LANDMARK BASED HEAD POSE ESTIMATION BENCHMARK AND METHOD

Philipp Werner *Frerk Saxen* *Ayoub Al-Hamadi*

Institute for Information Technology and Communications, University of Magdeburg, Germany
{Philipp.Werner, Ayoub.Al-Hamadi}@ovgu.de

ABSTRACT

Head pose estimation can help in understanding human behavior or to improve head pose invariance in various face analysis applications. Ready-to-use pose estimators are available with several facial landmark trackers, but their accuracy is commonly unknown. Following the goal to find the best landmark based pose estimator, we introduce a new database (called SyLaHP), propose a new benchmark protocol, and describe and implement a method to learn a pose estimator on top of any landmark detector (called HPFL). The experiments (including cross database) reveal that OpenFace comes with the best pose estimator. Further, HPFL models trained on top of landmark trackers outperform the respective built-in pose estimators. The SyLaHP database, source code, and trained models are publicly available for research.

Index Terms— Head pose estimation, facial landmarks, synthetic database, discriminative model.

1. INTRODUCTION

Head pose and gestures are known to play a considerable role in social interaction and nonverbal communication. We turn our head towards a conversational partner, nod to indicate understanding and agreement, and use additional gestures to indicate dissent, confusion, or consideration [1]. Studies found that head pose and movement correlate with several emotions, such as embarrassment and pride [2–5], with medical conditions, such as depression [6–8] and pain [9–11], and with perceived attractiveness [12]. Head pose estimation gives a coarse indication of a person's gaze direction and is required to estimate it accurately [1]. Knowledge of head pose can also help in face analysis applications, such as face recognition, expression recognition, or age/gender classification to gain invariance to head pose [9, 13].

Altogether, head pose estimation is an important component for understanding behavior in human computer interaction and video-based monitoring systems. We understand it as a tool that many researchers interested in human behavior want to use without a lot of effort. The goal of this work is to help them by comparing available head pose estimators and providing the opportunity to create new estimators easily. We focus on 2D landmark based methods due to the following reasons: (1) there is a growing number of landmark detectors/trackers that can be used for research purposes for free; (2) there is rapid progress in improving the landmark quality, including unconstrained scenarios with difficult lighting, out-of-plane head poses, and occlusions; (3) 2D methods do not need depth data, which is often not available. Further, landmark detection software often includes ready-to-use head pose estimators.

A survey on head pose estimation can be found in [1] (written in 2008). With improvements in depth sensors many recent works focused in 3D head pose (e.g. [14, 15]), which are out of scope here. Some recent 2D approaches, such as [16], learn head pose from texture features directly, but they typically only consider a very limited subset of the poses and/or only predict a coarse pose category. Landmark based approaches seem to be considered straightforward, as they are included in several landmark tracking software packages, but not detailed in publications. Chehra [17] and IntraFace [18] both include head pose estimators, but the underlying method and the accuracy are unknown. The CSIRO Face Analysis SDK [19], which is based on the face tracker by Saragih et al. [20], is open source software; this allows to figure out every detail of the method. To find the landmarks it fits a Constrained Local Model (CLM) to the image. The head pose is implicitly given in the global rigid transformation parameters of the underlying 3D shape model. OpenFace [21] is open source as well and also based on a 3D shape model. But it estimates the head pose after landmark detection by fitting the parametrized 3D shape model to the detected 2D landmarks using a pinhole camera model.

We did not find any information on the head pose estimation accuracy of IntraFace and CSIRO. OpenFace's accuracy has been evaluated on three datasets [21], outperforming Chehra. However, after looking at the source code we see some problems with the experimental results. First, they do not reflect the actual absolute difference between the estimation and ground truth (as implicitly stated in the paper), but between the respective movements relative to the sequence's first frame. Second, head pose estimation is only possible if the face was successfully detected or tracked, which is not always the case in challenging datasets; it is not clear how this

Funded by the German Research Foundation (project Al 638/3-2) and by the Federal Ministry of Education and Research (project 03ZZ0443G).

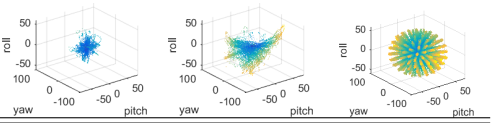
	BU [22]	BIWI [23]	**SyLaHP**
Subjects	6	20	**30**
Facial expr.	1	1	**20**
Samples	14k	16k	**101k**
Pose distribution			

Table 1. Comparison of BU, BIWI, and our new dataset. Pose angles in degree. Colors illustrate distance from zero pose.

was handled in the experiments of [21]. There are samples without a pose estimate, which cannot be included in the calculation of the mean absolute error measure. Thus, meaningful values can only be calculated from the subsets with successful estimations, which differ between methods; so their comparability is limited.

This work contributes the following: (1) We introduce a synthetic database (named SyLaHP) with perfect ground truth for benchmark and training of landmark based head pose estimators (Sec. 2), (2) we suggest a benchmark protocol aiming at better comparability (Sec. 3), (3) we describe a head pose estimation method that can be easily applied on top of any landmark detector (Sec. 4), and (4) we compare the accuracy several ready-to-use head pose estimators, including cross-database experiments (Sec. 5). The database, source code, and pose estimation models are publicly available for research purposes[1].

2. SYLAHP DATABASE

We introduce the **Sy**nthetic dataset for **La**ndmark based **H**ead **P**ose estimation (**SyLaHP**). Compared to other head pose databases, such as BU [22] and BIWI Kinect [23], it comprises more samples, more variation regarding identity, facial expression, and head poses, as well as perfect ground truth (see Table 1). It contains about 6,500 videos rendered using the FaceGen[2] 3D morphable model (3D-MM, similar to [24]).

Identity, facial expression, and head poses are varied systematically. Illumination and occlusions, which are other challenging factors for pose estimation, are not varied in the dataset, since they are handled better and better by current landmark detectors. Each of 30 subjects (with varying ethnicity, age, and gender) is combined with each of 20 facial expressions (including basic emotions, eye closure, and phonemes), resulting in 600 meshes (all created from 3D-MM). Each mesh is rendered in multiple head pose sequences. We define head pose as the orientation of a persons head relative to the view of a camera described by three angles (pitch, yaw, and roll) in a right-handed coordinate system (illustrated in Fig. 1 in [1]). The pose sequences start with a near-frontal pose and end in an extreme pose; frames in between were interpolated linearly with steps not greater than $5°$. Extreme poses were selected from the surface of an ellipsoid in the pose space with maximum pitch of $\pm 70°$, yaw of $\pm 90°$, and roll of $\pm 55°$. More precisely, we approximated the ellipsoid with an icosphere (2 subdivisions), which yielded 162 extreme poses (see scatter plot in Table 1), respectively 162 pose sequences. These were split into 15 subsets. Each mesh (subject/expression combination) was rendered with one of these subsets, i.e. with about 10 pose sequences. To further increase pose variability, we vary a sequence before rendering by adding a random number (range $\pm 2.5°$) to each pose angle.

Fig. 1. Example images from SyLaHP database with landmarks by OpenFace (top row), Chehra (2'nd row), IntraFace (3'rd row), and CSIRO (bottom row).

Next to the videos and pose ground truth, the database also includes the landmarks detected and tracked by OpenFace [21], Chehra [17], IntraFace [18], and CSIRO Face Analysis SDK [19, 20]. Fig. 1 shows some example images with tracked landmarks. The SyLaHP Database is publicly available for non-commercial research purposes.

3. BENCHMARK PROTOCOL

Pose estimation results depend on prior processing steps, i.e. on face detection and landmark localization. If any of these steps fails, no valid head pose estimation is possible. The remaining subset of the overall data, which is available for calculating the mean absolute error, usually varies, at least if different face detectors or landmark trackers are employed. So the mean errors may be calculated from very different sub-

[1] http://www.iikt.ovgu.de/LmHeadPoseEstBench.html
[2] http://facegen.com/

sets, limiting their comparability. Some authors handle this by reporting a "missed" rate, i.e. how many samples lack a pose estimate, and by computing the mean error measures on the rest of the data [15, 23]. Although this gives a clue on how robust the estimator is, the mean errors still vary with the miss rate, depending on how many of the more difficult samples are taken into account. So it is hard to compare the overall performance of a head pose estimation software with the mean error and miss rate.

As an alternative, we propose a new measure that can be calculated from absolute pose errors. It is based on the cumulative distribution function[3] (CDF), which is widely used in landmark localization and can be calculated from every error measure. We suggest to calculate it from absolute pitch, yaw and roll error, as well as from the mean of these three errors (for more compact presentation). To ensure comparability, the CDF is calculated from all samples, including those *without* a head pose estimate. As a consequence, the maximum value of the CDF is the proportion of samples *with* a valid estimate. The CDF plot also tells how errors are distributed; the steeper the function rises, the more samples have a low error. However, if there are many CDFs to compare, a graph will become crowded. As an easy-to-compare and compact measure we propose to calculate the area under the curve,

$$AUC_\alpha = \frac{1}{\alpha}\int_0^\alpha f(e)de, \tag{1}$$

where e is an error measure (e.g. absolute pitch error), $f(e)$ is the CDF of e on the sample set, and α is the upper bound of the integration, i.e. the maximum error that is considered acceptable. We divide by α to normalize the result so that the perfect estimator reaches a value of one. As we are mainly interested in high accuracy, we set $\alpha = 10°$

The CDF and AUC_α can also be computed for subsets of a database. Due to space limitations, we only plot CDF for the whole database. But we compute $AUC_{10°}$ for a lower pose angle subset, as some applications might only need estimation of close-to-frontal head poses.

4. HPFL METHOD

This Section describes a simple method for learning a **H**ead **P**ose estimator on top of a **F**acial **L**andmark detector (called **HPFL**). Although it might lack novelty, it is useful for the scientific community due to three reasons: (1) We describe the method in detail and publish the source code for training and prediction. For many other pose estimators, such as those coming with IntraFace or Chehra, neither information on the underlying method, nor open source code is available. (2) The method can be easily combined with *any* new facial landmark detector. Thus, head pose estimation can benefit from the rapid advances in landmark localization. (3) It performs best on both, SyLaHP-DB and BIWI Kinect Head Pose Database, see Sec. 5.

[3]The CDF represents the proportion of samples with an error less or equal to a threshold when varying this threshold. Example in Fig. 2.

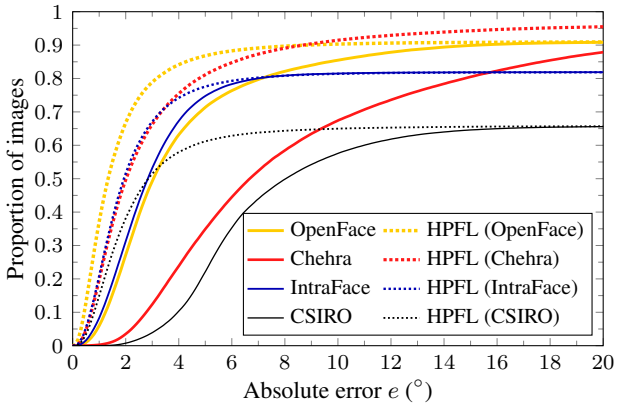


Fig. 2. Results on SyLaHP database (cumulative distribution function of pose estimation errors).

We predict the head pose through Support Vector Regression (SVR) [25] using the landmarks as features. To gain invariance regarding scale and translation, the landmarks are normalized based on the eyes center points (which we calculate from the inner and outer eye corners) in advance. We scale the points to an inter-ocular distance of one and translate them so that the center of both eyes is in the origin of the coordinate system. Further, we standardize the features, i.e. we subtract the mean of each feature and divide by its standard deviation. The resulting features are fed into an SVR with a radial basis function kernel. To keep the computation time in manageable magnitude, we learn the SVR on a randomly selected subset of the training set. SVR parameters C, γ, and ϵ are selected through a grid search on the training set [26]. For this purpose, the set is randomly split into two subsets without subject overlap and subsampled afterwards; one subset (with 500 samples) is used for training and one (with 5,000 samples) for validation. We validate each parameter combination with three different splits and average the results. After finding the best performing parameters, the SVR is retrained on the whole training set with more samples (see Sec. 5).

5. EXPERIMENTS

On two databases, we compare the performance of head pose estimators coming with OpenFace [21], Chehra [17], IntraFace [18], and CSIRO Face Analysis SDK [19, 20]. Further, we applied the proposed HPFL method on top of the 4 landmark trackers.

On SyLaHP database (see Sec. 2) the HPFL method was evaluated using 5-fold cross validation, in which we ensured that subjects occurring in a training set do not occur in the corresponding test set. For each fold, HPFL was trained on

Head Pose Estimatior	Landmarks	All data				Low angle subset			
		Pitch	Yaw	Roll	Mean	Pitch	Yaw	Roll	Mean
Provided with landmark tracker	OpenFace	**0.533**	0.576	0.683	0.579	0.622	**0.820**	0.862	**0.767**
	Chehra	0.325	0.409	0.443	0.325	0.461	0.599	0.565	0.522
	IntraFace	0.499	**0.608**	**0.684**	**0.593**	**0.641**	0.762	**0.902**	**0.767**
	CSIRO	0.261	0.291	0.310	0.248	0.395	0.485	0.433	0.414
HPFL	OpenFace	**0.704**	**0.760**	**0.803**	**0.752**	**0.830**	**0.901**	**0.944**	**0.891**
	Chehra	0.627	0.699	0.781	0.686	0.761	0.851	0.924	0.845
	IntraFace	0.570	0.665	0.727	0.650	0.750	0.867	0.935	0.851
	CSIRO	0.469	0.496	0.577	0.509	0.677	0.724	0.831	0.739

Table 2. Results on SyLaHP database (AUC_{10° measure). We compare head pose estimators coming with landmark tracking softwares and HPFL (see Sec. 4) trained on top the same landmark trackers (rows). The AUC_{10° measures were calculated from for pitch, yaw, and roll angle absolute errors as well as the mean of these absolute errors (columns) for all data of SyLaHP (left) and for a low angle subset (right), i.e. pitch, yaw, and roll not more than 30°.

6k samples randomly selected from the training set (due to computational complexity of SVR). Then, the CDF was calculated from predictions of all folds. The pose estimators coming with the considered landmark trackers do either not rely on training or could not be retrained, because the method is unknown; so their ready-to-use models were applied for all samples directly.

Fig. 2 shows the CDFs calculated on the mean of (each sample's) absolute pitch, yaw, and roll error. The values in which the curves saturate, reflect the number of samples for which the underlying landmark tracker provides output; essentially this is the robustness of the tracker regarding extreme poses. On SyLaHP, Chehra is the most robust tracker, followed by OpenFace, and IntraFace. However, Chehra does not provide the most accurate pose estimates. Regarding the estimators coming with trackers, Chehra is clearly outperformed by IntraFace and OpenFace. But pose estimation with Chehra landmarks can be much better, as we see a big improvement when applying HPFL. HPFL also improves results when applied on top of the other trackers (compared to the respective built-in pose estimators). Table 2 lists more results as AUC_{10° values. HPFL trained with OpenFace landmarks outperforms all others estimators in all error measures. This is probably due to good landmark quality combined with extra information from chin-line landmarks (which are not available for Chehra and IntraFace). The best built-in estimators are IntraFace and OpenFace, depending on the requirements.

In contrast to SyLaHP, the BIWI Kinect database [23] has no perfect ground truth (mean rotation error of 1° [23], systematic view-point errors due to distance between depth and RGB camera) and nearly no variation in facial expression. However, to show generalizability of the HPFL method, we evaluated it (and the other estimators) on the BIWI. We trained each HPFL variant on 20k samples randomly selected from the whole SyLaHP database and tested on BIWI afterwards. CDFs of the results are depicted in Fig. 3. OpenFace performs best. Applying HPFL reduces the error for OpenFace, Chehra, and CSIRO.

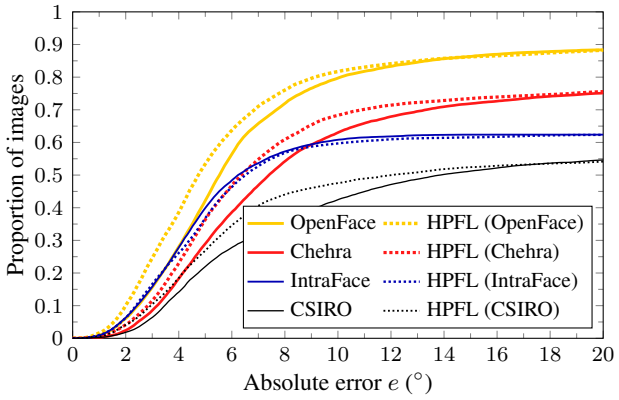


Fig. 3. Cross-database results on BIWI database (CDF of pose estimation errors). HPFL trained on SyLaHP database.

6. CONCLUSION

In this paper we introduced SyLaHP, a comprehensive synthetic database for landmark based head pose estimation. We described HPFL, a method for learning a head pose estimator on top of any landmark detector. The database, source code, and trained models are publicly available. Further, we proposed a new benchmark protocol for pose estimators. In experiments (including cross-database) we compared eight estimators that are available for non-commercial research.

OpenFace and IntraFace provide the most accurate out-of-the-box head pose estimation. However, OpenFace's face tracking performs better for extreme head poses, i.e. it still provides reasonable landmarks for many poses when IntraFace lost track. Nevertheless, if you are also interested in facial expression analysis, IntraFace might be the better choice, because OpenFace more often fails to track facial deformations (see Fig. 1). So far, CSIRO has been often used to estimate head poses in behavior analysis studies [6, 27, 28]; for future studies we recommend to avoid CSIRO, since all other options clearly performed better.

7. REFERENCES

[1] E. Murphy-Chutorian and M. M Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *TPAMI*, vol. 31, no. 4, pp. 607–626, 2009.

[2] A. Mignault and A. Chaudhuri, "The Many Faces of a Neutral Face: Head Tilt and Perception of Dominance and Emotion," *Journal of Nonverbal Behavior*, vol. 27, no. 2, pp. 111–132, 2003.

[3] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[4] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 17–34, Mar. 2009.

[5] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame.," *Journal of Personality and Social Psychology*, vol. 68, no. 3, pp. 441, 1995.

[6] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and Vision Computing*, vol. 32, no. 10, pp. 641–647, Oct. 2014.

[7] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

[8] P. Waxer, "Nonverbal cues for depression.," *Journal of Abnormal Psychology*, vol. 83, no. 3, pp. 319, 1974.

[9] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. Traue, "Automatic Pain Assessment with Facial Activity Descriptors," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, 2016.

[10] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic Pain Recognition from Video and Biomedical Signals," in *ICPR*, 2014.

[11] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges," in *BMVC*, 2013.

[12] E. Krumhuber, A. S. R. Manstead, and A. Kappas, "Temporal Aspects of Facial Displays in Person and Expression Perception: The Effects of Smile Dynamics, Head-tilt, and Gender," *Journal of Nonverbal Behavior*, vol. 31, no. 1, pp. 39–56, Dec. 2006.

[13] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild," in *CVPR*, 2015, pp. 787–796.

[14] R. Niese, P. Werner, and A. Al-Hamadi, "Accurate, Fast and Robust Realtime Face Pose Estimation Using Kinect Camera," in *International Conference on Systems, Man, and Cybernetics (SMC)*, 2013.

[15] 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 *CVPR*, 2015.

[16] B. Ma, X. Chai, and T. Wang, "A novel feature descriptor based on biologically inspired feature for head pose estimation," *Neurocomputing*, vol. 115, pp. 1–10, 2013.

[17] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental Face Alignment in the Wild," in *CVPR*, 2014.

[18] X. Xiong and F. De la Torre, "Supervised Descent Method and its Applications to Face Alignment," in *CVPR*, 2013.

[19] M. Cox, J. Nuevo, J. Saragih, and S. Lucey, "CSIRO face analysis SDK," 2013.

[20] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-Shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[21] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: an open source facial behavior analysis toolkit," in *Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[22] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models," *TPAMI*, vol. 22, no. 4, pp. 322–336, Apr. 2000.

[23] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random Forests for Real Time 3d Face Analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[24] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3d Faces," in *SIGGRAPH*, 1999.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM – A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[26] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," Tech. Rep., Department of Computer Science, National Taiwan University, 2003.

[27] Z. Hammal, J. F. Cohn, and D. T. George, "Interpersonal Coordination of Head Motion in Distressed Couples," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 155–167, Apr. 2014.

[28] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head Movement Dynamics during Play and Perturbed Mother-Infant Interaction," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 361–370, Oct. 2015.