https://doi.org/10.2352/ISSN.2470-1173.2017.19.AVM-025

Face Pose Estimation From Rigid Face Landmarks For Driver Monitoring Systems

Bhawani Shankar, Dakala Jayachandra, and Kalyan Kumar Hati; Path Partner Technology Pvt Ltd; Bangalore, India

Abstract

Face pose contains rich information about the intent of a person, hence, estimating the face pose is important in assessing the attention of the driver. Most of the methods for pose estimation derive some image features and then either model the appearance (3D or 2D) or apply regression on the features. But these methods have high computational costs. On the other hand, we aim to estimate pose from only the facial landmark locations. In most driver monitoring systems, the important facial landmarks are readily available as they are essential in assessing driver drowsiness. Therefore, we utilize the existing eye landmarks along with nose and mouth landmarks to estimate the face pose. For this, we propose to apply linear regression on features derived only from the 2D facial landmark locations. Instead of relying on a single linear regression model, we propose to apply a global linear model to predict the pose and then refine the predicted pose by applying a local model built for that pose region. Local models are built using partially overlapping subsets of training samples. The experiments on Pointing'04, MultiPIE, and Biwi Kinect datasets show that the proposed two-level models achieve accuracy comparable to that of the state-of-the-art methods. At the same time, the proposed method can process 2000 frames per second in Octave.

Index terms— Driver monitoring system, face pose, linear regression.

Introduction

It has been widely studied and understood that the driver drowsiness and inattention are the main causes for about $20-40\%$ of the road accidents every year. Hence, it is important to develop a reliable and cost effective driver monitoring system that can detect inattention and alert the driver. Face pose contains rich information about the attention of the driver. In this work, we aim to develop a cost effective pose predictor from the facial landmarks of the driver.

There has been a lot of work on face pose estimation, a good review can be found in [1]. Broadly, the approaches can be summarized as follows. Appearance based methods and 3D face model reconstruction methods [2] fit appearance and shape on the given 2D face image and then infer the pose. There are many methods, such as [3]-[8], that apply classification or regression on features extracted from the image pixels. Methods such as [9] exploit the fact that face detection models implicitly encode face pose information. Subspace embedding or local subspace learning methods, like in [10], learn local subspaces for making the models adaptable to the variations present across different poses. Geometric methods, such as in [11, 6], try to model the geometric relation between face pose and facial landmarks.

As mentioned, driver drowsiness detection is an important task for any driver monitoring system. For performing this, the system needs to assess activities like sleepy eyes, yawning and talking which essentially needs landmarks of eyes, nose and mouth. Hence, we assume that these important facial landmarks are already available. Now, without adding any significant computation, we aim to estimate the face pose using only the face landmark locations.

For geometric methods, it is difficult to explicitly and accurately model the complex geometric relationships. Instead, we propose to model the relationship as a linear relationship between features (extracted from facial landmark locations) and the face pose. To derive this relationship from the available data, we propose to use a regression method. However, regression methods typically fail to generalize (they overfit) when trained with limited data and they get biased when the training data is non-uniform. A linear regression model overcomes the generalization problem but a single linear model fails to capture finer variations in the predictor-outcome relationship over the full pose span. To overcome these limitations, we propose to build piece-wise linear regression models in a hierarchical way. The method in [8], first applies image feature based coarse pose prediction and then applies a pose regression on landmark locations combined with the predicted coarse pose. In our approach, we don't use any image features and limit only to the landmark locations. This eliminates the need to develop a robust pose predictor based on image features, just developing a robust landmark detector is sufficient. Also, choosing only linear regression steps in our model makes it computationally efficient.

Rest of the paper is organized as follows. The next section explains our proposed method. Subsequent section, explains the experiments and observations on MultiPIE [12], Pointing'04 [6] and BIWI-Kinect Head Pose datasets. We conclude the paper with a discussion on future work in the Conclusions section.

Proposed Method

In this section, we describe the landmarks that we choose for pose prediction, feature vector derived from landmarks, and the proposed regression applied on those feature vectors.

Approach: Our intuition is that the relative positions of facial landmarks and the corresponding face pose has an approximately linear relationship. We first build a global linear model using all the samples spanning over the complete pose variation under consideration. Then, we define a set of pose locations, denoted as model centers, at which we build local linear models using samples from a fixed subset of pose variations around the given model center. In summary, *we apply a global linear model to predict the pose and then refine the predicted pose by applying a local model close to the predicted pose.* Rest of this section explains the details of building one linear model.

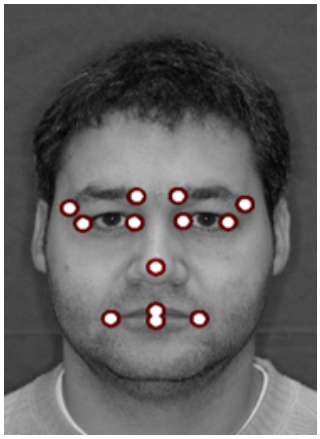

Figure 1. *Chosen face landmarks.*

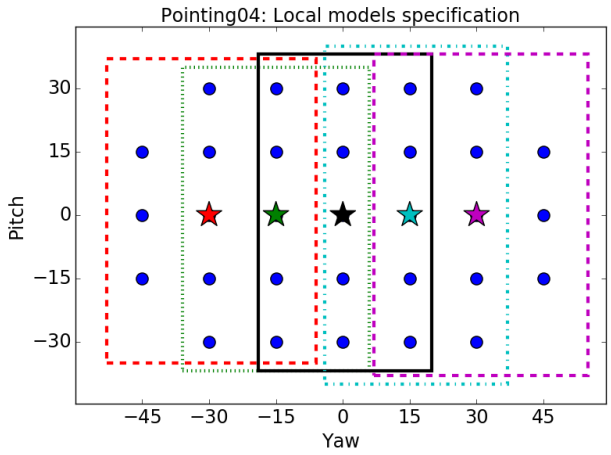

Figure 2. *The circular dots (and stars) indicate the selected pose range where the 13 landmarks are visible. On Pointing'04 [6], local models' pose centers and their corresponding pose ranges are indicated by the star symbols and their respective color rectangles.*

Feature vector: We choose the following 13 landmarks depicted in Figure 1: eye-brow corner points, eye corner points, nose-tip, mouth corner points, midpoint of upper lip and midpoint of lower lip. Except for the midpoints on lips, rest of the landmarks are rigid corners, hence, they can be detected with good accuracy. Intuitive reasons for the choice of the landmarks are as follows. Roll angle changes are reflected in the eye landmarks. Yaw angle changes are reflected in the relative distance between the nose-tip and eye corner points. Pitch angle changes are reflected in the change in depth information. To some extent, depth changes are implicitly encoded in the relative change in distances between the nose-tip and corners of eyes and mouth.

We derive normalized relative distances of the given landmarks as feature vector. For this, we define the landmark location of nose-tip as the reference point. Then, for each of the remaining landmarks, we compute the distance with respect to the reference point along both x and y directions. We form two vectors, one of all the distances along x and the other of all the distances along y. Denote them as $\mathbf{d_x}$ and $\mathbf{d_y}$, respectively. By concatenating both $\mathbf{d_x}$ and $\mathbf{d_y}$ we form a feature vector $\mathbf{d}'$.

Normalization: To make the feature vector invariant to the face size, we normalize the feature vector $\mathbf{d}'$ with a constant $k = max(\mathbf{d_y}) - min(\mathbf{d_y})$. The normalized feature vector $\mathbf{d} = \mathbf{d}'/k$. Notice that, the normalization constant k is derived from only the y components. This is chosen because, in general, the span of (chosen) facial landmarks is larger in the vertical direction and also relatively more consistent across different pose angles when compared to that in the horizontal direction.

Linear model: Given a set of landmark locations, we derive the normalized feature vector $\mathbf{d}$ and append it with a constant (=1, for the intercept). Denote the corresponding face pose as p, where p may be one of the pose components: yaw, pitch, and roll. Now, we model the relation between $\mathbf{d}$ and p as follows. $\beta^T\mathbf{d} = p$ Where, β is the model parameters to be learnt.

Training: Given the training samples, the model parameters in β are obtained as a closed form solution of linear least-squares regression, given by: $\beta = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{p}$ Where $\mathbf{D}$ is the feature-matrix. Each row of the matrix $\mathbf{D}$ contains feature-vector $\mathbf{d}^T$ derived from a single training sample. Column vector $\mathbf{p}$ has the corresponding target variable for each row of $\mathbf{D}$.

In our two-level model, we build one global linear model using all the training samples and then we build local linear models using partially overlapping local subsets of training samples.

Testing: Given a test face image with landmarks, the pose prediction using either global or local linear model is obtained simply by a single vector multiplication: $p' = \beta^T\mathbf{d}$

In our two-level model, we first apply the global model to obtain a coarse pose prediction. We then use Euclidean distance to find the model center nearest to this predicted pose and apply the corresponding local linear model to get the final pose prediction.

Experiments and Observations

In this section we evaluate the proposed method on three publically available datasets named MultiPIE [12], Pointing'04 [6] and Kinect [13], and compare with recently published results on these datasets. As our proposed method needs the 13 landmarks shown in Figure 1, we limit our experiments to the pose range where all these 13 landmarks are visible. Pose prediction from occluded landmarks for other pose angles will be taken up in future work. Based on the visibility of these 13 landmarks we choose the following yaw and pitch range in our experiments.

Yaw: +45°to -45°.

Pitch: +30°to -30°.

Rest of the section describes the experiments and the observations on results.

Evaluation metrics: We use two metrics to evaluate the proposed method. One is the mean absolute error (MAE) along with the standard deviation (SD) of error in the predicted pose. The other metric we use is the accuracy measured as a percentage of samples for which the magnitude of prediction error is less than some fixed tolerance.

Train-Test split: On all the three datasets we use the following train and test data split. Spreading uniformly over the pose angles in the given dataset, we pick 70% of the samples randomly as the training set, and the rest 30% for testing set.

Pointing'04

Pointing'04 [6] dataset has uniform distribution of samples over yaw ($\pm 90^\circ$) and pitch ($\pm 30^\circ$). The dataset does not provide facial landmarks annotations. We got the images manually annotated for 13 facial landmarks as in Figure 1. As mentioned earlier, based on the visibility of the chosen landmarks, we choose the pose range of $\pm 45^\circ$ for yaw and $\pm 30^\circ$ for pitch (excluding yaw = $\pm 45^\circ$ with pitch = $\pm 30^\circ$ combinations). We got around 1k images. From each discrete yaw and pitch combination, we randomly pick 70% of the samples as training set and the rest as testing set.

Local and global models: We build five local linear regression models and one global regression model. We use the complete training set to build the global regression model. This model serves as a coarse pose predictor and comprises the first level of our combined model. We build five local linear regression models centered at:
Center-1: (Yaw = -30°, Pitch = 0°),
Center-2: (Yaw = -15°, Pitch = 0°),
Center-3: (Yaw = 0°, Pitch = 0°),
Center-4: (Yaw = +15°, Pitch = 0°), and
Center-5: (Yaw = +30°, Pitch = 0°).

Each of these local linear regression models are built using the samples within $\pm 15^\circ$ of yaw and $\pm 30^\circ$ of pitch from their respective model center. For the local models, pose centers and their corresponding pose ranges are indicated by the star symbols and their respective color rectangles in Figure 2.

MultiPIE

MultiPIE [12] dataset has samples for discrete yaw angles ranging from -90° to +90° at uniform intervals of 15° with no variation in the pitch and roll components. The dataset does not provide facial landmarks annotations which we need for our pose model. We choose a subset of MultiPIE for which Zhu and Ramanan [7] have provided landmark annotations. We pick the pose (yaw angle) range of -45° to +45° based on the visibility of the 13 chosen landmarks (see Figure 1). We got around 4k images. From each discrete yaw bin, we randomly pick 70% of the samples as training set and the rest as testing set.

Local and global models: We build five local linear regression models and one global regression model. We use the complete training set to build the global regression model. We build five local linear regression models with the following model pose centers: Yaw = $\{$-30°, -15°, 0°, +15°, +30°$\}$. Each of these local linear regression models are built using the samples within $\pm 15^\circ$ of yaw from their respective model center.

BIWI-Kinect

Biwi-Kinect [13] head pose dataset has continuous real valued pose annotation for 15k images. This dataset does not provide facial landmark annotations. We use OpenFace [14] detector to obtain facial landmarks. We visualize the detected landmarks and manually filter out samples which have large landmark localization errors. Then, based on the visibility of our chosen 13 facial landmarks, we choose the pose range of $\pm 45^\circ$ for yaw, $\pm 30^\circ$ for pitch and $\pm 20^\circ$ for roll. We got around 6.5k samples. Figure 3 shows their distribution for yaw and pitch combination. We randomly pick 70% of the samples uniformly spread over the pose ranges as training set and the rest as testing set.

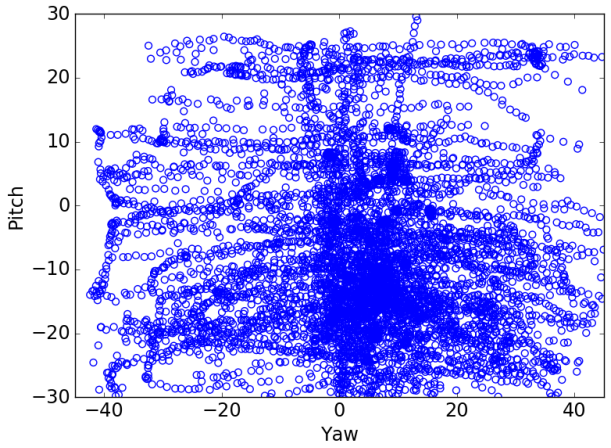

Figure 3. *Data distribution on Kinect [13] dataset for yaw and pitch variations.*

Local and global models: We build 18 local linear models and one global regression model. We train the global model using the complete training set. For building the local linear models, the 18 possible combinations of 3 (-22.5°, 0°, +22.5°) yaw angles, 3 (-15°, 0°, +15°) pitch angles and 2 (-5°, +5°) roll angles are used as model centers.

Each of these local linear regression models are built using the samples within $\pm 22.5^\circ$ of yaw , $\pm 30^\circ$ of pitch and $\pm 10^\circ$ of roll from their respective model center.

Observations

Table 1 and Table 2, respectively, shows the mean absolute error (MAE) and accuracy on all the three datasets. On all the datasets, the proposed two level model performs better than the single global model.

Roll angle variations are there only in Kinect [13] dataset. As can be seen from Table 1, for the roll angle, the proposed method achieved a mean absolute error (MAE) of $1.47^\circ \pm 2.0^\circ$. This suggests that the face landmarks are good enough to predict the roll angle. For yaw angle, both on MultiPIE [12] and Kinect [13], the proposed method achieved MAE of $2.9^\circ \pm 4.6^\circ$ and for pitch on Kinect [13] it achieved MAE of $3.01^\circ \pm 4.16^\circ$. These results are comparable to the state-of-the-art methods. On Pointing'04 [6], for yaw, the proposed method achieved MAE of $5.50^\circ \pm 7.07^\circ$ and for pitch it achieved MAE of $7.76^\circ \pm 9.71^\circ$. Whereas, the best results on Pointing'04 [6] are MAE of $4.24^\circ \pm 0.17^\circ$ for yaw and MAE of $2.69^\circ \pm 0.19^\circ$ for pitch by X. Geng et al., in [3]. As can be seen for Table 2, for a tolerance of $\pm 15^\circ$, the proposed method has achieved more than 95% accuracy for yaw and close to 90% accuracy for pitch on Pointing'04 dataset, and more than 95% for yaw and pitch on both MultiPIE and Kinect datasets. Note that the results for most of the recent methods are on entire datasets, whereas, the above mentioned results are obtained by applying our proposed method to a subset of these datasets. However, the results obtained are very promising.

Figure 4 shows the pose prediction results for some of the test images of Kinect [13] dataset where the proposed method achieved good results. Pointing'04 [6] dataset, pose labels are obtained by asking the person to look at some pre-defined mark-

MAE and SD results on three datasets. The proposed single regression model is named as "Global model" whereas the proposed two level model that combines global model and the local models is named as "Combined model."

Dataset	Model	Yaw		Pitch		Roll	
		MAE°	SD°	MAE°	SD°	MAE°	SD°
MultiPIE [12]	Global model	5.02°	6.70°	-	-	-	-
	Combined model	**2.93**°	4.11°	-	-	-	-
	D. Huang [10]	4.33°	-		-	-	-
Pointing'04 [6]	Global model	7.23°	9.17°	8.73°	10.51°	-	-
	Combined model	5.50°	7.07°	7.76°	9.71°	-	-
	Hulens [9]	11.25°	-	-	-	-	-
	N. Gourier [6]	5 to 15°	-	-	-	-	-
	X. Geng [3]	**4.24**°	0.17°	**2.69**°	0.19°	-	-
	N. Alioua [4]	6.1°	-	4.6°	-	-	-
	G. L. Marcialis [11]	9.6°	-	13.6°	-	-	-
BIWI-Kinect [13]	Global model	4.83°	6.92°	4.48°	6.17°	1.59°	2.12°
	Combined model	**2.98**°	4.67°	**3.01**°	4.16°	**1.47**°	2.00°
	A. Schwarz [15]	5.1°	9.5°	3.9°	6.1°	4.2°	7.3°
	Gabriele Fanelli [2]	3.8°	6.5°	3.5°	5.8°	5.4°	6.0°
	J. Chen [5]	9.9°	12.4°	12.9°	17.2°	6.9°	9.8°

Accuracy results on three datasets. The proposed single regression model is named as "Global model" whereas the proposed two level model that combines global model and the local models is named as "Combined model."

Dataset	Model	Yaw			Pitch		
		$< 5°$	$< 10°$	$< 15°$	$< 5°$	$<10°$	$< 15°$
MultiPIE [12]	Global model	60.03 %	88.55 %	96.21 %	-	-	-
	Combined model	84.36 %	96.93 %	99.35 %	-	-	-
	Zhu [7]	91.4 %	97.0 %	99.99 %	-	-	-
Pointing'04 [6]	Global model	41.29 %	74.89 %	90.68 %	31.98 %	61.13 %	86.23 %
	Combined model	56.27 %	83.40 %	95.14 %	37.65 %	68.42 %	89.87 %
	Hulens [9]	-	-	72 %	-	-	-
	X. Geng [3]	73.3 %	-	-	86.24 %	-	-
BIWI-Kinect [13]	Global model	66.09 %	88.24 %	94.51 %	67.46 %	88.87 %	97.79 %
	Combined model	83.29 %	95.39 %	98.57 %	81.67 %	97.20 %	99.31 %

ers on a wall. The way each person may orient their head while looking at the given marker is subjective, hence, the ground truth pose labels may be inconsistent. Figure 5 shows the results for some of the images from the test sets. Ground truth pose given in the datasets are same for the images in each row. Clearly, we can see the inconsistency between the given ground truth pose and the observable actual face pose. Notice that for such cases the predicted pose of the proposed method looks perceptually more appropriate than the ground truth pose.

The proposed method takes only one second to predict the pose for about 2000 images in Octave. In summary, in spite of inconsistent pose labels and non-uniform sample distribution, the results of the proposed two-level linear regression model is very encouraging. By preparing dataset with uniform sample distribution over pose using a more accurate pose sensor, the results for the proposed method may improve further. Moreover, recent works on landmarks detection, such as in [16], detect the landmarks visibility score. Utilizing the visibility score under different pose ranges, we foresee that the proposed method can be extended to work in all pose variations.

Conclusions

Using 13 visible rigid facial landmarks, we have proposed to apply a global linear regression followed by a local linear regression to predict the pose in the range of $\pm 45°$ yaw, $\pm 30°$ pitch, and $\pm 20°$ roll. The results on the datasets MultiPIE, Pointing'04, and Biwi-Kinect are comparable to the state-of-the-art methods. On Biwi-Kinect dataset, which has continuous pose values, the proposed method achieved pose prediction with mean absolute error within $5°$. If we build our models on a dataset with more uniform samples with continuous pose values the results may improve further. For assessing the attention of a driver from face pose, pose prediction with mean absolute error within $5°$ would be reasonably sufficient. Notice that below $5°$ even a human may not be able to differentiate well. Assuming a thorough customization with more uniform and consistent data, we conclude that the proposed method is good enough for face pose prediction for driver monitoring systems.

Encouraged by the results of the proposed method and its low computational complexity, we will work on collecting uniform samples with more accurate pose sensors, and then work on extending the proposed method to cover all pose variations by

utilizing the landmark visibility score.

References

[1] E. Murphy-Chutorian and M. M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. *In IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607-626, Apr. 2009.

[2] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Gool. Random Forests for Real Time 3D Face Analysis. *Int. J. Comput. Vision*, vol. 101, no. 3, pp. 437-458, Feb. 2013.

[3] X. Geng and Y. Xia. Head Pose Estimation Based on Multivariate Label Distribution. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1837-1842.

[4] N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, and Md. Rziza. Driver head pose estimation using efficient descriptor fusion. *J. Image Video Proc.*, vol. 2016, no. 1, pp. 1-14, Jan. 2016.

[5] J. Chen, J. Wu, K. Richter, J. Konrad, and P. Ishwar. Estimating head pose orientation using extremely low resolution images. *In IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, 2016, pp. 65-68.

[6] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. *In Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, 2004.

[7] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2879-2886.

[8] J. Whitehill and J. R. Movellan. A discriminative approach to frame-by-frame head pose tracking. *8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1-7.

[9] D. Hulens, K. Van Beeck, and T. Goedemi. Fast and Accurate Face Orientation Measurement in Low-resolution Images on Embedded Hardware. *In 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016, pp. 538-544.

[10] D. Huang, M. Storer, F. De la Torre, and H. Bischof. Supervised local subspace learning for continuous head pose estimation. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2921-2928.

[11] G. L. Marcialis, F. Roli, and G. Fadda. A novel method for head pose estimation based on the Vitruvian Man. *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 1, pp. 111-124, Feb. 2014.

[12] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[13] R. Min, N. Kose, and J. L. Dugelay. KinectFaceDB: A Kinect Database for Face Recognition. *In IEEE Trans. Syst. Man Cybern. A., Syst. Humans*, vol. 44, no. 11, pp. 1534-1548, Nov. 2014.

[14] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. *CMU-CS-16-118, CMU School of Computer Science, Tech. Rep.*, 2016.

[15] A. Schwarz, Z. Lin, and R. Stiefelhagen. HeHOP: Highly efficient head orientation and position estimation. *In IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-8.

[16] Brandon M. Smith and Charles R. Dyer. Efficient Branching Cascaded Regression for Face Alignment under Significant Head Rotation. *Arxiv*, Nov, 2016.

Author Biography

Bhawani Shankar *received his B. Tech degree in electronics and communications engineering from Gandhi Engineering College, Bhubaneswar, India in 2013. He is currently working with Path Partner Technology Pvt Ltd, India. His interests include multimedia-codecs, statistical learning and artificial intelligence.*

Dakala Jayachandra *received the B.Tech degree in electronics and communications engineering from NIT, Warangal, India in 2003 and received PhD in electrical and electronic engineering from Nanyang Technological University, Singapore in 2015. Currently, he is heading the ADAS vision team at Path Partner Technology Pvt Ltd, India. His research interests include 1-D and 2-D signal representations and their application to computer vision problems.*

Kalyan Kumar Hati *received the M.Tech degree from the Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela, India, in 2013. He is currently working as a Senior Software Engineer in PathPartner Technology. Since 2011, he has been working in various computer vision problems. His research interest include image processing, pattern recognition, and machine learning methods in computer vision.*

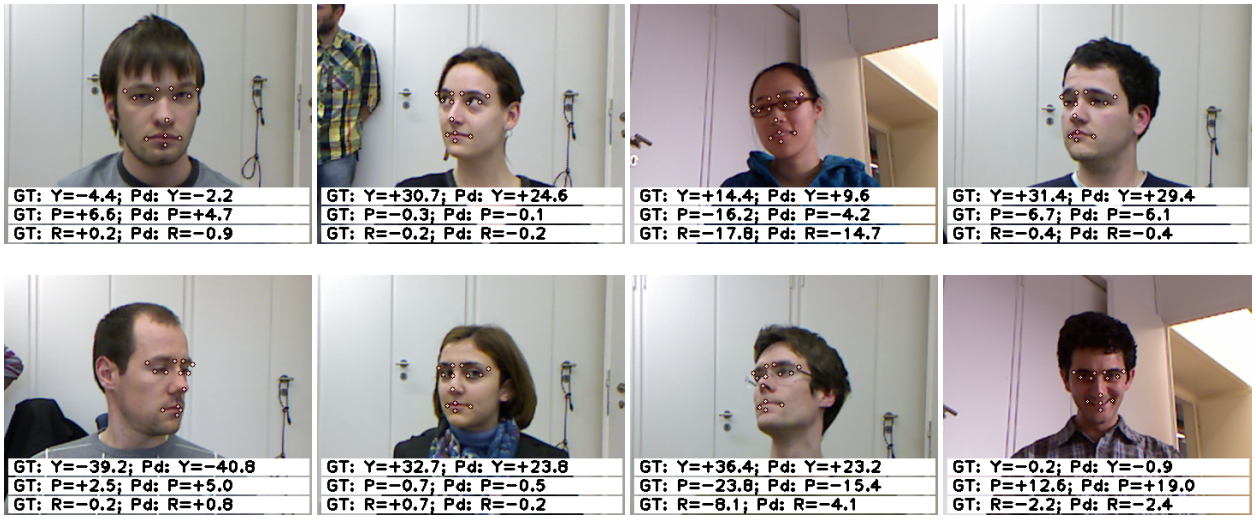

Figure 4. *Sample images from the test set of Biwi Kinect dataset [13]. Notations: GT = ground truth pose, Pd = predicted pose; Y = yaw, P = pitch, and R = roll.*

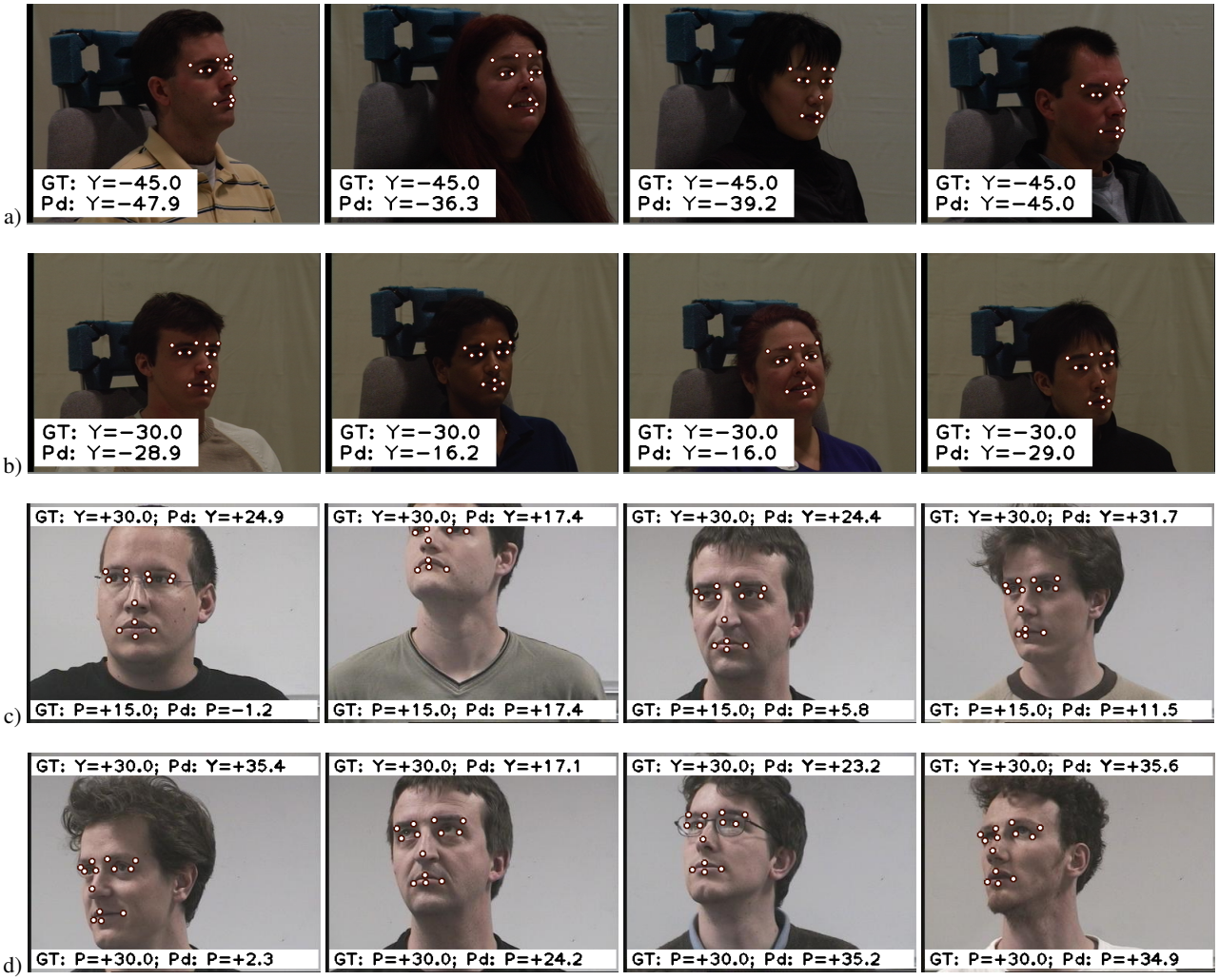

Figure 5. *Sample images from the test set of MultiPIE [12] (rows a and b) and Pointing'04 [6] (rows c and d). Notations: GT = ground truth pose, Pd = predicted pose, Y = yaw.*