Advancement in the head pose estimation via depth-based face spotting

Anwar Saeed, Ayoub Al-Hamadi, and Sebastian Handrich
Institute for Information Technology and Communications (IIKT)
Otto-von-Guericke-University Magdeburg
D-39016 Magdeburg, P.O. Box 4210 Germany
Email: {Anwar.Saeed, Ayoub.Al-Hamadi, Sebastian.Handrich}@ovgu.de

***Abstract*—Head pose estimation is not only a crucial pre-processing task in applications such as facial expression and face recognition, but also the core task for many others, e.g. gaze; driver focus of attention; head gesture recognitions. In real scenarios, the fine location and scale of a processed face patch should be consistently and automatically obtained. To this end, we propose a depth-based face spotting technique in which the face is cropped with respect to its depth data, and is modeled by its appearance features. By employing this technique, the localization rate was gained. additionally, by building a head pose estimator on top of it, we achieved more accurate pose estimates and better generalization capability. To estimate the head pose, we exploit Support Vector (SV) regressors to map Histogram of oriented Gradient (HoG) features extracted from the spotted face patches in both depth and RGB images to the head rotation angles. The developed pose estimator compared favorably to state-of-the-art approaches on two challenging D-RGB databases.**

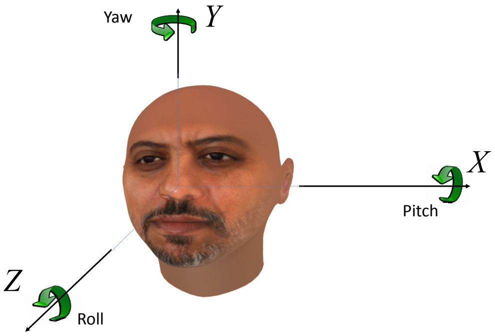

Fig. 1: The head pose rotation angles. Yaw is the rotation around Y-axis, pitch around X-axis, roll around Z-axis [15].

I. Introduction

Head pose estimation is a crucial pre-processing step for several computer vision systems, e.g. face and facial expression recognitions. One challenge for the facial analysis systems is to cope with uncooperative persons, whose faces are in arbitrary in-depth rotations. These variations in pose are larger than inter-person variations, thus impairing the face identification and the facial expression recognition as well. Zhang and Gao [17] provide a review of face recognition across poses in which the pose is estimated as a pre-processing step in many approaches. Niese et al. [12] use the estimated pose to project extracted features, distances between the facial points and optical flows, onto a frontal face and then perform a pose-invariant facial expression recognition. To minimize the intra-class variations of the facial expression according to the head pose, Moore and Bowden [9] recognize the facial expression for discrete groups of poses, where the subject pose is estimated first. On the other hand, head pose estimation is the core task for many other computer vision systems, in which a continuous estimate of the head pose over an image sequence is required. Head gesture recognition is one of these applications [3]. Additionally, gaze detection [5], focus of attention recognition [2] and driver assistance systems [11] are mainly built on top of head pose estimators.

Considering the human head as a rigid object, we express the pose in terms of three rotation angles: pitch; roll; and yaw, as shown in Fig. 1. Several approaches have been proposed to automatically estimate the head pose in frame- and video-based (tracking) modes. In the frame-based mode, the head pose is inferred mostly via machine learning approaches that map various types of feature, extracted from a cropped face patch, onto the rotation angles. By contrast in the tracking mode, the approaches estimate the head pose at the first frame (or assume the tracking starts from zero rotation angles) and then accumulate estimates of the pose difference between the consecutive frames. Fanelli et al. [7] propose a method to detect the head and estimate his pose by applying the discriminative random regression forest method on depth data stemmed from D-RGB sensor (Kinect camera). Yang et al. [16] employ a multi-linear SV classifier to locate the face using the D-RGB data, then estimate the pose using a Multi-Layer Perception (MLP) network. The authors in [15], [14] employ the Viola and Jones (VJ) face detector to locate the face, then utilize several depth- and appearance-based features to infer the head pose. They cope with the inconsistent face cropping of the VJ detector by confining the face search across scales using the depth data [14], and by performing the face search in two stages [15]. Niese et al. [13] apply the **I**terative **C**losest **P**oint (ICP) algorithm to fit current head pose to person-specific head model within a limited range of poses around the previous pose estimate.

Jimenez et al. [8] employ the **RAN**dom **SA**mple **C**onsensus (RANSAC) and the **P**ose from **O**rthography and **S**caling with **IT**erations (POSIT) algorithms to infer the head pose from tracked 2D points.

The appearance-based approaches for the facial analysis rely heavily on the face cropping, where inconsistent cropping shall ruin any further processing. The main contribution of this work is the proposal of a depth-based technique to locate and crop the face. The use of it led to an improvement in the cross-database localization rate by at least 10%. Additionally by developing a pose estimator on top of it, we advanced the estimation accuracy by at least 11.9% for pitch, 1.5% for yaw and roll angle. Finally, the generalization capability of the developed pose estimator was enhanced as well, as proven by a cross-database evaluation where the estimation accuracy is improved by at least 13.5% for pitch, and 9% for yaw.
The remainder of this work is organized as follows. In section 2, a description of the proposed approach is given. In section 3, we present the experimental results. Finally, the conclusion and future perspectives are given in section 4.

II. The Proposed Approach

D-RGB sensors provide depth data along with the 2D RGB image. Obviously, the depth date facilitate the foreground segmentation and the estimation of the unknown object scale as well. Knowing the depth data, we perform only a face search across potential spatial regions in this work, where the scale of each search window is determined with respect to the depth value of its center. The face localization takes place in the RGB image due to the rich and unique texture of the face in the RGB image in comparison to the face 3D structure available from the depth image, as empirically evaluated. Then, features encoding the facial appearance (from RGB image) and the face 3D structure (from depth image) are used to estimate the head pose via SV regressors.

A. Appearance-based features

The human face has its own distinctive appearance. Additionally, it is obviously seen from the cropped face patches in Fig. 2 that the facial appearance and face 3D structure are significantly changing across the head poses, as they were captured using a fixed mounted camera. The aforementioned characteristics can be read from the face patches in both depth and RGB images using a low-level texture-based feature extractor. In this work, we use the HoG features [6] as they were proven to be more effective for the head pose estimation among other types [15]. To extract the HoG descriptor, we divide the face patch into smaller regions named cells. The pixels of each cell are used to build a histogram of their orientation with a predefined number of bins, where each pixel orientation is weighted by its magnitude. The neighbor cells form block, in which their histograms are normalized to be illumination invariant. Finally, we concatenate the blocks' features to generate the patch descriptor.

B. The face localization process

Building a head pose estimator on top of an automatic face detector is necessary for real-world applications. The main shortages of most available face detectors are their inconsistent face cropping across scales and their limitation to a small range of poses. In this work, we propose a method to spot the face in the D-RGB images. To this end, we built a Gaussian Mixture Model (GMM) for the face under pose variations. Exploiting the Biwi database [7], we divided the entire pose range into discrete groups spaced by 5 degrees in each angle (pitch, yaw, roll) direction. Then, for each cube we selected one sample from each subject, if available. Next, we annotated those samples by enclosing each face sample with a box of fixed size in real-world units. This box is centered at the nose tip. Annotation samples of three subjects are shown in Fig. 2. Let B_z denote the distance from the face center to the camera. B_n is the head width in the real-world units that was empirically set to 150 mm. Then using a simple pinhole camera model, we calculated the corresponding head width (B_m) in pixels as follows.

$$B_m = \frac{f_x \times B_n}{B_z}, \tag{1}$$

where f_x is the camera focal length multiplied by the scale parameter in x- direction, and is measured in pixels. Next the faces were cropped, each with the corresponding B_m. Features extracted from those patches were used to build a multivariate GMM face model, where the likelihood of a face feature vector is calculated as follows.

$$p(\mathbf{x}|\Phi) = \sum_{i=1}^{m} \alpha_i p_i(\mathbf{x}|\phi_i), \tag{2}$$

where $\mathbf{x} = (x_1, ..., x_d)' \in \mathbb{R}^{d\times 1}$ is the feature vector encoding the face patch. $\phi_i = (\mu_i, \Sigma_i)$, $\Phi =$ $(\alpha_1, ..., \alpha_m, \phi_1, ..., \phi_m)$ is the face model, which is estimated via the Expectation Maximization (EM) algorithm. Each p_i is a d-dimensional multivariate Gaussian distribution given by

$$\begin{aligned} p_i(\mathbf{x}|\phi_i) \quad &= \quad \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{1/2}} \\ &\quad \times \exp\{-\frac{1}{2}(\mathbf{x}-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\mu_i)\}, \end{aligned} \tag{3}$$

$\mu_i \in \mathbb{R}^{d\times 1}$ is the mean vector of the i^{th} subpopulation; where Σ_i is its $d \times d$ covariance matrix. $\alpha_i \in [0, 1]$ for all i and the α_i's are constrained to sum to one. To spot a face inside an image, we evaluate all potential face locations and the window of $\mathbf{x}^*$ of maximum $p(\mathbf{x}|\Phi)$ is considered the cropped face.

$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x}} p(\mathbf{x}|\Phi). \tag{4}$$

The patch of $\mathbf{x}^*$ is then used for the pose estimation process. A satisfactory correct localization rate was achieved with a face model of five subpopulations. The face patch is scaled to 100 × 100 pixels, with cell size of 20 × 20 pixels, block of 40 × 40 pixels, block spacing stride of 20 pixels and eight bins orientation histogram. Therefore, the final

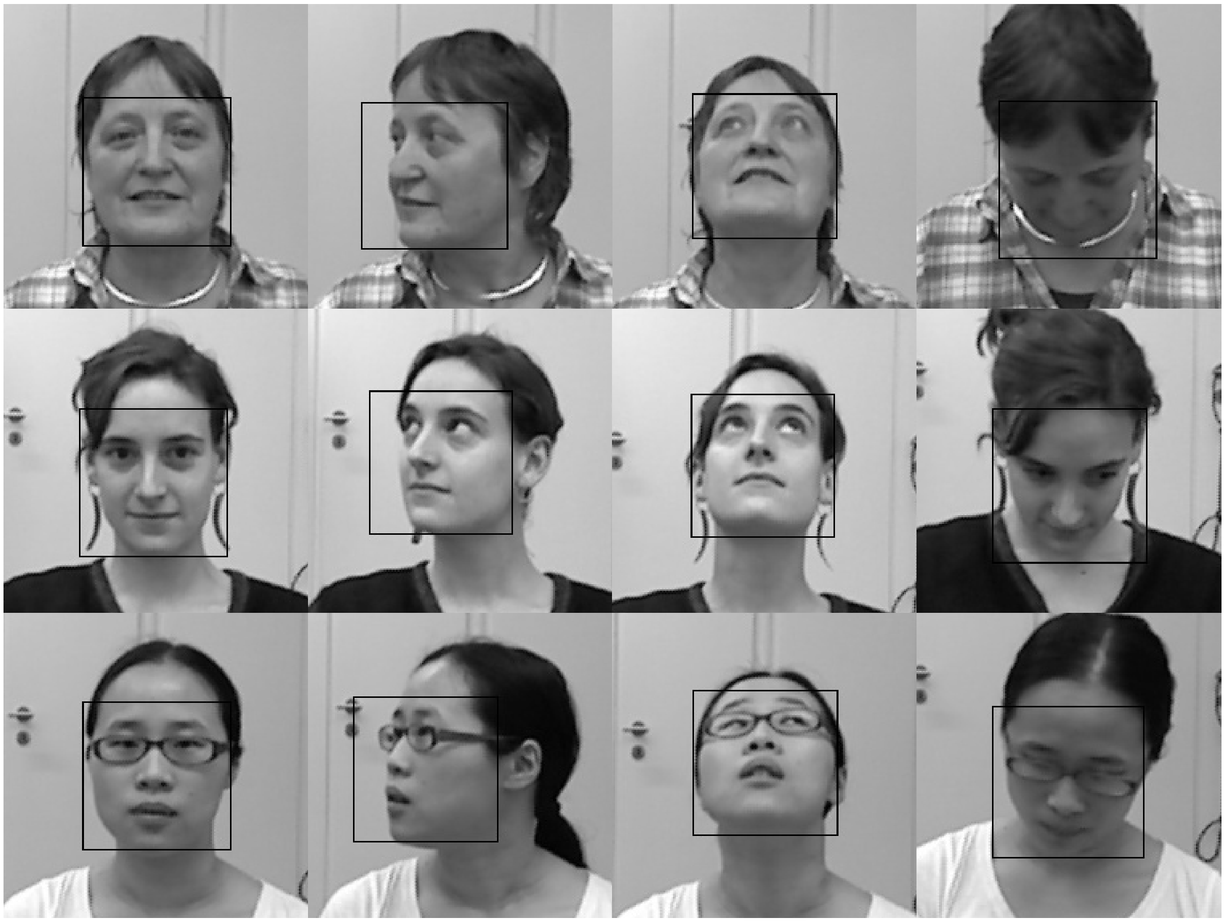

Fig. 2: Samples of our annotations on three subjects, taken from the Biwi database, at different poses.

TABLE I: Localization rates resulting from using different feature sources. HoG_g is HoG features extracted from grayscale image of the color image, HoG_d from the depth image, and HoG_{g+d} from both. This evaluation was conducted on Biwi database.

Feature	**Localization rate** $\%$
HoG_g	98
HoG_d	80
HoG_{g+d}	88

HoG descriptor has a length of 512. By performing a cross-validation evaluation on Biwi database, the face localization rates were 98%, 80%, and 88% of utilizing features from the grayscale image (derived from the RGB image) HoG_g, depth image HoG_d, and both HoG_{g+d}, respectively, as shown in Table I. The face localization is considered correct if the intersection area of the predicted patch and its corresponding ground truth patch is at least 60% of the union of the two. The localization rate using HoG_g is better than that using HoG_d due to the distinctive unique texture of the face in the grayscale image. Additionally, the search with only one scale precludes many false-positive detections. Although the depth-based features show better performance on estimating the head pose [15] [14] [7], they perform poorly in the face localization, where the face pattern is always confused with parts from the body as they look like a face in profile views. The HoG_d vector adversely affects the localization rate when it is concatenated with HoG_g. Consequently for further processes, only HoG_g is employed as a feature vector in Eq. (2), while the patch size is determined based on its depth data. With a goal of maximizing the face localization rate, we set the parameters for the aforementioned process (the number of Gaussian subpopulations, HoG parameters, B_n) through a grid search with cross-validation evaluation conducted on the Biwi database.

C. Head pose estimation

As clearly shown in Fig. 2, the facial appearance and the face 3D structure within the annotated boxes vary according to the head pose. To encode these variations, we extracted HoG features from the cropped face patches, in the RGB and depth images. To this end, each face patch is resized to 160×160 pixels, with cell size of 20×20 pixels, block of 40×40 pixels, block spacing stride of 20 pixels and eight bins orientation histogram. Hence, the final descriptor from one patch is of length 1568. Concatenating both descriptors results in vector (HoG_{g+d}) of length 3136 (2×1568).

The head pose is estimated by employing three SV regressors, each maps the extracted features (HoG_{g+d}) to one angle. SV is well known for its generalization capability and overfitting avoidance in multi-class classification and regression as well [1]. The regression parameters (penalty cost (400), kernel function (radial basis)) along with the HoG parameters (e.g. cell, block, and stride sizes) were optimized using a grid-search with cross-validation experiment conducted on the training data with a goal of maximizing the accuracy of the head pose estimation.

III. Experimental results

To assess the effectiveness and the generalization capability of the proposed head pose estimator, we exploited two databases, publicly available, on which comparisons with state-of-the-art approaches were performed in both within-database and cross-database scenarios.

A. Within-database evaluation

For this evaluation, we utilized the Biwi database [7]. It comprises 24 D-RGB sequences (of 20 people) captured using a Kinect camera. The authors provided frame-based annotations (head pose angles), which were obtained by fitting a personalized 3D head model to the point cloud of the face at each frame using the ICP algorithm. We split the database into two training and testing sets of 18 and 2 subjects, respectively, where samples of the same person were not used in both sets at the same time. Both the face spotter and pose estimator were trained using the training set before being evaluated on the testing set. Finally, we reported the absolute estimation error, averaged over the test sets, in terms of its mean and standard deviation values. Table II summarizes our results in comparison with those of state-of-the-art approaches; the best configuration of each approach was always selected. The participation in the comparison was limited to the approaches that incorporate face localization and work in a frame-based mode as well, except the approach in [10] it was built on top of manual annotations. Saeed et al. [14] employed the VJ face detector whose parameters were adjusted using the depth data. Saeed et al. [15] applied the VJ detector on foreground images, where the background was removed using the depth data. Yang et al. [16] built his own face detector that exploits both RGB and depth information. Fanelli et al. [7] built a face detector based on the depth data only. The aforementioned methods annotate the face by a box exactly enclosing it, while in our method the box is centered at the nose tip and its size is depth-based calculated. Our approach provided a better estimation accuracy, where the major improvement was achieved for the pitch angle by at least 11.9% $((4.19 - 4.76)/4.76)$. The accuracy improvement can be attributed to the employed method for the face localization and cropping. The depth-based cropping is considered as a compromised solution to arrange the faces with a goal of minimizing intra-class variations (of faces with similar poses) while maximizing inter-class variations (of faces with different poses). Clearly shown in Fig. 2 that within a box of fixed real-world size, the variations of the face texture and 3D structure across the head poses are greater than those among individuals of the same pose. Although the size of the projected (observed) face changes across the pitch angles, the cropping size stills fixed leading to more accurate pitch estimates. Fig. 3 shows pose estimates over an image sequence. The estimation error increases when the head is under higher pose angles as the appearance variations are barely visible at those rotation values. Interestingly, it is apparent that our approach is qualified to be used in applications of head gesture recognition.

TABLE II: The mean/standard deviation of the absolute error for each head pose angle. Within-Biwi database evaluation.

Approach	**Pitch Er °**	**Yaw Er °**	**Roll Er °**
Saeed et al.[15]	5.12 / 5.3	4.6 / 4.5	4.2 / 4.1
Saeed et al. [14]	5.0 / 5.8	3.9 / 4.2	4.3 / 4.6
Fanelli et al. [7]	8.5 / 9.9	8.9 / 13.0	7.9 / 8.3
Yang et al. [16]	9.12 / 7.40	8.92 / 8.27	7.42 / 4.9
Mukherjee and Robertson [10]	4.76/-	5.32/-	-/-
Our	4.19 / 4.30	3.84 / 3.9	4.13 / 4.4

B. Cross-database evaluation

An important metric to assess any approach is its performance in a cross-database evaluation. To this end, we evaluated a pre-trained head pose models, provided from Sec. III-A, on the ICT-3DHP database [4]. It encompasses 10 D-RGB sequences (6 male, 4 female) captured using a Kinect camera. Frame-based ground truths were obtained by tracking the head using Polhemus Fastrack flock of birds tracker attached to a cap the participants were wearing. As the annotations are actually the angles difference from the first frame, we considered the estimation of the first frame as a bias. We applied our approach along with those of [7], [15], [14], all were trained using the Biwi database, on each frame of the ICT-3DHP database (approximately 14200). Table III summarizes the localization rates stemmed from the cross-database evaluation. We achieved a localization rate of 95% compared with 82% of [7], 85% of [15] and [14], an improvement by at least 10% proves the better generalization capability of our depth-based method for the face localization. Table IV shows the pose estimation results, where we only considered the frames in which the face was correctly detected by the four approaches. The online available code of Fanelli et al. [7] approach does not estimate the roll angle; therefore, their error estimate of this angle is not presented in Table IV. Our results are superior to those of the state-of-the-art approaches, which emphasizes the better generalization capability of the developed pose estimator as well. Fig. 4 shows pose estimates over a sequence of images taken from the ICT-3DHP database. Similar to the results in Sec. III-A, the estimation error increases only when the face experiences greater rotation angles.

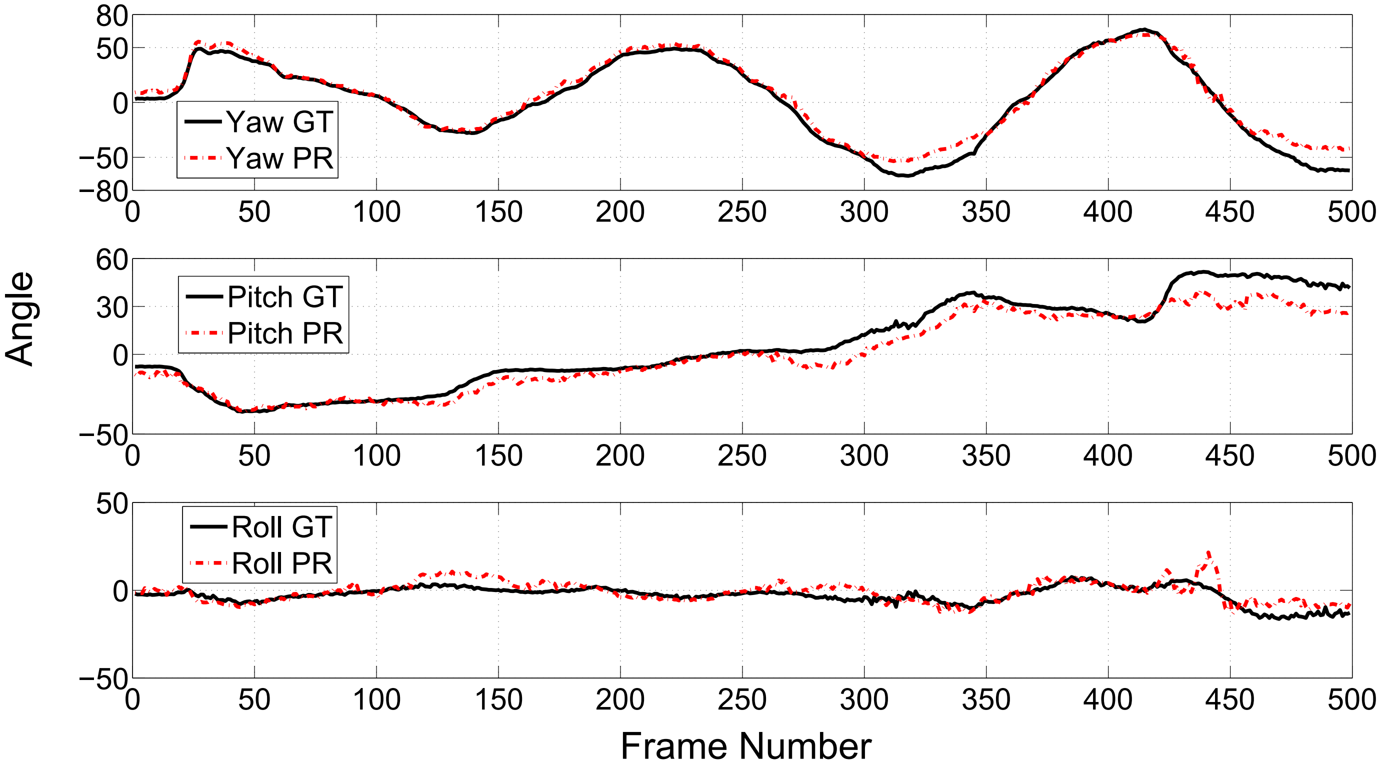

Fig. 3: Samples of the head pose estimation over an image sequence. They were taken from the cross-validation experiment, conducted on the Biwi database. GT denotes the ground truth, PR the predicted angles.

TABLE III: Face localization rates resulting from applying developed face detector on ICT-3DHP database. The detectors in this work and in [7] were completely developed based on Biwi database, while [15], [14] employed the VJ detector whose parameters were set with respect to Biwi database.

Approach	**Localization rate** %
Saeed et al. [15], [14]	85
Fanelli et al. [7]	82
Our	95

TABLE IV: The mean/standard deviation of the absolute error for each ICT-3DHP head pose angle stemmed from the cross-database validation. These head pose estimators were trained on the Biwi database and tested on the ICT-3DHP database.

Approach	**Pitch Er** °	**Yaw Er** °	**Roll Er** °
Saeed et al. [15]	5.32 / 5.7	5.3 / 5.5	4.3 / 4.5
Saeed et al. [14]	4.9 / 5.3	5.1 / 5.4	4.4 / 4.6
Fanelli et al. [7]	5.9 / 6.3	6.3 / 6.9	-
Our	4.23 / 4.41	4.64 / 4.9	4.33 / 4.6

IV. Conclusions and Future Work

In this work, we have presented a depth-based method to locate and crop the face patch. This method is supposed to improve the performance of any further facial analysis. It has been proven that it increases the face localization rate. Implementing a human head pose estimator on top of it led to better estimation accuracy and better generalization capability. The face is represented by HoG features extracted from the RGB data and modeled by a multivariate GMM, where the patch size is calculated with respect to its distance from the camera. After cropping the face, the head pose is estimated by mapping HoG features extracted from the face two patches in D-RGB images to the rotation angles using SV regressors. Our next research step is to do a head gesture recognition on the basis of the pose estimator developed here.

Acknowledgment

This work is supported by Transregional Collaborative Research Centre SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by DFG and OvG-University Magdeburg.

References

[1] Abe, S.: Support Vector Machines for Pattern Classification (Advances in Pattern Recognition). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2010)

[2] Ba, S., Odobez, J.M.: Recognizing visual focus of attention from head pose in natural meetings. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39(1), 16–33 (2009)

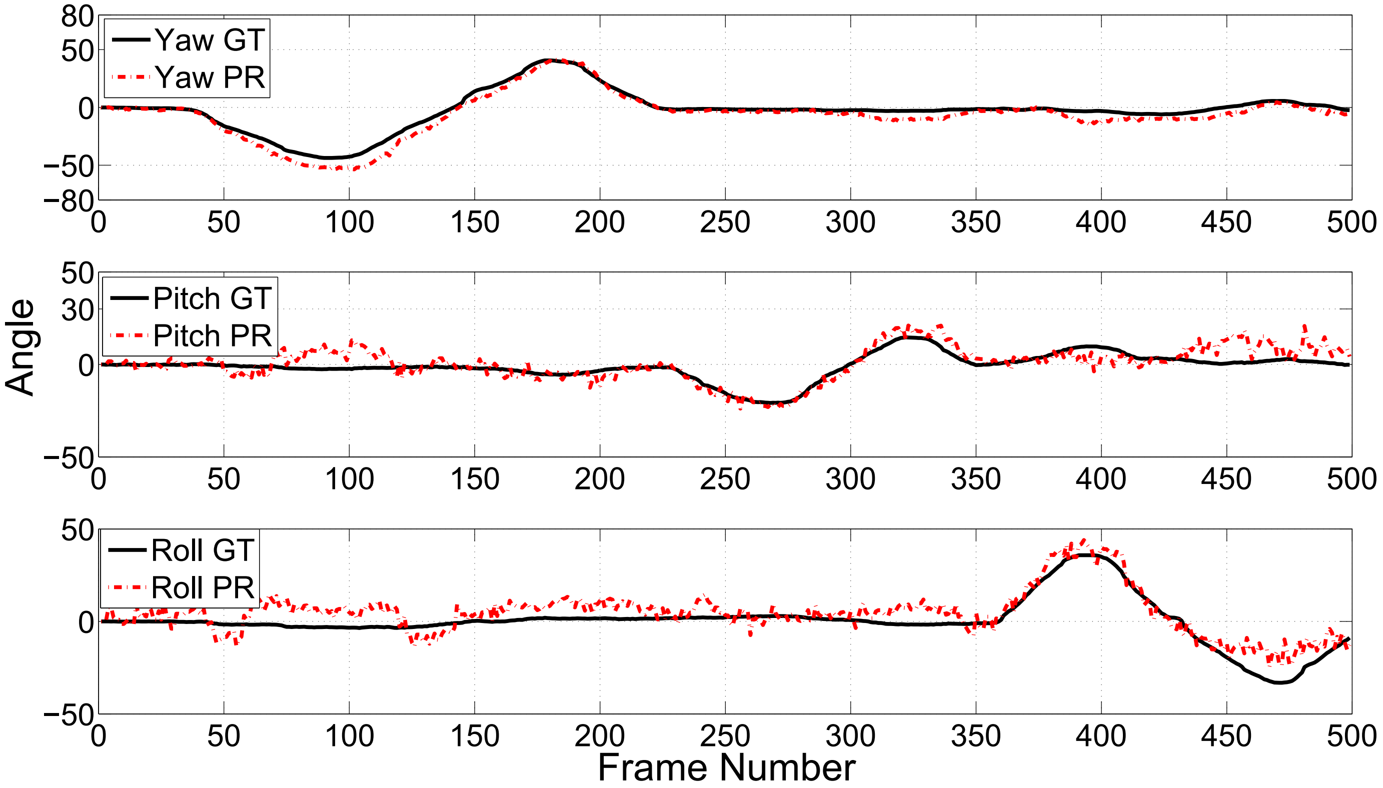

Fig. 4: Samples of the head pose estimation over an image sequence. They were taken from cross-database validation; the pose models were trained using the Biwi database and tested on the ICT-3DHP database. GT denotes the ground truth, PR the predicted angles.

[3] Baltrusaitis, T., McDuff, D., Banda, N., Mahmoud, M., El Kaliouby, R., Robinson, P., Picard, R.: Real-time inference of mental states from facial expressions and upper body gestures. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. pp. 909–914 (March 2011)

[4] Baltrusaitis, T., Robinson, P., Morency, L.: 3d constrained local model for rigid and non-rigid facial tracking. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2610–2617 (June 2012)

[5] Cazzato, D., Leo, M., Distante, C.: An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation. Sensors (Basel, Switzerland) 14(5), 8363–8379 (May 2014)

[6] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893 vol. 1. San Diego, CA, USA (June 2005)

[7] Fanelli, G., Weise, T., Gall, J., Gool, L.V.: Real time head pose estimation from consumer depth cameras. In: Proceedings of the 33rd International Conference on Pattern Recognition. pp. 101–110. DAGM'11, Springer-Verlag, Berlin, Heidelberg (2011)

[8] Jimenez, P., Nuevo, J., Bergasa, L.: Face pose estimation and tracking using automatic 3d model construction. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on. pp. 1–7. Anchorage, AK, USA (June 2008)

[9] Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. Comput. Vis. Image Underst. 115(4), 541–558 (Apr 2011)

[10] Mukherjee, S.S., Robertson, N.M.: Deep head pose: Gaze-direction estimation in multimodal video. IEEE Transactions on Multimedia 17(11), 2094–2107 (Nov 2015)

[11] Murphy-Chutorian, E., Doshi, A., Trivedi, M.: Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In: Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE. pp. 709–714 (Sept 2007)

[12] Niese, R., Al-Hamadi, A., Farag, A., Neumann, H., Michaelis, B.: Facial expression recognition based on geometric and optical flow features in colour image sequences. Computer Vision, IET 6(2), 79 –89 (march 2012)

[13] Niese, R., Werner, P., Al-Hamadi, A.: Accurate, fast and robust realtime face pose estimation using kinect camera. In: 2013 IEEE International Conference on Systems, Man and Cybernetics - SMC. pp. 487–490. Manchester, UK (Oct 2013)

[14] Saeed, A., Al-Hamadi, A.: Boosted human head pose estimation using kinect camera. In: Image Processing (ICIP), 2015 IEEE International Conference on. pp. 1752–1756 (Sept 2015)

[15] Saeed, A., Al-Hamadi, A., Ghoneim, A.: Head pose estimation on top of haar-like face detection: A study using the kinect sensor. Sensors 15(9), 20945–20966 (2015)

[16] Yang, J., Liang, W., Jia, Y.: Face pose estimation with combined 2d and 3d hog features. In: Pattern Recognition (ICPR), 2012 21st International Conference on. pp. 2492–2495. Tsukuba, Japan (November 2012)

[17] Zhang, X., Gao, Y.: Face recognition across pose: A review. Pattern Recognition 42(11), 2876–2896 (Nov 2009)