pose estimation. Specifically, the visibility of facial landmark limits the diversity and range of sample angle, since profile landmark detection which perplexes researchers, still has not been addressed successfully. Besides, the estimation of head pose is directly

related to the accuracy of 3D human head model which brings extra errors during ad-hoc fitting procedures. Third, it is too time-costing to label facial landmark. Thus, it is attractive to propose a landmark-free method to avoid these problems.



**Fig. 1.** Example pose estimation using our method. The blue axis points towards the front of the face, green pointing downward and red pointing to the side.

Recently, a fined-grained head pose estimation method[1] without landmarks has been proposed. It predict head pose Euler angle directly from image using a multi-loss network, where three angles are trained together and each angle loss is of two parts: a bin classification and regression component. Classification and regression components are connected through multi-loss training. However, they do not deal with the extra errors brought by coarse bin classification.

In our proposed method, we pose higher restriction for bin classification, in order to get better result of regression. Based on our observation, the classification converges much faster than regression, which weakens the usefulness of multi-loss training scheme. But a direct refined bin classification may counteract the benefit of problem reduction. Therefore, we introduce a hybrid coarse-fine classification framework, which proves to not only be helpful to refined bin classification but also improve the performance of prediction. Proposed network is shown in Figure 2. The main contributions for our work are summarized as below. Code will be released later.

- Use stricter fine bin classification to reduce the error brought by coarse bin classification.
- Propose our hybrid coarse-fine classification scheme to make better refined classification.

\*This work was done when Haofan Wang interned at Horizon Robotics.

- State-of-the-art performance for head pose estimation using CNN-based method on AFLW2000 and BIWI datasets, and close the gap with state-of-the-art on AFLW.

## 2. RELATED WORK

Head pose estimation has been widely studied and diverse traditional approaches have been proposed, including Appearance Template Models [2], Detector Arrays[3] and Manifold Embedding[4]. Until now, approaches to head pose estimation have adapted to deep neural network and been divided into two camps: landmark-based and landmark-free.

Landmark-based methods utilize facial landmarks to fit a standard 3D face. 3DDFA [5] directly fits a 3D face model to RGB image via convolutional neural networks, and aligns facial landmarks using a dense 3D model, 3D head pose is produced in the 3D fitting process. SolvePnP tool [6] also produces head pose in analogous way. However, this method usually use a mean 3D human face model which introduces intrinsic error during the fitting process.

Another recent work done by Aryaman et al. [7] achieves great performance on public datasets. They propose to use a higher level representation to regress the head pose while using deep learning architectures. They use the uncertainty maps in the form of 2D soft localization heatmap images over selected 5 facial landmarks, and pass them through a convolutional neural network as input channels to regress the head pose. However, this approach still cannot avoid problem of landmark invisibility even though they use coarse location, especially when considering that their method only involves with five landmarks which make their method very fragile to the invisible condition.

Landmark-free methods treat head pose estimation as a sub-problem of multi-task learning process. M. Patacchiola[8] proposes a shallow network to estimate head pose, and provide a detailed analysis on AFLW dataset. KEPLER[9] uses a modified GoogleNet and adopts multi-task learning to learn facial landmarks and head pose jointly. Hyperface[10] also follows multi-task learning framework, which detect faces and gender, predict facial landmarks and head pose at once. All-In-One[11] adds smile prediction and age estimation to the former method.

Chang et al.[12] regresses 3D head pose by a simple convolutional neural network. However, they focus on face alignment and do not explicitly evaluate their method on public datasets. Ruiz et al.[1] is another landmark-free work which performs well recently, they divide three branches to predict each angle jointly, each branch is combined by classification and integral regression. Lathuiliere et. al[13] proposed a CNN-based model with a Gaussian mixture of linear inverse regressions to regress head pose. Drouard et. al[14] further exploits to deal with issues in [13], including illumination, variability in face orientation and in appearance, etc. by combining the qualities of unsupervised manifold learning and in-

verse regressions.

Although recent state-of-the-art landmark-based method has better prediction given the ground truth of landmark, they suffer from landmark invisibility and the accuracy of landmark under real scene. Robust landmark-free method introduces extra error which limits its performance. In our work, we follow the landmark-free scheme and propose hybrid coarse-fine classification scheme which intends to solve the problem of extra error introduced by coarse classification in [1].

## 3. PROPOSED METHOD

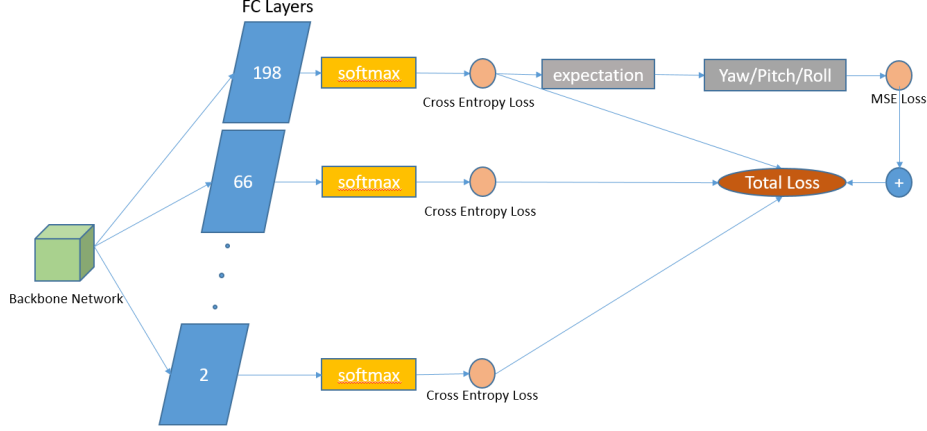
### 3.1. Hybrid coarse-fine classification

Multi-task learning (MTL) has led to success in many applications of machine learning. [15] demonstrates that MTL can be treat as implicit data augmentation, representation bias, regularization and etc. However, hybrid classification scheme for a single regression problem still does not receive enough attention as far as we know.

Although [1] contribute great work on head pose estimation using landmark-free method, but their work still meet with some issues. First of all, they do coarse bin classification before integrate regression. Bin classification relaxes a strict regression problem into a coarse classification problem, as classification is easier to learn for neural network, on the other hand, it introduces extra error which limits the performance of precise prediction. One of ways to make up such gap is to introduce a new offset, but this need add one more regression branch to regress offset, which may counteract the usefulness of problem reduction.

Therefore, we introduce our hybrid coarse-fine classification scheme into network, architecture is shown in Figure 2. At first, we use more refined bin classification at the highest level through increasing the bin number, which can improve the regression accuracy to bits, but this operation may counteract the benefit of problem reduction as it add more restriction on the learning process. Thus, we propose our hybrid coarse-fine classification scheme to offset the influence of refined classification. We relax the problem multi times on different scales by adding several full connected layers behind the backbone network, in order to ensure precise prediction under different classification scale, as we take both coarse bin classification and relatively fine bin classification into account, each FC layer represents a different classification scale and compute its own cross-entropy loss later. In the integrate regression component, we only utilize result of the most refined bin classification(the top branch in Figure 2.) to compute expectation and regression loss. One regression loss and multiple classification losses are combined as a total loss. Each angle has such a combined loss and share the previous convolutional layers of the network.

Our proposed hybrid coarse-fine classification scheme can



**Fig. 2.** Hybrid Coarse-fine Classification for head pose estimation. The upper classification branch is fine classification and used for get final prediction through integrate regression. Other branches are coarse classification and used to assist the learning.

be easily added into former framework and bring performance up without much extra computing resources. The final loss for each angle is the following:

$$Loss = \alpha * MSE(y, y^*) + \sum_{i=1}^{num} \beta_i * H(y_i, y_i^*) \quad (1)$$

where  $H$  and  $MSE$  respectively designate the cross-entropy loss and mean squared error loss functions,  $num$  means the number of classification branch.

We experiment with different coefficients for the regression component and hybrid classification component, our results are presented in Table 4. and Table 5.

### 3.2. Integrate regression

Xiao et al.[16] introduces integrate regression into human pose estimation to cope with non-differentiable post-processing and quantization error. Their work shows that a simple integral operation relates and unifies the heat map representation and joint regression. Ruiz et al.[1] utilizes integrate regression in head pose estimation.

$$J_k = \arg \max_p H_k(p) \quad (2)$$

$$J_k = \int_{p \in \Omega} p * H_k(p) \quad (3)$$

$\Omega$  is domain and  $p$  is the index of bin. This scheme treats a direct regression problem as two steps process, a multi-class classification followed by integrate regression, by modifying the “taking-maximum” operation of classification as shown in (1) to “taking-expectation” as shown in (2), and a fine-grained predictions is obtained by computing the expectation of each output probability for the binned output.

We follow this setting in our network and use the same backbone network as [1] in order to fairly compare. Intuitively, such scheme can be seen as a way of problem reduction, as bin classification is a coarse annotation rather than precise label, classification and the output are connected through multi-loss learning which makes the classification also sensitive the output. Another explanation is that bin classification uses the very stable softmax layer and cross-entropy loss, thus the network learns to predict the neighborhood of the pose in a robust fashion.

## 4. EXPERIMENTS

### 4.1. Datasets for Pose Estimation

We demonstrate that datasets under real scene with precise head pose annotations, numerous variation on pose scale and lighting condition, is essential to make progress in this filed. Three benchmarks are used in our experiments.

300W-LP [5]: is a synthetically expanded dataset, and a collection of popular in-the-wild 2D landmark datasets which have been re-annotated. It contains 61,225 samples across large poses, which is further expanded to 122,450 samples with flipping.

AFLW2000 [17]: contains the first 2000 identities of the in-the-wild AFLW dataset, all of them have been re-annotated with 68 3D landmarks.

AFLW [17]: contains 21,080 in-the-wild faces with large-pose variations (yaw from  $-90^\circ$  to  $90^\circ$ ).

BIWI [18]: is captured in well controlled laboratory environment by record RGB-D video of different people across different head pose range using a Kinect v2 device and has better pose annotations. It contains about 15, 000 images with  $\pm 75^\circ$  for yaw,  $\pm 60^\circ$  for pitch and  $\pm 50^\circ$  for roll.

Method	Yaw	Pitch	Roll	MAE
3DDFA[5]	5.400	8.530	8.250	7.393
Ruiz et al.[1]	6.470	6.559	5.436	6.155
<b>Ours</b>	<b>4.820</b>	<b>6.227</b>	<b>5.137</b>	<b>5.395</b>

**Table 1.** Mean average error of Euler angles across different methods on the AFLW2000 dataset.

#### 4.2. Pose Estimation on the AFLW2000

Same backbone network as [1] is adopted. Network was trained for 25 epochs on 300L-WP using Adam optimization [19] with a learning rate of  $10^{-6}$  and  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\varepsilon = 10^{-8}$ . We normalize the data before training by using the ImageNet mean and standard deviation for each color channel. Our method bins angles in the  $\pm 99^\circ$  range we discard images with angles outside of this range. Results can be seen in Table 1.

Method	Yaw	Pitch	Roll	MAE
Liu et al.[20]	6.0	6.1	5.7	5.94
Ruiz et al.[1]	4.810	6.606	3.269	4.895
Drounard[14]	4.24	5.43	4.13	4.60
DMLIR[13]	<b>3.12</b>	4.68	3.07	3.62
MLP + Location[7]	3.64	4.42	3.19	3.75
CNN + Heatmap[7]	3.46	3.49	<b>2.74</b>	3.23
<b>Ours</b>	3.4273	<b>2.6437</b>	2.9811	<b>3.0174</b>

**Table 2.** Mean average error of Euler angles across different methods on the BIWI dataset with 8-fold cross-validation.

Method	Yaw	Pitch	Roll	MAE
Patacchiola et al.[8]	11.04	7.15	4.40	7.530
KEPLER[9]	6.45	5.85	8.75	7.017
Ruiz et al.[1]	6.26	5.89	3.82	5.324
MLP + Location[7]	6.02	5.84	3.56	5.14
<b>Ours</b>	6.18	5.38	3.71	5.090
CNN + Heatmap[7]	5.22	4.43	2.53	4.06

**Table 3.** Mean average error of Euler angles across different methods on the AFLW dataset.

#### 4.3. Pose Estimation on the AFLW and BIWI Datasets

We also test our method on AFLW and BIWI dataset with same parameter setting as 4.2. Results can be seen in Table 2. and Table 3. Our method achieves state-of-the-art on BIWI, MAE of our method is lower the base network [1] by large scale, and also achieves better performance than the recent landmark-based CNN + Heatmap [7] method. We demonstrate that our method perform well on BIWI as it is captured

in controlled environment, and has better ground truth annotations, but this verifies the usefulness of our hybrid coarse-fine classification scheme when the annotation is precise.

We also surpass all landmark-free methods, and achieve competing performance over all methods on AFLW, following testing protocol in [9] (i.e. selecting 1000 images from testing and remaining for training.)

#### 4.4. AFLW2000 Multi-Classification Ablation

$\alpha$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	MAE
2	1	0	0	0	0	5.7062
2	3	1	1	1	1	5.6270
2	5	3	1	1	1	5.6898
2	7	5	3	1	1	<b>5.3953</b>
2	9	7	5	3	1	5.5149

**Table 4.** Ablation analysis: MAE across different classification loss weights on the AFLW2000 dataset.

$\alpha$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	MAE
0.1	7	5	3	1	1	5.4834
1	7	5	3	1	1	5.6160
2	7	5	3	1	1	<b>5.3953</b>
4	7	5	3	1	1	5.6255

**Table 5.** Ablation analysis: MAE across different regression loss weights on the AFLW2000 dataset.

In this part, we present an ablation study of the hybrid coarse-fine classification. We train ResNet50 using different coefficient setting. Results can be seen in Table 4. and Table 5. We observe the best results on the AFLW2000 dataset when the coefficient is 2,7,5,3,1,1 by order.  $\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$  correspond to the weight of regression, 198 classes, 66 classes, 18 classes, 6 classes and 2 classes. We demonstrate increased accuracy when weighting each loss differently, and the magnitude of regression weight has little effect on the overall performance.

## 5. CONCLUSION

We present a hybrid classification scheme for precise head pose estimation without facial landmarks. Our proposed method achieves state-of-the-art on BIWI and AFLW2000 dataset, and also achieves promising performance by comparing with other state-of-the-art methods on AFLW dataset. The hybrid coarse-fine classification framework is proved to be good at head pose estimating. We believe that our accurate head pose method without facial landmarks can be easily combined into other problem as an extra branch to provide 3D angle information, and we will work on it later.

## 6. REFERENCES

- [1] Nataniel Ruiz, Eunji Chong, and James M Rehg, “Fine-grained head pose estimation without keypoints,” *arXiv preprint arXiv:1710.00925*, 2017.
- [2] Jeffrey Huang, Xuhui Shao, and Harry Wechsler, “Face pose discrimination using support vector machines (svm),” in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*. IEEE, 1998, vol. 1, pp. 154–156.
- [3] Edgar Osuna, Robert Freund, and Federico Girosit, “Training support vector machines: an application to face detection,” in *Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on*. IEEE, 1997, pp. 130–136.
- [4] Vineeth Nallure Balasubramanian, Jieping Ye, and Sethuraman Panchanathan, “Biased manifold embedding: A framework for person-independent head pose estimation,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–7.
- [5] 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*, 2016, pp. 146–155.
- [6] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng, “Complete solution classification for the perspective-three-point problem,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.
- [7] Aryaman Gupta, Kalpit Thakkar, Vineet Gandhi, and PJ Narayanan, “Nose, eyes and ears: Head pose estimation by locating facial keypoints,” *arXiv preprint arXiv:1812.00739*, 2018.
- [8] Massimiliano Patacchiola and Angelo Cangelosi, “Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods,” *Pattern Recognition*, vol. 71, pp. 132–143, 2017.
- [9] Amit Kumar, Azadeh Alavi, and Rama Chellappa, “Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors,” in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 258–265.
- [10] R Ranjan, VM Patel, and R Chellappa, “Hyperfacer: A deep multi-task learning framework for face detection,” *landmark localization, pose estimation, and gender recognition*, 2016.
- [11] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa, “An all-in-one convolutional neural network for face analysis,” in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 17–24.
- [12] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni, “Faceposenet: Making a case for landmark-free face alignment,” in *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1599–1608.
- [13] Stéphane Lathuilière, Rémi Juge, Pablo Mesejo, Rafael Muñoz-Salinas, and Radu Horaud, “Deep mixture of linear inverse regressions applied to head-pose estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 3, p. 7.
- [14] Vincent Drouard, Radu Horaud, Antoine Deleforge, Sil-eye Ba, and Georgios Evangelidis, “Robust head-pose estimation based on partially-latent mixture of linear regressions,” *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1428–1440, 2017.
- [15] Sebastian Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv preprint arXiv:1706.05098*, 2017.
- [16] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, “Integral human pose regression,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 529–545.
- [17] 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 localization,” in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.
- [18] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool, “Real time head pose estimation from consumer depth cameras,” in *Joint Pattern Recognition Symposium*. Springer, 2011, pp. 101–110.
- [19] D Kinga and J Ba Adam, “A method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, 2015, vol. 5.
- [20] Xiabing Liu, Wei Liang, Yumeng Wang, Shuyang Li, and Mingtao Pei, “3d head pose estimation with convolutional neural network trained on synthetic images,” in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1289–1293.