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

Aryaman Gupta, Kalpit Thakkar, Vineet Gandhi and P J Narayanan

Centre for Visual Information Technology, KCIS, IIIT Hyderabad

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 (which provides only coarse and approximate annotations). 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). Most present day deep learning approaches which learn a regression function directly on the input images fail to do so. To this end, we propose to use a higher level representation to regress the head pose while using deep learning architectures. More specifically, we use the uncertainty maps in the form of 2D soft localization heatmap images over five facial keypoints, namely left ear, right ear, left eye, right eye and nose, and pass them through an convolutional neural network to regress the head-pose. We show head pose estimation results on two challenging benchmarks BIWI and AFLW and our approach surpasses the state of the art on both the datasets.
Index Terms— Image analysis, Pose estimation

1. 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.

Formally, head pose estimation entails computing the 3D orientation of head with respect to the camera pose using digital images. Initial approaches estimated only one or two angles for head pose while assuming other angles are fixed or fixed discrete values for head pose angles to be estimated [5, 6]. However, head pose estimation with *three* degrees of freedom, viz. (yaw, pitch and roll), is more useful than discrete head pose and recent methods have been aimed at estimating the three head pose angles. With the availability of well annotated datasets captured using Kinect sensors such as BIWI [7], 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 [8], which takes RGB images as input and learns the parameters of an inverse regression network using a Mean Squared Error (MSE) loss. As BIWI [7] 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 [7] fail to generalize on images in the wild (as illustrated in Figure 1). On the other hand, datasets like AFLW [9] 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 [7].

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 [10]. Since it is easier to annotate 2D keypoints directly on images, huge labelled datasets are now available [11] and have lead to development of powerful methods [12] 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 uncertainty 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

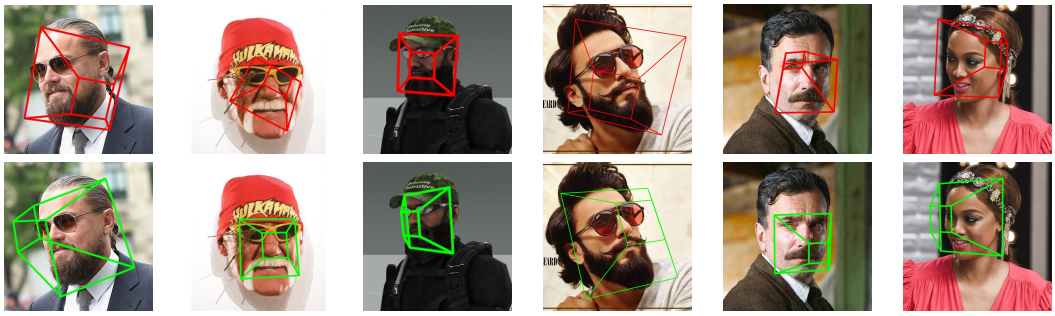

Fig. 1. Estimation of head pose using three different models (all trained on BIWI), on unseen images taken from the web. **Top row**: Results for CNN-based model [4] which takes RGB images as input, **Bottom row**: Results for our CNN-based framework which takes heatmaps of five facial keypoints locations as input.

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. We show that our baseline approach achieves competitive performance, while CNN-based framework surpasses state-of-the-art. The contributions of this paper are as follows:

- A hypothesis on learning a model for head pose estimation which relies only on five facial keypoint locations and abstracts out the dependency on appearance of the subject.
- A baseline approach that uses the exact keypoint locations (sampled from their distribution) and employs a MLP for regression of pose angles.
- A CNN-based framework which uses the probability distribution of keypoint locations in the form of heatmap images, as input to regress the head pose.
- State-of-the-art performance for head pose estimation using the CNN-based framework on the BIWI [7] and AFLW [9] datasets.

2. RELATED WORK

Previous approaches to head pose estimation can be classified into two categories: RGB and RGBD based (2D vs 3D input). We limit our discussion to RGB input only. Earlier methods for head pose estimation used appearance templates that use a set of exemplars to find the pose of an input image, by finding the closest exemplar [5]. The assumption that similarity in image space equates similarity in pose is the major drawback of such methods. Extending appearance templates, several methods using multiple pose detectors (each corresponding to one discrete pose) have been proposed [6]. However, detector-based methods require several detectors and non-face samples (negative samples) for successful training, which is burdensome. Manifold embedding methods were later introduced, which project an input sample to a lower dimension using an embedding function and regress pose in the embedding space. Techniques like PCA [13], Isomap [14] and several combinations [15] of dimensionality reduction approaches are used for head pose estimation. Learning useful low-dimensional representations needs proper training data having balanced samples.

With the transition to deep learning based methods, several former drawbacks have been mitigated. One of the earliest efforts in this area was by Osadchy et. al [16]. They extract CNN features from images and regress pose using them. Patacchiola and Cangelosi [17] test the effect of dropout and adaptive gradient-based methods combined with CNNs for head pose estimation, where they propose to use adaptive gradients in conjunction with a CNN. On the other hand, Ruiz et. al [18] propose a CNN with 3 separate branches, each with combined classification and regression for the respective head pose angle. Both these methods aim to improve performance of head pose estimation in the wild. Lathuiliére et. al [8] proposed a CNN-based model with a Gaussian mixture of linear inverse regressions. They use an Imagenet-pretrained CNN to learn face features and train a pose regressor on them. An extension of this approach by Drouard et. al [19] proposes to cope with changing illumination conditions, variability in face orientation and in appearance, etc. by combining the qualities of unsupervised manifold learning and inverse regressions. However, as the CNN-based methods estimate the pose angles directly from RGB images, it makes them prone to poor generalization on account of illumination ass well as appearance changes. Geometric models regress the pose using facial features such as keypoints, nose angle, etc. and have been proposed in previous literature [20]. Similar in spirit, we propose to use a higher-level feature to drive the pose regression, viz. the heatmaps of five facial keypoints extracted from face images (or exact 2D locations) using a keypoint localization routine [12]. The performance of our models prove our hypothesis of facilitating abstraction over illumination and appearance dependent image data by achieving state-of-the-art results for head pose estimation and demonstrating good generalization capability.

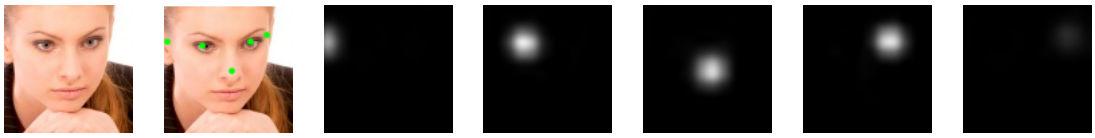

Fig. 2. Example of a face image, detected keypoints and respective heatmaps of each keypoint computed using [12].

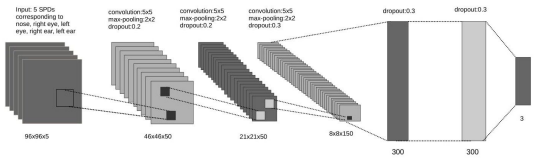

Fig. 3. The architecture consists of 3 convolutional layers (conv1, conv2, conv3) followed by two fully connected layers (fc1, fc2). The input has 5 channels: one each for the nose, left eye, right eye, left ear and right ear (heatmap images for these keypoints). The network outputs the estimated values of the three intrinsic Euler angles (yaw, pitch, roll).

3. HEAD POSE ESTIMATION VIA KEYPOINT LOCALIZATION

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 [12]). 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 [12]. 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 keypoint locations (Figure 2 illustrates an image and corresponding uncertainty maps for the five different facial keypoints used in our work). 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 [21] use heatmap images of 2D joint locations to infer 3D human pose using an Expectation Maximization framework. Wu [22] 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 [21, 22] 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 [21, 22] 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 [12] to compute the uncertainty maps for the five facial keypoint locations as illustrated in Figure 2. 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, the architecture of which is shown in Figure 3, 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:

$$\mathbf{L}_{\text{mse}} = \frac{1}{3}\sum_{i=1}^{3}\left(\mathbf{\Theta}_i - \hat{\mathbf{\Theta}}_i\right)^2 \tag{1}$$

where, $\mathbf{\Theta}_i$ is the vector consisting of the predicted values for intrinsic Euler angles and $\hat{\mathbf{\Theta}}_i$ is the vector consisting of the values of ground truth angles.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup and Datasets

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 used same architecture used in Liu[23] 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

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)

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.

[7] contains over 15K samples spread over 24 sequences, captured in a controlled environment. The range of head pose angles in the dataset vary from $\pm 75^\circ$ for yaw, $\pm 60^\circ$ for pitch and $\pm 50^\circ$ for roll. **AFLW** [9] 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.

4.2. Results

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 (as illustrated in Figure 1). 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 shown in Table 1 which shows that the MLP model achieves competitive performance, 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 [18, 17]. We do not assume any such constraints and show 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. We also perform experiment following testing protocol in [26] (i.e. selecting 1000 images from testing and remaining for training) and present the results in Table 3. The numbers of other methods in both tables are reported directly from the associated papers (aligned with corresponding protocol).

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 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.

5. CONCLUSIONS

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 [7] and AFLW [9].

Acknowledgements

This work was supported in part by Early Career Research Award, ECR/2017/001242, from Science and Engineering Research Board (SERB), Department of Science Technology, Government of India

6. REFERENCES

[1] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz, and Elisa Ricci, "Joint estimation of human pose and conversational groups from social scenes," *IJCV*, 2018.

[2] K. Wang, R. Zhao, and Q. Ji, "Human computer interaction with head pose, eye gaze and body gestures," in *FG*, 2018.

[3] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen, "Driveaheada large-scale driver head pose dataset," in *CVPRW*, 2017.

[4] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson, "Face alignment assisted by head pose estimation," in *BMVC*, 2015.

[5] J. Huang, X. Shao, and H. Wechsler, "Face pose discrimination using support vector machines (svm)," in *ICPR*, 1998.

[6] Jeffrey Ng Sing Kwong and Shaogang Gong, "Composite support vector machines for detection of faces across views and pose estimation," *Image Vision Computing*, 2002.

[7] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *DAGM*, 2011.

[8] S. Lathuilire, R. Juge, P. Mesejo, R. Muoz-Salinas, and R. Horaud, "Deep mixture of linear inverse regressions applied to head-pose estimation," in *CVPR*, 2017.

[9] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[10] Hugh R Wilson, Frances Wilkinson, Li-Ming Lin, and Maja Castillo, "Perception of head orientation," *Vision Research*, 2000.

[11] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[13] Junwen Wu and Mohan M. Trivedi, "A two-stage head pose estimation framework and evaluation," *Pattern Recogn.*, 2008.

[14] Nan Hu, Weimin Huang, and S. Ranganath, "Head pose estimation by non-linear embedding and mapping," in *ICIP*, 2005.

[15] Shuicheng Yan, Zhenqiu Zhang, Yun Fu, Yuxiao Hu, Jilin Tu, and Thomas Huang, "Learning a person-independent representation for precise 3d pose estimation," in *Multimodal Technologies for Perception of Humans*, 2008.

[16] Margarita Osadchy, Yann Le Cun, and Matthew L. Miller, "Synergistic face detection and pose estimation with energy-based models," *JMLR*, 2007.

[17] Massimiliano Patacchiola and Angelo Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognition*, 2017.

[18] Nataniel Ruiz, Eunji Chong, and James M. Rehg, "Fine-grained head pose estimation without keypoints," *CoRR*, 2017.

[19] Vincent Drouard, Radu Horaud, Antoine Deleforge, Sileye Ba, and Georgios Evangelidis, "Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions," *TIP*, 2017.

[20] Jian-Gang Wang and Eric Sung, "Em enhancement of 3d head pose estimated by point at infinity," *Image Vision Comput.*, pp. 1864–1874, 2007.

[21] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis, "Sparseness meets deepness: 3d human pose estimation from monocular video," in *CVPR*, 2016.

[22] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman, "Single image 3d interpreter network," in *ECCV*, 2016.

[23] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3d head pose estimation with convolutional neural network trained on synthetic images," in *ICIP*, 2016.

[24] K. Sundararajan and D. L. Woodard, "Head pose estimation in the wild using approximate view manifolds," in *CVPRW*, June 2015, pp. 50–58.

[25] Roberto Valle, José Miguel Buenaposada, Antonio Valdés, and Luis Baumela, "Head-pose estimation in-the-wild usingarandom forest," Cham, 2016, pp. 24–33, Springer International Publishing.

[26] Kumar et al, "Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors," in *FG*, May 2017.