Беженарь Александр Васильевич,
Тюменский государственный университет,
Институт математики и компьютерных наук,
студент МОиАИС-174,
Sbezhenar43@gmail.com

Сизова Людмила Владимировна,
Тюменский государственный университет,
Институт социально-гуманитарных наук ,
Кафедра иностранных языков и межкультурной
профессиональной коммуникации,
старший преподаватель,
l.v.sizova@utmn.ru

НОС, ГЛАЗА И УШИ: ОЦЕНКА ПОЗЫ ГОЛОВЫ ПО ОБНАРУЖЕНИЮ ХАРАКТЕРНЫХ ТОЧЕК ЛИЦА

Bezhenar Aleksandr Vasilevich,
the University of Tyumen,
Institute of Mathematics and Computer Science,
student of M&ISA-174,
Sbezhenar43@gmail.com

Lyudmila V. Sizova,
the University of Tyumen,
Institute of Humanities and Social Studies,
Department of Foreign Languages and
Intercultural Professional Communication,
senior lecturer,

l.v.sizova@utmn.ru

NOSE, EYES AND EARS: HEAD POSE ESTIMATION BY LOCATING FACIAL KEYPOINTS

АННОТАЦИЯ: Монокулярная оценка позы головы требует изучения модели, которая вычисляет внутренние углы Эйлера для позы (рысканье, тангаж, крен) из входного изображения человеческого лица. Обозначать углы позы головы в реальной ситуации для изображений на практике сложно и требует специальных процедур подгонки. Это подчеркивает необходимость подходов, которые могут тренироваться на данных, полученных в контролируемой среде, и обобщать изображения в естественной среде (с различным внешним видом и освещением лица). Авторы статьи предлагают использовать представление более высокого уровня, чтобы регрессировать позу головы при использовании архитектур глубокого обучения. Более конкретно, они используем карты неопределенности в виде двухмерных изображений тепловой карты мягкой локализации для пяти ключевых точек лица, а именно левого уха, правого уха, левого глаза, правого глаза и носа, и пропускаем их через сверточную нейронную сеть для регрессии позы головы. Результаты оценки позы головы показываются на двух сложных контрольных показателях BIWI и AFLW.

Ключевые слова: анализ изображения, оценка позы, невербальная коммуникация.

ABSTRACT: Monocular head pose estimation requires learning a model that computes the intrinsic Euler angles for pose (yaw, pitch, roll) from an input image of human face. Annotating ground truth head pose angles for images in the wild is difficult and requires ad-hoc fitting procedures. This highlights the need for approaches which can train on data captured in controlled environment and generalize on the images in the wild (with varying appearance and illumination of the face). The authors of the article propose to use a higher level representation to regress the

head pose while using deep learning architectures. More specifically, they use the uncertainty maps in the form of 2D soft localization heatmap images over five facial key points, namely left ear, right ear, left eye, right eye and nose, and pass them through a convolutional neural network to regress the head-pose. The authors show head pose estimation results on two challenging benchmarks BIWI and AFLW.

***Key words**:* Image analysis, pose estimation, non-verbal communication.

Introduction: The ability of humans to comprehend non-verbal communication by effortlessly estimating the orientation and movements of human head is fascinating. In order to humanize machines by bringing them closer to human-like perception and understanding, accurately estimating the human head orientation using visual imagery presents an important challenge. Head pose relates to the visual attention and interest of a person, which is crucial for many applications in computer vision. Estimating head pose has been actively pursued in problems like social event analysis [1], Human Computer Interaction (HCI) [2], driver assistance systems [3] etc., which are an important part of present day technologies.

With the availability of well an-notated datasets captured using Kinect sensors such as BIWI [4], monocular head pose estimation with 3-DOF has seen good improvements in recent years. The state-of-the-art method relies on end-to-end convolutional regression networks [5], which takes RGB images as input and learns the parameters of an inverse regression network using a Mean Squared Error (MSE) loss. As BIWI [4] is captured in a controlled environment for accurate ground truth annotation which is dependent on precise 3D reconstruction of face, methods using RGB input directly for head pose estimation on BIWI [4] fail to generalize on images in the wild. On the other hand, datasets like AFLW [6] only provide coarse approximation of ground truth angles as annotation of ground truth on images in the wild is challenging. Hence, an important property for head pose estimation algorithms is generalization on face images in the wild when trained on precisely annotated datasets like BIWI [4].

While computer vision based pose estimation approaches have focused predominantly on appearance-based solutions that compute human pose directly from digital images, there have been methods based on psychophysical experiments. These consider the human perception of head pose to rely on cues such as deviation of nose angle and the deviation of the head from bilateral symmetry [7]. Since it is easier to annotate 2D keypoints directly on images, huge labelled datasets are now available [8] and have led to development of powerful methods [9] for localizing keypoints like nose, eyes and ears. We hypothesize that we can learn a head pose estimation model using only five facial keypoint locations. Such a model implicates an abstraction over the appearance and illumination dependent image data which is a hindrance for generalization capability of head pose estimation methods. The abstraction limits the dependencies of the model to scale and configuration of a few keypoint locations.

Our first baseline approach takes as input the keypoint locations and directly predicts the head-pose using a Multi-Layer Perceptron (MLP). However, we notice that the facial keypoint locations have inherent uncertainty in their estimation. Hence we propose a second framework, which first computes the un-certainty maps for the five points in the form of heatmap images capturing their soft localization (in other words, the probability distribution of all possible locations of that keypoint). The five images are then stacked together and provided as input to a Convolutional Neural Network (CNN) for estimation of head pose angles.

Our baseline approach is to employ a Multi-Layer Perceptron (MLP) which regresses the 3D head-pose directly using the predicted locations of the five keypoints (detected using [9]). Each of the keypoint is parameterized by its 2D location and prediction likelihood, resulting in an input vector of 15 dimensions, which is used to regress a 3D vector representing the yaw, pitch and roll. Undetected keypoints are represented by a vector of zeroes.

MLP-based method is based on the assumption that the locations of five facial keypoints estimated from the face image are accurate. However, in practice there is inherent uncertainty in predicting the locations of keypoints such as eyes, ear and

nose, using an optimization based approach [9]. One possible way to account for this uncertainty in localization is to treat the image locations of the facial keypoints as latent variables. From a representation perspective, uncertainty maps (heatmap images) can be used to depict latent variables, which capture the soft localization of 2D

Fig. 1. Example of a face image, detected keypoints and respective heatmaps of each keypoint computed using [9].

Keypoint locations. An image-based representation of the facial keypoint locations facilitates the use of CNN-based approaches for learning the head pose. Uncertainty maps over locations of keypoints (or joints) in human body or an object skeleton, present in an image, have been successfully used in previous literature where the exact locations of the keypoints were noisy or unknown. Zhou [10] use heatmap images of 2D joint locations to infer 3D human pose using an Expectation Maximization framework. Wu [11] use heatmaps of 2D skeleton keypoints of an object as an intermediate representation to recover 3D structure of an object and bridge the gap between synthetic and real data. Interestingly, both these works [10, 11] use heatmaps over 2D spatial locations to infer 3D structure/pose. Deriving motivation from these efforts, we propose an algorithm which takes 2D uncertainty maps over the facial keypoints as input and regresses the 3D head pose.

Unlike previous efforts [10, 11] that use heatmaps as an intermediate representation and do not have ground truth data, we have ground truth pose angles available. This allows us to directly train a convolutional regression network using ground truth supervision for head pose estimation. Specifically, we use OpenPose [9] to compute the uncertainty maps for the five facial keypoint locations as illustrated in Figure 1. Each heatmap image is considered as a separate channel and the channels are stacked together, which generates a 5-channel feature map. This feature map is used as an input to the CNN, to learn a head pose estimation model. The final

layer gives the values of three pose angles obtained as a result of the convolutional regression. We use a MSE loss to train the convolutional regression network, which can be written as follows:

$$L_{mse} = \frac{1}{3}\sum_{i=1}^{3}(\theta_i - \widehat{\theta_i})^2$$

where θ_i is the vector consisting of the predicted values for intrinsic Euler angles and $\widehat{\theta_i}$ is the vector consisting of the values of ground truth angles.

MLP-based Model Our network consists two hidden layers of size 30 neurons each. We set learning rate of 0.00001 and train for 500 epochs using Adam optimizer with a weight decay of 0.0001 and batch size 64.

CNN-based Model We use a CNN architecture with 3 convolution layers and 2 fully connected layers (we have use same architecture used in Liu[12] but with 5 input channels). Training is run for 1200 epochs with Adam optimizer and set learning rate of 0.00001. We set the batch size to 32.

All the experiments are run on a single Nvidia GTX 1080Ti GPU.

We use two benchmark datasets to measure the performance of our models and test them. BIWI Kinect Headpose Dataset [4] contains over 15K samples spread over 24 sequences, captured in a controlled environment. The range of head pose angles in the dataset vary from ±75◦ for yaw, ±60◦ for pitch and ±50◦ for roll. AFLW [6] Annotated Facial Landmarks in the Wild (AFLW) provides a large-scale collection of annotated face images gathered from the web, exhibiting a large variety in appearance (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions. In total about 25K faces are annotated with up to 21 landmarks per image.

Results on BIWI dataset: As BIWI is captured in controlled conditions and has better ground truth annotations, better performance is achieved on this dataset. The motivation for designing our frameworks is to train a model on a dataset like BIWI and use it to generalize to face images in the wild. In order to demonstrate the ability of our frameworks, we predict the head pose on unseen images taken from

the web. Our results show the presence of a perceptually better sense of pose than a model learned directly on the RGB images. Quantitative results for the dataset in terms of Mean Absolute Error (MAE) from ground truth annotations are while the CNN based approach surpasses the state of the art.

Results on AFLW dataset: Given the large variations in AFLW dataset, most of the previous methods compute results for head pose estimation on this dataset by constraining the range of angles, using a subsampled set of images or creating a very small test set [13, 14]. We do not assume any such constraints and show

Method	Yaw	Pitch	Roll	MAE
Liu [23]	6.0	6.1	5.7	5.94
Ruiz et al. [18]	4.810	6.606	3.269	4.895
Drouard [19]	4.24	5.43	4.13	4.6
DMLIR [8]	**3.12**	4.68	3.07	3.62
MLP with location (Ours)	3.64	4.42	3.19	3.75
CNN + Heatmaps (Ours)	3.46	**3.49**	**2.74**	**3.23**

Table 1. Results on BIWI with 8-fold cross-validation (21 randomly selected videos for training and the remaining 3 videos for test such that no person appears both in training and test sets)

the results using a standard five-fold validation process on the entire dataset, where the samples are randomly divided into train and test sets with 80% samples ending up in training set (Table 2).

Method	Yaw	Pitch	Roll	MAE
View manifolds [24]	-	-	-	17.52
Random Forests [25]	-	-	-	12.26
Pata. and Cang.∗ [17]	11.04	7.15	4.4	7.53
MLP + Locations (Ours)	9.56	6.64	4.68	6.96
CNN + Heatmaps (Ours)	**6.19**	**5.58**	**3.76**	**5.18**

Table 2. Results on AFLW dataset with 5-fold cross validation. ∗ : Constrains the angles to a certain range.

We also perform experiment following testing protocol in [15] (i.e. selecting 1000 images from testing and remaining for training) and present the results in Table 3.

Method	Yaw	Pitch	Roll	MAE
Kepler [26]	6.45	7.05	5.85	6.45
Ruiz et al. [18]	6.26	5.89	3.82	5.324
MLP + Locations (Ours)	6.02	5.84	3.56	5.14
CNN + Heatmaps (Ours)	**5.22**	**4.43**	**2.53**	**4.06**

Table 3. Results on AFLW using testing protocol in [26].

The numbers of other methods in both tables are reported directly from the associated papers (aligned with corresponding protocol).

The results clearly show that our CNN-based framework achieves the lowest MAE, significantly improving on the previous state-of-the-art on both the protocols. Interestingly, the MLP based approach also gives competitive performance as compared to previous work. We believe that the exact locations of the facial keypoints, as used in case of MLP, makes it prone to overfitting while the heatmaps act as a regularizer in that sense, giving an edge to CNN based framework. Overall, the experiments provide a strong empirical evidence towards the hypothesis pursued in this paper.

Conclusion: In this paper, we present a hypothesis that using an intermediate representation such as locations of five facial keypoints instead of face images can help achieve better pose estimation and generalization performance. We propose two frameworks (a baseline approach employing MLP and a CNN over uncertainty maps) to support our claim. Although, minimal the MLP based approach gives competitive performance and we believe that it will improve with improvement in localization of keypoints. Owing to presence of noise in localization estimates, our CNN-based approach uses it as an advantage by representing the uncertainty as heatmaps and regressing the head pose with the heatmaps as input. The CNN-based framework surpasses state-of-the-art for head pose estimation on two challenging benchmarks BIWI [4] and AFLW [6].

REFERENCES

1. Varadarajan J., Subramanian R., Bulo S.R., Ahuja N., Lanz O., Ricci E. (2018) Joint estimation of human pose and conversational groups from social scenes. *IJCV*. pp. 410-429.

2. Wang K., Zhao R., Ji Q. (2018) Human computer interaction with head pose, eye gaze and body gestures. *FG*. p. 789.

3. Schwarz A., Haurilet M., Martinez M., Stiefelhagen R. (2017) Driveaheada large-scale driver head pose dataset. *CVPRW*. pp. 1-10.

4. Fanelli G., Weise T., Gall J., Gool L.V. (2011) Real time head pose estimation from consumer depth cameras. *DAGM*. pp. 617-624.

5. Lathuilire S., Juge R., Mesejo P., Munoz-Salinas R., Horaud R. (2017) Deep mixture of linear inverse regressions applied to head-pose estimation. *CVPR*. pp. 4817-4825.

6. Roth P.M., Koestinger M., Wohlhart P., Bischof H. (2011) Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies. Barcelona. DOI: 10.1109/ICCVW.2011.6130513 .

7. Wilson H.R. et al. (2000) Perception of head orientation. Vision Research, pp. 459-472.

8. Andriluka M., Pishchulin L., Gehler P., Bernt Schiele. (2014) 2d human pose estimation: New benchmark and state of the art analysis. *CVPR*. pp. 3686-3693.

9. Cao Z., Simon T., Wei S.E., Sheikh Y. (2017) Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*. DOI: 10.1109/CVPR.2017.143.

10. Zhou X., Zhu M., Leonardos S., Derpanis K.G., Daniilidis K. (2016) Sparseness meets deepness: 3d human pose estimation from monocular video. *CVPR*. pp. 4966-4975.

11. Wu J., Xue T., Lim J.J., Tian Y., Tenenbaum J.B., Torralba A., Freeman W.T. (2016) Single image 3d interpreter network. *ECCV*. pp.365-382.

12. Liu X., Liang W, Wang Y., Li S., Pei M. (2016) 3d head pose estimation with convolutional neural network trained on synthetic images. *ICIP*. DOI:10.1109/ICIP.2016.7532566.

13. Ruiz N., Chong E., Rehg J.M. (2017) Finegrained head pose estimation without keypoints. *CoRR*. pp. 2074-2083.

14. Patacchiola M., Cangelosi A. (2017) Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition7*. DOI: 10.1016/j.patcog.2017.06.009.

15. Kumar et al. (2017) Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. *FG*. DOI:10.1109/FG.2017.149