NosePose: a competitive, landmark-free methodology for head pose estimation in the wild

Flávio H. B. Zavan, Antonio C. P. Nascimento, Olga R. P. Bellon and Luciano Silva
IMAGO Research Group - Universidade Federal do Paraná
{flavio,antonio.paes,olga,luciano}@ufpr.br

***Abstract*—We perform head pose estimation solely based on the nose region as input, extracted from 2D images in unconstrained environments. Such information is useful for many face analysis applications, such as recognition, reconstruction, alignment, tracking and expression recognition. Using the nose region has advantages over using the whole face; not only it is less likely to be occluded by acesssories, it is also visible and proved to be highly discriminant in all poses from profile to frontal. To this end, we propose and compare two different approaches, based on Support Vector Machines (SVM-NosePose) and on Convolutional Neural Networks (CNN-NosePose) such that no landmarks are needed to perform pose estimation, favoring success in extreme pose and environment where landmark detection is non-trivial. Our NosePose methodology was applied to four publicly available uncontrolled image datasets (McGillFaces, AFW, PaSC and IJB-A). Results show that both SVM-NosePose and CNN-NosePose approaches are competitive, through thoughtful and comprehensive experiments, when compared against state-of-the-art works on head pose estimation.**

***Keywords*-Head pose estimation; Nose pose estimation; Face image analysis; Support vector machines; Convolutional neural network**

I. Introduction

The head pose estimation problem can be defined as determining at least one of the three parameters that configures the face relative to its three degrees of freedom, yaw, pitch and roll (Figure 1) and the camera [1]. The growing interest in head pose estimation is mainly due to the advantages it brings to facial analysis tasks. Estimating the head pose can lead to higher accuracy rates in other computer vision problems, such as gaze estimation [2], face quality assessment [3], face frontalization [4], face recognition [5], facial landmark detection [6], 3D face reconstruction [7] and facial expression recognition [8].

Most of the previous works use 2D information from the whole face to perform head pose estimation [1]. Recently, due to the advent of real-time and low-cost 3D sensors, the focus of many researchers shifted towards estimating the head pose on facial depth images [9] [10]. However, one cannot rely on having depth information in unconstrained environments, where there is no control over the sensor that is being used to capture the images. According to Zhu and Ramanan [11], not only estimating extreme head poses is a difficult problem, but even face detection. Additionally, training in-the-wild pose estimators is not trivial as there is no reliable ground-truth [6]. Such poses are likely to be found in unconstrained environments and are not considered in many published works regarding face analysis. In our work, the focus is kept on 2D RGB images, including those with extreme poses.

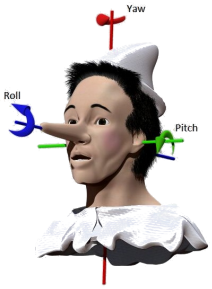

Fig. 1. The head yaw, pitch and roll

Pawelczyk and Kawulok [12] extract gradient information from the nose and use SVM to classify the pose into a discrete set of angles. This approach was only applied to controlled environment dataset. For estimating the pose in uncontrolled environments, Demirkus *et al.* [13] propose using a set of facial features to estimate a probability density function over the pose on each frame and aggregating the results using temporal information.

In this work we show that the nose region can be successfully used for head pose estimation in unconstrained environments. Unlike the eyes and ears, it is visible even in profile faces; unlike the mouth, it cannot be easily deformed by speech and expressions; it is also less likely to be partly occluded by accessories and facial traits, such as sunglasses and beards; when compared to using the whole face.

We developed a methodology, NosePose, composed of two different approaches for estimating the head pose only based on the nose region. The first one, SVM-NosePose, uses Support Vector Machines (SVM) trained with the output of the Local Gradient Increasing Pattern (LGIP) filter [14] on the nose region. The second approach, CNN-NosePose, makes use of Convolution Neural Networks (CNN). Both were tested on four unconstrained datasets to evaluate their repeatability. NosePose is landmark-free, does not take advantage of temporal information, treats pose estimation as a classification problem and estimates the angles based on a predefined set of discrete poses that depends on the dataset used for training.

II. Head Pose Estimation from The Nose Region

While SVM has already been shown effective to estimate the head pose [12], as it can be modelled as a classification

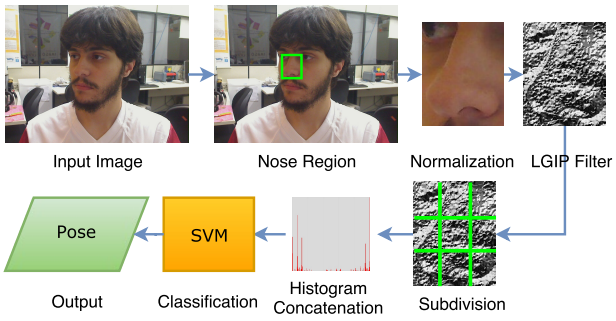

Fig. 2. SVM-NosePose

problem with a small number of classes, Convolutional neural networks appear as a suitable candidate for solving the same problem. Not only are they efficient at classification problems with a limited number of classes and have been used successfully for face recognition [15] [16], they also have the advantage of not having to explicity define the descriptors that are being used to extract features. When using SVM, the choice of the descriptor is, at times, empirical, using a CNN provides a solution for this. Unbalanced data is handled better by SVM [17], but CNNs can take advantage of very large datasets (100,000 images or more), while training SVM with too many images would be impractical as the training time increases considerably.

A. SVM-NosePose for Head Pose Estimation

Our SVM-NosePose strategy uses support vector machines for classifying a vector of extracted features into a discrete pose. Pawelczyk and Kawulok [12] propose the use of the raw gradient values as the feature vector. However, after conducting more tests, we found that histograms of the LGIP descriptor [14] can be applied to achieve higher head pose classification accuracy, due to its ability to describe the shape while still being robust to some variation. We exhaustively searched the number of subregion histograms for each dataset, to achieve maximum accuracy, however using 49 subregions yields good results on all datasets and only minor improvements were obtained using a different number.

Extracting the histogram of the nose subregions instead of the whole region enables some of the spatial information to be kept, while allowing some variations to occur, resulting in higher recognition rates. Our SVM-NosePose method uses C-SVM with a radial basis function kernel, is trained with 10-fold cross validation and is shown as a diagram in Figure 2.

B. CNN-NosePose for Head Pose Estimation

We developed a CNN architecture for estimating the head pose of a subject given the nose region. The CAFFE (Convolutional Architecture for Fast Feature Embedding) framework [18] was used for all experiments. It was chosen due to it being free, consistently documented, frequently updated, easy to use and having an active community[19] [20] [21].

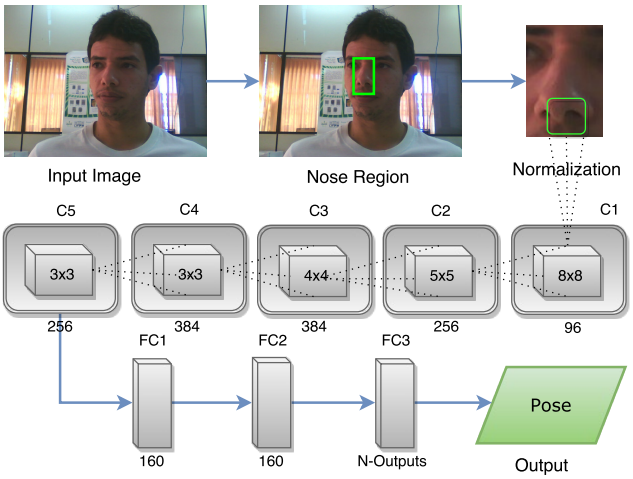

Fig. 3. CNN-NosePose

The proposed CNN architecture is similar to Krizhevsky *et al.*'s [22] and Hu *et al.*'s [23], the main difference lies in the network width and parameter definition, based on experiments performed on the IJB-A dataset [24], then applied to all datasets. The architecture is a deep convolutional neural network composed of five convolutional layers, followed by three fully connected layers (Figure 3).

We evaluate CNN-NosePose's performance using two techniques, cross-validation (CV) and splitting the data into three subsets, training, validation and testing. In both cases, the network is trained from scratch with 100,000 backpropagation iterations. After the training is finished, fine-tuning is performed for each different dataset in this work.

III. Experimental Results

Ground-truth nose regions were used on both training and testing subsets, which allows assessing the pose estimation performance without the influence of a nose detector. When reporting our accuracy, we also provide our weak score on datasets annotated with more than five classes. Weak scores are calculated considering off-by-one misses as hits.

A. McGillFaces

The McGillFaces database [25] consists of 18,000 frames extracted from video sequences of 60 unique subjects and their corresponding labels (face mask, gender and head yaw). However, only 10,500 frames are available publicly and only 6,665 frames have the head yaw annotation. The pose annotations is discretized into 9 possible angles (from -90 to 90 in steps of 23.5 degrees). During recording, the subjects were placed in different illumination and background conditions and were allowed free movement and object interaction. This resulted in a variety of arbitrary face scales, expressions, viewpoints and occlusions.

We manually annotated the nose region (bounding box) in all 6,665 images that have the pose annotation. For our tests, 3,208 images for training and 3,457 for testing, without overlapping subjects, were used. When using CNN-NosePose

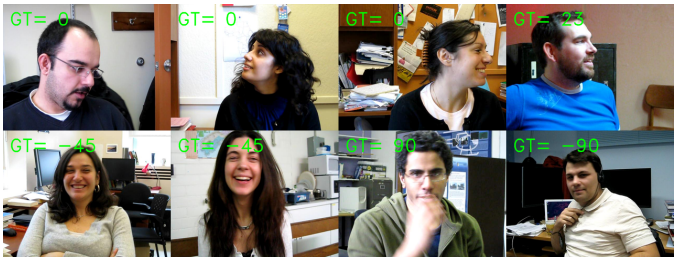

Fig. 4. Examples of inconsistent ground-truth annotations on the McGillFaces dataset

without cross-validation, the training subset is the same, the validation subset has 2,169 images and the testing subset, 1,288. Table I shows our results compared to Demirkus *et al.*, who reported, via personal communication, to have used all 18,000 images for training and testing. Because of this, the comparison is, unfortunately, biased as both of our methods would have benefited from using almost three times as many images.

B. Filtered McGillFaces

We investigated and evaluated the reliability of the provided pose annotations, since they were annotated semi-automatically [25]. Each image in the dataset with a label was evaluated by at least two different people, one by one in random order and was tagged either good or inconsistent. This visual analysis of the provided ground-truth annotation showed that approximately one fifth of the images were assigned inconsistent labels (Figure 4). Because of this, we also evaluate our algorithms using a filtered version of the McGillFaces dataset, containing only the images tagged as good. It contains 5,329 total images, 2,475 for training and 2,854 for testing (1,692 for validating and 1,162 for testing when not using cross-validation). The increase in accuracy was evident.

When performing our annotations, we learned that estimating the head pose is not a trivial task for humans, specially when there are more than 5 classes. Each annotated dataset requires multiple people, hours and revision rounds. Even when calibration is performed to establish the pose boundaries for the people annotating, hours of corrections are still necessary.

To evaluate the effects of using a larger training subset, we added 5,748 training images from the PaSC datasetand retrained. We call this experiment CNN+. A clear increase in accuracy can be noticed. All results are available in Table I.

TABLE I
COMPARATIVE RESULTS WHEN ESTIMATING THE YAW ON MCGILLFACES

	Strict	Weak
SVM (Original McGill) (3,457 images)	59.24%	83.34%
SVM (Filtered McGill) (2,854 images)	70.71%	92.68%
CNN (Original McGill) (3,457 images)	59.76%	88.08%
CNN (Filtered McGill) (2,854 images)	68.47%	94.53%
CNN (Original McGill w/o CV) (1,288 images)	70.50%	90.76%
CNN (Filtered McGill w/o CV) (1,162 images)	77.45%	97.50%
CNN+ (Filtered McGill) (2,854 images)	72.81%	94.57%
CNN+ (Filtered McGill w/o CV) (1,162 images)	85.46%	97.50%
[13] (18,000 images)	79.02%	–

C. PaSC Experiments

PaSC (Point-and-Shoot Challenge) [26] is an in-the-wild dataset with both videos and still frames subsets for face recognition with no pose annotations. It contains 9,376 challenging images of 293 subjects of different ethnic backgrounds in different environments, illumination conditions, poses and sensors.

The PaSC dataset is pre-divided into training and testing subsets optimized for evaluating face recognition. This subdivision proved to be poor for evaluating head pose estimation, as the distribution of the poses in the subsets varies greatly. We redivided the images in a way that this difference would be less noticeable while guaranteeing that no subject is present in both subsets.

For this experiment, we used only the still images which we were able to manually annotate the nose region and the head yaw (into 5 classes $[-90, 45, 0, 45, 90]$), resulting in 5,784 training and 6,243 testing images. When training CNN-NosePosewithout cross-validation, 3,172 images are used for validating and 3,071 for testing. Table II shows our results for both approaches.

TABLE II
COMPARATIVE RESULTS WHEN ESTIMATING THE YAW ON PASC

	Strict
SVM-NosePose (6,243 images)	86.91%
CNN-NosePose (6,243 images)	88.42%
CNN-NosePose w/o CV (3,071 images)	90.85%

D. AFW Experiments

The annotated faces in-the-wild (AFW) [11] dataset contains 468 faces with landmark and head yaw annotations in 13 different poses (from -90 to 90 in steps of 15 degrees). Large variations in the background, pose, expression and subject appearance are present, as the images were extracted from Flickr and are all from real world in-the-wild scenarios. We manually annotated all the corresponding nose regions.

TABLE III
COMPARATIVE RESULTS WHEN ESTIMATING THE YAW ON AFW

	Strict	Weak
SVM-NosePose	44.71%	81.21%
CNN-NosePose	49.43%	86.96%
CNN-NosePose w/o cross-validation	46.00%	86.00%
[11]	–	81.00%

To perform training, we augmented the AFW dataset, by mirroring and rotating the images, we were able to increase the number of images 14-fold, allowing us to use the dataset for both training and testing. When using cross-validation, 3752 images were used for training and 2800 for testing, without cross-validation, the size of the training subset is the same, 1800 images are used for testing and 1000 for validating. Our results are compared to Zhu and Ramanan's in Table III.

E. IJB-A Experiments

IJB-A (IARPA Janus Benchmark A) [24] is an in-the-wild dataset for face detection and recognition. The dataset contains a large geographic distribution, pose variation, occlusions and illumination. It contains 500 subjects distributed in 5,712 still images and 2,085 videos, with an average of 11.4 images and 4.2 videos per subject and no pose annotations. The images and videos were extracted performing searches on Creative Commons licensed image datasets.

A large subset of the IJB-A dataset was manually annotated (nose region and yaw) with a precision of 45 degrees. From all 10,014 annotated images, two subsets were generated, 2,414 images for testing and 5,820 for training, totalling 7,234 images. When cross-validation is not used for training CNN-NosePose, 1,500 images from the training subset are used for validation. To evaluate the performance on a larger dataset, the training subset is augmented ten-fold, by rotating and flipping. These experiments were not performed with SVM-NosePose, as the large number of images influences the total training time significantly. The achieved accuracy with both our methods is summarized in Table IV.

TABLE IV
COMPARATIVE RESULTS WHEN ESTIMATING THE YAW ON IJB-A

	Accuracy
SVM-NosePose (2,414 images)	76.47%
CNN-NosePose (2,414 images)	78.42%
CNN-NosePose augmented (2,414 images)	79.62%
CNN-NosePose w/o cross-validation (2,414 images)	76.39%
CNN-NosePose augmented w/o cross-validation (2,414 images)	76.93%

IV. Final Remarks

We presented our landmark-free NosePose methodology for head pose estimation solely based on the nose region using two approaches, SVM-NosePose and CNN-NosePose. Both approaches were tested on four different publicly available datasets and compared to state-of-the-art works, achieving favourable results. As part of future work, a complete system is being developed, including nose and face detection, tracking, pose estimation and expression and face recognition. All stages use landmark-free approaches, favoring robustness in extreme cases. Not only the combination of these different solutions will increase the pose estimation accuracy (e.g. by including temporal information), but the estimated pose will contribute to the accuracy of the system as a whole.

Acknowledgment

The authors would like to thank CAPES and CNPq for supporting this research.

References

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE TPAMI*, vol. 31, no. 4, pp. 607–626, 2009.

[2] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE TBME*, vol. 53, no. 6, pp. 1124–1133, 2006.

[3] Y. Lee, J. Phillips, J. J. Filliben, J. R. Beveridge, and H. Zhang, "Generalizing face quality and factor measures to video," in *IEEE IJCB*. IEEE, 2014, pp. 1–8.

[4] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *IEEE CVPR*, June 2015, pp. 4295–4304.

[5] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM CSUR*, vol. 35, no. 4, pp. 399–458, 2003.

[6] B. Martinez and M. Pantic, "Facial landmarking for in-the-wild images with local inference based on global appearance," *Image and Vision Computing*, vol. 36, pp. 40 – 50, 2015.

[7] J. Choi, G. Medioni, Y. Lin, L. Silva, O. Regina, M. Pamplona, and T. C. Faltemier, "3d face reconstruction using a single or multiple views," in *IEEE ICPR*, Aug 2010, pp. 3959–3962.

[8] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE TPAMI*, vol. PP, no. 99, pp. 1–1, 2016.

[9] S. Tulyakov, R.-L. Vieriu, S. Semeniuta, and N. Sebe, "Robust real-time extreme head pose estimation," in *IEEE ICPR*. IEEE, 2014, pp. 2263–2268.

[10] C. Papazov, T. K. Marks, and M. Jones, "Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features," in *IEEE CVPR*, 2015, pp. 4722–4730.

[11] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE CVPR*. IEEE, 2012, pp. 2879–2886.

[12] K. Pawelczyk and M. Kawulok, "Head pose estimation relying on appearance-based nose region analysis," in *ICCVG*. Springer, 2014, pp. 510–517.

[13] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, "Probabilistic temporal head pose estimation using a hierarchical graphical model," in *ECCV*. Springer, 2014, pp. 328–344.

[14] Z. Lubing and W. Han, "Local gradient increasing pattern for facial expression recognition," in *IEEE ICIP*. IEEE, 2012, pp. 2601–2604.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE CVPR*. IEEE, 2014, pp. 1701–1708.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, M. W. J. Xianghua Xie and G. K. L. Tam, Eds. BMVA Press, 2015, pp. 41.1–41.12.

[17] R. C. Prati, G. E. Batista, and D. F. Silva, "Class imbalance revisited: a new experimental setup to assess the performance of treatment methods," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 247–270, 2015.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[19] Z. Huang, R. Wang, S. Shan, and X. Chen, "Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning," *Pattern Recognition*, vol. 48, no. 10, pp. 3113–3124, 2015.

[20] J. Xu, A. Schwing, and R. Urtasun, "Tell me what you see and i will show you where it is," in *IEEE CVPR*, 2014, pp. 3190–3197.

[21] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*. Springer, 2014, pp. 184–199.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[23] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Li, and T. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," in *IEEE ICCVW*, 2015, pp. 142–150.

[24] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *IEEE CVPR*. IEEE, 2015, pp. 1931–1939.

[25] M. Demirkus, J. J. Clark, and T. Arbel, "Robust semi-automatic head pose labeling for real-world face video sequences," *Multimedia Tools Applications*, vol. 70, no. 1, pp. 495–523, May 2014.

[26] J. R. Beveridge, J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *IEEE BTAS*. IEEE, 2013, pp. 1–8.