



Evaluation of Camera Pose Estimation Using Human Head Pose Estimation

Robert Fischer¹ · Michael Hödlmoser² · Margrit Gelautz¹

Received: 21 June 2022 / Accepted: 25 January 2023
© The Author(s) 2023

Abstract

We introduce and evaluate a novel camera pose estimation framework that uses the human head as a calibration object. The proposed method facilitates extrinsic calibration from 2D input images (NIR and/or RGB), while merely relying on the detected human head, without the need for depth information. The approach is applicable to single cameras or multi-camera networks. Our implementation uses a fine-tuned deep learning-based 2D human facial landmark detector to estimate the 3D human head pose by fitting a 3D head model to the detected 2D facial landmarks. Our work focuses on an evaluation of the proposed approach on real multi-camera recordings and synthetic renderings to determine the accuracy of the pose estimation results and their applicability. We assess the robustness of our method against different input parameters, such as varying relative camera positions, variations of head models, face occlusions (by masks, sun glasses, etc.), potential biases and variance among humans. Based on the experimental results, we expect our approach to be effective for numerous use cases including automotive attention monitoring, robotics, VR/AR and other scenarios where ease of handling outweighs accuracy.

Keywords Evaluation · Camera pose estimation · Camera networks · Head pose estimation · Extrinsic calibration

Introduction

Camera pose estimation or extrinsic calibration refers to the calculation of both the translation and rotation of cameras relative to a well-defined coordinate system. For a multi-camera extrinsic calibration, the reference coordinate system is usually the coordinate system defined by one of the cameras. Camera pose estimation [1, 2] needs to be frequently applied when cameras employed in the context of real-world

phenomena, as is the case for virtual/augmented reality applications [3], various automotive applications [4], rapid prototyping [5] and robotics [6]. To estimate the camera pose, some known calibration object is usually used to correlate the recorded pixels with the corresponding transformation [7]. Popular calibration objects are rigid planar surfaces featuring a checkerboard pattern [8] or a circle grid pattern [9]. Problems arise in use cases and scenes where no calibration object is present or none can be conveniently used but the pose of a camera still needs to be estimated. To cope with such cases, we propose our novel camera pose estimation technique which measures the pose of a single camera or multiple cameras using the human head as a reference object for calibration. Figure 1 illustrates a proposed application of our camera pose estimation approach in a car cockpit, in which the camera pose can be estimated via the pose of the driver's head. Our method observes human heads using a single NIR/RGB camera or multiple NIR/RGB time-synchronized cameras. For each of the cameras' 2D projections, our approach extracts 2D landmarks using a deep-learning-based convolutional neural network. By assuming a fixed pre-defined 3D human head model (as shown in Fig. 2), we can use the facial landmarks to estimate the poses of the human heads, as well as the poses of the cameras. Hence,

This article is part of the topical collection “Advances on Computer Vision, Imaging and Computer Graphics Theory and Applications” guest edited by Kadi Bouatouch, Augusto Sousa, Mounia Ziat and Helen Purchase.

✉ Robert Fischer
robert.fischer@tuwien.ac.at

Michael Hödlmoser
mho@emotion3d.ai

Margrit Gelautz
margrit.gelautz@tuwien.ac.at

¹ Visual Computing and Human-Centered Technology, TU Wien, Favoritenstrasse 9, 1040 Vienna, Austria

² emotion3D GmbH, Rainergasse 1/8, 1040 Vienna, Austria

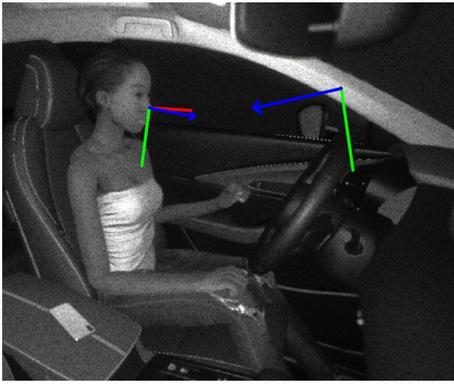


Fig. 1 Application of proposed camera pose estimation method. The left 3-axes represent the XYZ directions of the estimated head pose. The right XYZ-axes represent the XYZ directions of the calculated camera pose

our proposed approach is especially suited for camera setups where human heads are common, such as settings within the cabin of a vehicle, plane or train, in which the passengers are observed by the cameras. We evaluate the usefulness and accuracy of our method by performing extensive experiments using a multi-camera setup and synthetic renderings.

There is a serious need for our method in applications where ease of calibration has priority over high calibration accuracy. Examples of such use cases include region-based attention monitoring [4], robot attention tracking [6], automated shopping systems [10] or rapid-prototyping of new camera arrangements. For such systems, it is often impractical to enforce the calibration of cameras by users with an additional calibration target before usage. In contrast, our approach is applicable ad-hoc for a variety of extrinsic camera calibration tasks. It is worth noting that extrinsic calibration of cameras is more fragile than its intrinsic counterpart, as stronger vibrations, winds or other external forces might invalidate the camera's extrinsics. Hence, being able to calibrate a camera extrinsically without requiring dedicated calibration objects is particularly desirable in such situations, as said object might not always be readily available. Furthermore, the defined 3D head model allows our method to be independent of additional depth information. Instead, our multi-camera pose estimation technique merely depends on RGB and/or NIR 2D input data.

The main differences with respect to our previous conference paper [11] are: In addition to the synthetic data used in [11], we created a larger dataset for validation, showing six different participants. Each participant was recorded under different potentially challenging modalities, including facial occlusions (masks, sun glasses), grimacing, different head rotations and three camera positions using the Opti-Track Motive camera system [12] (“Evaluation” section). To improve camera pose estimation accuracy, we additionally

implemented a camera pose aggregation over multiple frames, which has been shown to significantly improve accuracy in “Aggregation Experiment” section and Table 5. In the current work, we also fine-tuned and optimized a Stacked Hourglass 2D facial landmarks detection model [13] and compared it with various state-of-the-art alternatives (“2D Facial Landmark Detection” section and Table 2). We performed the experiments from [11] again using the optimized Stacked Hourglass 2D facial landmarks detection model and the extended dataset (“Core Experiment”, “Checkerboard Experiment”, “Impact of Relative Camera Poses”, “Bias Towards Skin Color and Gender”, “Qualitative Evaluation” section and Tables 3, 4, 6, 9). The newly added dataset allowed us to design and perform additional experiments concerning the impact of grimacing (“Impact of Head Model Divergence” section and Table 7), the impact of face occlusions (“Impact of Face Occlusions” section and Table 8) and the variance among individual participants (“Variance Among Individual Participants” section and Table 10). We extended “Related Work” section to cover additional related work regarding various 2D facial landmarks models and non-standard calibration methods.

Related Work

Due to its high practical relevance in computer vision, robotics and augmented/virtual reality, multi-camera pose estimation has become a relatively intensively studied research topic [1, 2]. A well-established technique of camera pose estimation was proposed by Zhang [14]. The technique extracts the poses from a plane with an unknown orientation. Later, Zhang [8] extended his technique to use a planar surface with squares, resembling a checkerboard pattern. Further, Abad et al. [9] added to the planar projection board approach by implementing the detection of concentric circles instead of squares. Adnan et al. [15] proposed a method using both point and line correspondences for camera pose estimation. Later, Manolis et al. [16] applied a model-based approach to this problem by using a rigid 3D model instead of a planar calibration object. Other non-planar 3D calibration objects include wands with multiple collinear points [17] or point-like objects [18]. In more recent years, Camposeco et al. [19] proposed to leverage both structure-less (2D-3D) and structure-based (2D-2D) correlations. Finally, Noll et al. [20] presented a comprehensive overview of PnP-based, RANSAC-based and other more traditional camera pose estimation techniques. In summary, performing extrinsic camera calibration while exclusively relying on the human head as a calibration target is a novel approach to the problem of multi-camera pose estimation.

Our method relies on the accurate estimation of the human head pose in 3D from a 2D input image [21].

Before the advent of deep-learning-based computer vision, head pose estimation was often performed by using manifolds [22–24]. Many proposed methods [25–30] also used random forests with RGB and depth images. Deep-learning-based approaches have since largely replaced the traditional computer-vision-based methods for both RGB and depth input images [31–38]. For in-car automotive applications, convolutional neural networks (CNNs) in combination with depth cameras have shown to be successful [31]. Regressing the head pose directly using CNNs has also proven to be effective [32]. Wu et al. [33] combined face detection and head pose estimation within the same network. Xia et al. [39] built an efficient multitask architecture for face alignment, head pose estimation and face tracking. Deep-learning-based head pose estimation often suffers from too little training data - synthetic data has shown to be an effective solution to this problem [34]. Furthermore, Patacchiola et al. [35] investigated multiple CNN architectures for head pose estimation. Recently, Liu et al. [37] proposed to handle pose errors in the ground truth distribution and to leverage asymmetric relation cues by assigning different weights on the yaw- and pitch rotation direction. Moreover, graph CNNs have been applied successfully to head pose estimation from 3D point clouds [36]. The visual transformer architecture (ViT) [40] has also shown to be promising for many computer vision tasks including head pose estimation. Hu et al. [41] introduced a spatiotemporal vision transformer enabled head tracker. Dhingra [42] proposed an architecture for head pose estimation consisting of a mixture of depthwise separable convolutional layers and transformer encoder.

Oftentimes convolutional neural networks are applied for facial landmarks extraction, which is commonly part of a head pose estimation model [34, 35]. In general, CNNs consist of two main components: the feature extractor and the discriminator. In the following, we provide a selection of relevant state-of-the-art feature extractors for 2D facial landmarks extraction. HRNet [43] maintains high-resolution representations throughout the whole feature extraction process by gradually adding high-to-low-resolution convolution streams into the network. MobileNet [44] is tailored for embedded applications focused on leveraging depthwise convolutions. Stacked Hourglass networks [13] process features across multiple scales and have been shown to capture spatial relationships relatively well. The ResNet [45] architecture extends deep neural networks with skip connections and enables much deeper neural networks. SCNet [46] introduces the use of self-calibrating convolutions. Continuing with the success of ViT architectures [40], Li et al. [38] exploit temporal contexts using strided transformers. Our approach uses an optimized model derived from a Stacked Hourglass architecture [13]. Our evaluation in “2D Facial Landmark Detection” section shows that in our context the

fine-tuned model performs better than off-the-shelf 2D facial landmarks detectors.

The need for multi-camera pose estimation in the absence of a dedicated calibration object is widespread in computer vision. Bleser et al. [47] propose to use a model created with CAD software to estimate the camera pose. Rodrigues et al. [48] attain the camera pose by detecting planar mirror reflections in the scene. Hödlmoser et al. utilize pedestrians on zebra crossings to calculate the camera poses [49]. The proposed methods by Puwein et al. [50], Takashi et al. [51] and Moliner et al. [52] are similar to our approach and usually expect the full human body to be detectable by the cameras. While the aforementioned methods are suited for camera setups in surveillance or studio-like environments, they are not readily applicable to use cases where primarily the human head is visible (and other body parts are possibly occluded). In contrast to our method, those techniques rely on finding pixel correspondences using the entire human body pose. Instead, our approach exploits merely the human head pose estimated independently from each camera without requiring pixel correspondences between different camera images.

Camera Pose Estimation

In the following subsections, we describe the entire calibration pipeline of our multi-camera pose estimation technique using head pose estimation. Each synchronized camera within the camera network captures an image of the current scene, on which we then perform human head pose estimation. After that, we derive the camera poses from the computed head pose transformations, allowing us to transform the shared coordinate system into the respective camera coordinate systems and vice versa.

General Workflow

Camera pose estimation is a common problem in computer vision and can be time-consuming to perform. Traditionally, it is first necessary to physically prepare and print some calibration object, for example an adequate checkerboard, and then validate that the calibration object satisfies certain conditions, such as being rigid and unbendable. Normally, the calibration process must then be carried out, in particular the parametrization of the calibration algorithm is required. Finally, the camera pose estimation itself can be performed by capturing the calibration data from the calibration object, which needs to be visible to all cameras. In contrast to this procedure, our approach only requires a single person to be present in the scene to calculate the head-pose-based camera pose estimation. This condition is inherently met in many use cases. From the computational point of view, estimating the head pose is more complex than finding a checkerboard

Table 1 Runtime of head pose estimation on ARM Cortex A57 (2.035 GHz) per camera (From [11])

No. Cores:	1	2	3	4
Runtime	25.1 ms	14.9 ms	12.1 ms	9.9 ms

pattern in the recorded images. We counteract this problem by running the head pose estimation algorithm on a graphical processing unit (GPU), resulting in comparable execution times. Table 1 gives an overview of the runtimes on different numbers of cores. It shows that the computation time needed for head pose estimation is approximately inversely proportional to the number of cores.

Camera Pose Estimation With Human Head Pose Estimation

Our method assumes camera intrinsics to be available. The first step of our camera pose estimation technique consists of extracting the human head pose. In principle, any head pose estimator that returns a proper translation and orientation for a human head can be used for our method. For our implementation, we focus on facial-landmarks-based head pose estimation. Our approach is particularly motivated by driver cabin applications (e.g. cars, trains, planes), where the assumption that the driver's face is visible to the camera usually holds. After detecting the face with an off-the-shelf face detector, we apply a CNN-based deep learning facial landmarks detector on the recorded and cropped 2D image from each camera. More traditional approaches not based on deep learning are applicable as well, but are not state-of-the-art anymore [21]. In "2D Facial Landmark Detection" section, we perform a comparison of various facial landmark detectors. We evaluate our custom trained fine-tuned Stacked Hourglass [13] based model in comparison to several off-the-shelf models and observe that the custom model outperforms the off-the-shelf models. Our modifications and re-training of the 2D facial landmarks detector can be summarized as follows: We fine-tuned the hyperparameters (learning rate, data augmentation, regularization, etc.) of the Stacked Hourglass [13] model and trained it on the COCO 2014 dataset [53] with grayscale pre-processing. The prediction framework of Stacked Hourglasses by Newell et al. [13] provides a flexible approach for training image-dependent spatially-aware models, while supporting long-range feature dependencies for the detected facial landmarks. In the original publication, the authors applied their model architecture for general human pose estimation but as we will show in our evaluation in "Evaluation" section, it can also be fine-tuned for the task of camera pose estimation. We increased the accuracy of 2D facial landmark detection by achieving a

more accurate detection of the bounding box of the detected faces by using a CNN-based off-the-shelf face detector.

Using the extracted 2D facial landmarks, we then fit an independently measured static 3D head model. We use iterative Perspective-n-Point (PnP) to fit the estimated 2D points to the assumed 3D head model [54]. Figure 2 visualizes the correspondence of the 2D facial landmarks and their corresponding 3D head model points. We define the origin of the shared coordinate system (Refer SCS in Fig. 3) as the tip of the nose of the 3D head model (refer to 3D head model illustration in Fig. 2), but other locations are equally applicable and could be chosen instead. Next, we use the estimated head pose to calculate the transformation from the head coordinate system into the camera coordinate system. The accuracy of this transformation mainly depends on (1) the quality of the facial landmarks and (2) the similarity between the actual head and the presumed 3D head model. A key factor for (1) are occlusions. More occluded 2D facial landmarks result in fewer points to use for the PnP-step, leading to less optimal estimations. For example, in the presence of rotations or occluding objects (hair, sunglasses, etc.) noses are visible more often than ears. In "Impact of Face Occlusions" section, we find that the trained Stacked Hourglass [13] model can cope relatively well with such scenarios and estimates a reasonable location for occluded facial landmarks. Regarding (2), we assume a pre-defined 3D head model, so if the actual head shape diverges significantly from the assumed model, the PnP-step might find a suboptimal solution, which has negative impact on the camera pose estimation. Nonetheless, our experiments show empirically that a generic head model is applicable to a wide scope of different head types.

In the following, we discuss how to construct the multi-camera network. Figure 3 visualizes the multi-camera setups with the corresponding coordinate systems and transformations. Figure 3a depicts a general camera network of our approach and Fig. 3b shows the camera network for the

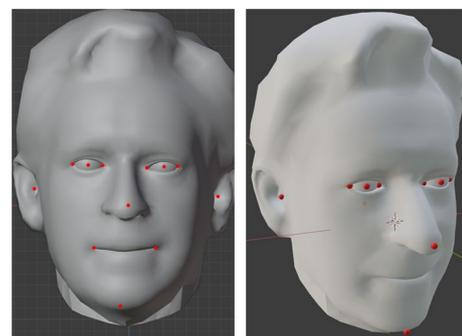
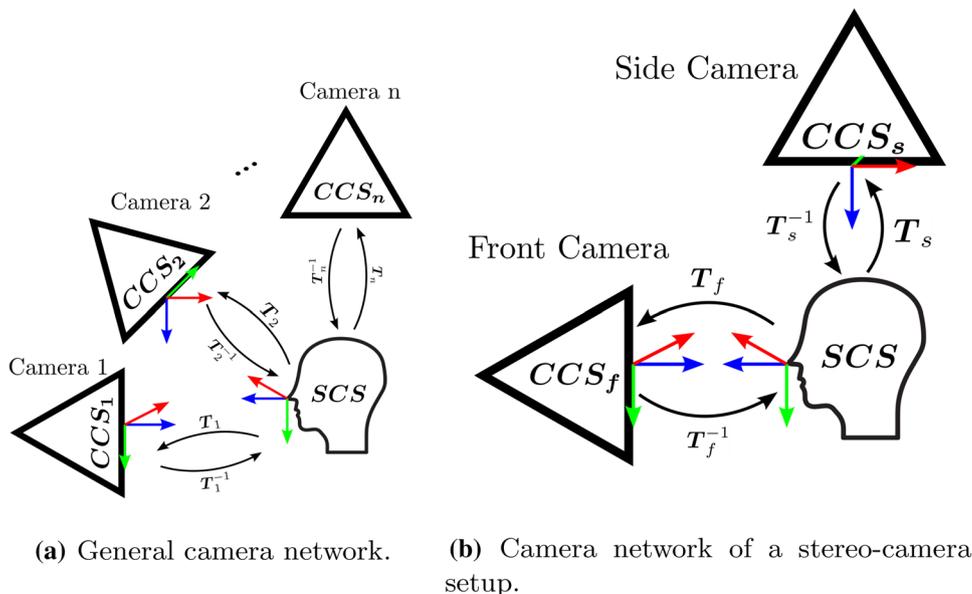


Fig. 2 Correspondence of 2D facial landmarks and the 3D head model. Left: 2D facial landmarks on 3D head. Right: 3D head model fitted according to facial landmarks. (From [11])

Fig. 3 Schematic visualization of camera networks supported by our camera pose estimation method with the corresponding camera coordinate systems (CCS), a shared coordinate system (SCS), and the corresponding transformations between them. (From [11])



stereo-camera setup we use for experiments in “Evaluation” section. Limitations of the camera network are primarily of physical nature and depend on the specific installed hardware. In particular, limitations stem from camera synchronization, placement conditions and bandwidth of the network. Our test setup consists of three time-synchronized cameras and is capable of capturing full HD resolution images while simultaneously performing camera pose estimation.

A head pose consists of a 3x3 rotation matrix R and a 1x3 translation matrix t and transforms from a camera coordinate system’s origin to the head coordinate system’s origin. We construct a corresponding 4x3 transformation matrix T as defined in Eq. 1. Equation 2 defines the inverse transformation from the head coordinate system’s origin into a camera coordinate system’s origin.

$$T = [R | t] \tag{1}$$

$$T^{-1} = [R^T | -R^T * t] \tag{2}$$

We denote a camera as CCS_i and the shared coordinate system defined by the head pose as SCS . We perform the transformation from the shared coordinate system SCS into CCS_i by constructing T_i using the estimated head pose defined by the rotation matrix R_i and translation matrix t_i for camera i . Conversely, in order to transform from CCS_i into SCS , we perform the inverse transformation T_i^{-1} .

To apply the estimated camera pose, we show how to transform both points and rotations within the camera

network’s camera coordinate systems into the shared coordinate system.

$$t_{SCS} = T_i^{-1} * t_i \tag{3}$$

$$R_{SCS} = rot(T_i)^{-1} * R_i * rot(T_i) \tag{4}$$

$$rot([R | t]) = R \tag{5}$$

Equation 3 defines the transformation of an arbitrary translation t_i from the coordinate system of camera i CCS_i into the shared coordinate system defined by the head pose SCS , resulting in the transformed translation t_{SCS} . It can be seen that we can transform from the SCS into CCS_i by applying the transformation T_i . Equation 4 shows how to transform an arbitrary rotation R_i from the coordinate system of the camera i CCS_i into the shared coordinate system defined by the head pose SCS , resulting in the transformed rotation R_{SCS} . The function $rot(T)$ returns the 3x3 rotation matrix R of the transformation T (see Eq. 5).

We propose a relatively simple way to improve the camera pose estimation accuracy by exploiting temporal coherence typically present in video sequences. More precisely, we aggregate multiple head poses over time into the transformation matrix T . This technique reduces the impact of head pose estimation outliers. In our current implementation, we choose to calculate the mean over all translations and rotations. Calculating such mean over rotations is an instance of the single rotation averaging problem [55], which we address by applying the technique of the Geodesic L_2 mean [55]. In “Aggregation Experiment” section, we show that aggregation performs well but with the added constraint that multiple image frames are required for camera pose estimation.

Evaluation

In this section, we investigate the performance of our novel camera pose estimation method from various points of view. We define and illustrate the setup we use throughout our evaluation in ‘[Experiment Setups](#)’ section. In ‘[2D Facial Landmark Detection](#)’ section, we draw comparisons between several state-of-the-art 2D facial landmark detectors and our custom fine-tuned and optimized model. In the core experiment presented in ‘[Core Experiment](#)’ section, we explore and compare the overall accuracy of our camera pose estimation technique using a dataset containing a sample of all our recordings featuring six participants, filmed from three different camera poses (0°, 45° and 90°), including face occlusions, head rotations (pitch, yaw and roll) and grimaces. In ‘[Checkerboard Experiment](#)’ section, we perform a baseline comparison with camera pose estimation using a checkerboard pattern. In ‘[Aggregation Experiment](#)’ section, we evaluate the effects of aggregating multiple head poses over time on the estimation of the camera poses. In ‘[Impact of Relative Camera Poses](#)’ section, we compare how different camera poses impact the estimation accuracy. In ‘[Impact of Head Model Divergence](#)’ section, we evaluate the impact of head model divergence by comparing the results of the generic head model against the true head model using synthetic data. In the same section, we also analyze how grimaces impact estimation accuracy. We analyze the impact of face occlusions caused by masks, sunglasses and manual drop-out in ‘[Impact of Face Occlusions](#)’ section. We then investigate potential biases towards gender or skin color using real and synthetic data in ‘[Bias Towards Skin Color and Gender](#)’ section. We further perform an analysis focusing on the variance of our camera pose estimation results derived from individual participants in ‘[Variance Among Individual Participants](#)’ section. Finally, we provide additional qualitative evaluation results in ‘[Qualitative Evaluation](#)’ section.

Experiment Setups

We performed most experiments using real near infrared (NIR) cameras. Additionally, some experiments were performed or augmented using rendered images of simulated NIR cameras. To distinguish between real and synthetic data, we explicitly denote the use of synthetic data in our experiments. Using a mixture of mainly real and some synthetic images allows us to evaluate our method from several viewpoints, which would not be possible with just a single data source. NIR cameras are often used in cockpit-like environments. In these scenarios it is normally

possible to emit near-infrared light from a custom light source, resulting in NIR cameras not being dependent on external illumination. Our method is not limited to NIR images, as our 2D facial landmark extraction model is capable of handling color RGB images as well. To evaluate the accuracy of the estimated camera poses, we need ground truth camera poses. For the synthetic data, we compute the ground truth camera poses from the scene graph of the render engine. For the real data, we use the Opti-Track Motive camera system [12]. According to the manufacturer, this camera system returns the ground truth camera pose with less than 0.2 mm error.

Our experiment setup consists of a stereo camera inside a cabin-like environment mimicking a car. Each experiment is set up with a front and a side camera. Refer to Fig. 3b for an illustration of the camera setup. Assuming the person sits neutrally straight and is looking forward, the front camera is placed roughly one meter in front of that person. The pose of the side camera varies between the experiments. For most experiments, the camera is positioned roughly 90 degrees to the right relative to the person. In some of the experiments, we test our approach with other camera positions as well. Using the estimated pose and the ground truth, we establish evaluation metrics that facilitate intuitive comparison and are designed to reflect the accuracy of our camera pose estimation method. To this end, we evaluate the error of transformation between two cameras within the camera network. In ‘[Camera Pose Estimation](#)’ section, we describe how to insert new cameras into the camera network. We perform the camera pose estimation independently for each camera. Hence, there is no positive nor negative impact of adding more cameras. We split the transformation evaluation metric into its translational and rotational components.

Given a point in the shared coordinate system SCS , we transform the point p_{SCS} into CCS_1 to p_1 and CCS_2 to p_2 using corresponding ground truth camera pose data. Then we transform the point p_1 using the estimated camera pose T_1^{-1} for camera 1 into SCS resulting in $p_{SCS \text{ from } 1}$. Analogously, we transform p_2 using the estimated camera pose T_2^{-1} for camera 2 into SCS resulting in $p_{SCS \text{ from } 2}$. If the estimated camera poses match the ground truth camera poses exactly, $p_{SCS \text{ from } 1} = p_{SCS \text{ from } 2} = p_{SCS}$ holds, meaning that both points transform to the same position in the shared coordinate system SCS . Comparing the two transformed points $p_{SCS \text{ from } 1}$ and $p_{SCS \text{ from } 2}$ with each other allows us to measure the degree of inaccuracy introduced by the camera pose transformation. We then compare the mean Euclidean distance (Eq. 6) of the two points.

$$L_2(p_{SCS \text{ from } 1} - p_{SCS \text{ from } 2}) \quad (6)$$

Similarly, for the rotation errors, given a rotation in the shared coordinate system SCS , we transform the rotation

R_{SCS} into CCS_1 to R_1 and CCS_2 to R_2 using ground truth camera pose data. Then we transform the rotation R_1 using the estimated camera pose T_1^{-1} for camera 1 into SCS resulting in $R_{SCS \text{ from } 1}$. Analogously, we transform R_2 using the estimated camera pose T_2^{-1} for camera 2 into SCS resulting in $R_{SCS \text{ from } 2}$. As in the previous point transformation, if the estimated camera poses match the ground truth camera pose exactly, $R_{SCS \text{ from } 1} = R_{SCS \text{ from } 2} = R_{SCS}$ holds, meaning that the rotations transform to the same rotation in the shared coordinate system SCS . Afterwards, we convert the rotation matrices into pitch, yaw and roll Euler angles in degrees, as they are intuitive to understand for humans. We then calculate the mean absolute circle difference of all Euler angles [56].

Our method is tailored for applications where quick and easy camera pose estimation without the need for a dedicated calibration object is more important than higher estimation accuracy. In these applications, the acceptable accuracy trade-off depends on the actual use case. For example, for non-safety critical attention monitoring within car cockpits, we consider a mean Euler difference of below 15 degrees and a mean distance of up to 20 cm to be acceptable.

2D Facial Landmark Detection

In this section, we explore the impact of various state-of-the-art facial landmark detectors and compare them to our fine-tuned Stacked Hourglass [13] model. We chose various pre-trained state-of-the-art 2D facial landmarks detector models, trained with different publicly available datasets. The models explored are HRNet v2 [43], MobileNet v2 [44], Stacked Hourglass [13], ResNet [45] and SCNet [46]. The datasets selected for training are 300W [57], AFLW [58], COCO-WholeBody-Face [59] and WFLW [60]. Some models use DarkPose [61] or Adaptive Wingloss [62]. We sample 1500 images from the core experiment validation dataset

(refer to “Core Experiment” section) for this facial landmark detection evaluation. Table 2 shows the evaluated pre-trained models and compares them to our Stacked Hourglass [13] based fine-tuned model. We find that our Stacked Hourglass approach outperforms the pre-trained models in each of the applied evaluation metrics. Hence, we perform all further experiments using the fine-tuned model.

Core Experiment

The core experiment evaluates the overall performance of our camera pose estimation method by randomly sampling from every recording available of our participants. Additionally, we also compare results of our method being applied on various synthetic 3D renderings. The synthetic data is described in more detail in “Experiment Setups” section. It consists of six 3D models of different humans turning their head by 90 degrees (see Fig. 7) captured from various camera angles (refer to Fig. 5) within a simulated car cockpit. We call this experiment “Synthetic experiment”. In Table 3 we see the results of our camera pose estimation technique for the core- and synthetic experiment. The core experiment contains some challenging scenarios including six different people, occluded faces, grimaces and relatively extreme camera angles in relation to the recorded persons’ faces. Refer to Fig. 4 for an overview of the different modalities contained within our dataset. Nonetheless, the mean distance is 18 cm and the mean Euler difference is 5.17° , which indicates that our approach for camera pose estimation using human head pose estimation is of sufficient accuracy for selected use cases where ease of calibration outweighs estimation accuracy. For the synthetic experiment, we observe a translational error of 10 cm and a rotational error of 5.12° . The aggregation experiment in “Aggregation Experiment” section shows that there are circumstances

Table 2 Comparison of 2D facial landmark detection models

Model	Mean Dist. [m]	Mean Euler [deg]	Std. Dist. [m]	Std. Euler [deg]
Stacked Hourglass + COCO WholeBody Face	0.37	13.90	0.48	12.13
HRNet v2 + WFLW + DarkPose	1.12	41.05	0.43	20.30
HRNet v2 + COCO WholeBody Face	0.37	13.93	0.43	12.33
HRNet v2 + WFLW + Adaptive Wingloss	1.13	40.71	0.46	20.55
HRNet v2 + WFLW	1.11	40.77	0.41	20.47
HRNet v2 + AFLW	0.75	23.42	1.91	26.50
HRNet v2 + 300W	0.68	21.46	1.88	25.02
HRNet v2 + COCO WholeBody Face + DarkPose	0.36	13.74	0.32	11.85
SCNet + COCO WholeBody Face	0.36	13.67	0.31	12.03
ResNet50 + COCO WholeBody Face	0.36	13.74	0.41	12.19
MobileNet v2 + COCO WholeBody Face	0.36	13.93	0.28	12.06
Our (Stacked Hourglass + COCO 14)	0.13	5.02	0.20	7.06

Fig. 4 Overview of dataset used for evaluation

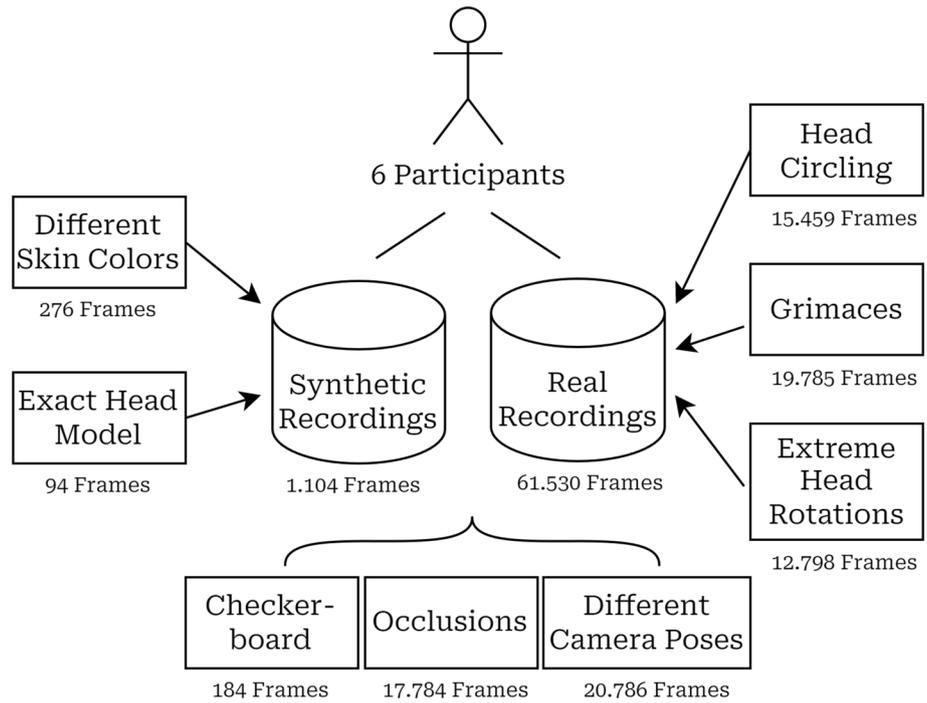


Table 3 Comparison of core experiment and experiment with synthetic data

Experiment	Mean	Mean	Std.	Std.
	Dist.	Euler	Dist.	Euler
	[m]	[deg]	[m]	[deg]
Core experiment	0.18	5.17	0.97	9.96
Synthetic experiment	0.10	5.23	0.11	8.78

Table 4 Comparison of our camera pose estimation approach with the checkerboard pattern approach

Experiment	Mean	Mean	Std.	Std.
	Dist.	Euler	Dist.	Euler
	[m]	[deg]	[m]	[deg]
Our approach (Neutral)	0.10	2.81	0.05	2.12
Checkerboard	0.02	0.17	0.01	0.04

in which our approach performs even more closely to the checkerboard-based approach in terms of accuracy. In summary, the observed accuracy is below the thresholds we consider acceptable (see “[Experiment Setups](#)” section). The impact of the previously mentioned challenging scenarios on the accuracy will be explored in more detail in the following sections.

Checkerboard Experiment

In this subsection, we compare our camera pose estimation technique with the traditional method introduced by Zhang [8], which we implement using a rigid planar surface imprinted with a checkerboard pattern. To the best of our knowledge, there is no pre-existing method for camera pose estimation using a human head which could be compared to our approach in a meaningful way. Thus, we selected the following approach to establish a baseline for accuracy.

We asked each of the participants to assume a neutral sitting posture and turn their head from the neutral forward

position to looking towards their right side. We call this motion “Neutral”. In each recorded frame, the person moves their head slightly towards the final head rotation. Afterwards, we capture a similar motion, but this time we replace the human head with a checkerboard which we rotate from facing forward to facing 90 degrees to the right. As the motion and camera setup are essentially the same, we can compare the accuracy of these two approaches meaningfully. We calculate the metrics described in “[Experiment Setups](#)” section for both the head-pose-based and the checkerboard-based approach. In Table 4, we can see the mean distance of the checkerboard-based camera pose estimation, which is 2 cm, and the mean Euler angle, which is 0.01 degrees, as well as the mean distance and mean Euler angle for our approach, which are 10 cm and 2.81 degrees, respectively. When comparing the results, it is important to note that the checkerboard method requires a special calibration object to be visible to all cameras, whereas our approach takes advantage of human faces, which are usually omnipresent in our envisioned applications.

Aggregation Experiment

In this subsection, we explore ways to further improve the estimation accuracy by exploiting temporal coherence present in video sequences by averaging over multiple frames. Our camera pose estimation framework has shown to estimate poses with relatively low translational and rotational errors using the “Neutral” recordings from “[Checkerboard Experiment](#)” section. Next, we aggregate by calculating the mean translation and rotation (as described “[Camera Pose Estimation](#)” section) for the “Neutral” recordings. For the purpose of comparability, we do the same for the images of the core- and synthetic experiment (refer to “[Core Experiment](#)” section). In [Table 5](#), we see considerable improvements in scenarios with aggregated neutral head poses. The accuracy in such cases is approximately 3 cm mean Euclidean distance and 1.33° mean Euler angle difference. In each of the performed experiments, we observe an improvement when comparing the aggregated results with their non-aggregated counterparts. For further comparison with a well-known baseline, we also provide results for camera pose estimation using a checkerboard pattern. We observe that the translational error difference is 1 cm and the rotational error difference is slightly above 1 degree. In summary, we find that aggregation improves our pose estimation results consistently.

Impact of Relative Camera Poses

Camera pose estimation should be robust against a wide range of relative camera poses in relation to the calibration object. For this experiment, we created a dataset of several recordings with different relative rotations between the human face and the recording cameras. We placed three NIR cameras into the scene: one directly in the front, another approximately 45° to the right side and the last one approximately 90° to the right side of the recorded participants. To increase the variance of relative rotations between the

Table 5 Comparing the technique of camera pose aggregation by using a single frame for each estimation of the camera pose

Experiment	Mean	Mean	Std.	Std.
	Dist.	Euler	Dist	Euler
	[m]	[deg]	[m]	[deg]
Neutral, aggregation	0.03	1.33	0.01	0.44
Neutral, no aggregation	0.10	2.81	0.05	2.12
Core experiment, aggregation	0.11	4.24	0.11	4.05
Core experiment, no aggregation	0.18	5.17	0.97	9.96
Synthetic experiment, aggregation	0.08	3.32	0.06	2.83
Synthetic experiment, no aggregation	0.10	5.23	0.11	8.78
Checkerboard	0.02	0.17	0.01	0.04

cameras and the human heads within the dataset camera poses, we asked the test subjects to rotate their heads around the pitch, yaw and roll axes, separately. In addition, we use a synthetic dataset containing renderings of a car cockpit from different camera poses relative to the driver. [Figure 5](#) indicates the relative positions of the cameras within our synthetic dataset. The results in [Table 6](#) show that the camera pose estimation from different relative camera poses is relatively stable. The core experiment’s translational error is 18 cm, while the errors from different camera poses are 11 cm and 12 cm, respectively. Similarly, we obtain a mean Euler angle difference of 5.17° for the core experiment and 3.69° and 3.91° for the rotational error metrics of the different camera poses. We observe similar results for the synthetic case, apart from an outlier for the “Synthetic near relative camera pose” configuration. In this case, we found that the 2D facial landmarks detector returned inaccurate 2D locations, which might be caused by missing training data of faces in close proximity. In summary, our method has proven robust against a wide range of relative rotations, but small

Table 6 Estimation accuracy resulting from different relative camera poses. Synthetic camera poses correspond with the camera positions of [Fig. 5](#): Camera 1 is used for the camera pose with label *Side*, camera 2 for *Far Side* and camera 3 for *Near*

Experiment	Mean	Mean	Std.	Std.
	Dist.	Euler	Dist.	Euler
	[m]	[deg]	[m]	[deg]
Relative camera pose 45°	0.11	3.69	0.12	5.21
Relative camera pose 90°	0.12	3.91	0.11	4.90
Synthetic <i>Far Side</i> relative camera pose	0.11	3.60	0.08	4.17
Synthetic <i>Near</i> relative camera pose	0.31	19.61	0.19	20.68
Synthetic <i>Side</i> relative camera pose	0.06	2.69	0.08	3.77
Core experiment	0.18	5.17	0.97	9.96
Synthetic experiment	0.10	5.23	0.11	8.78

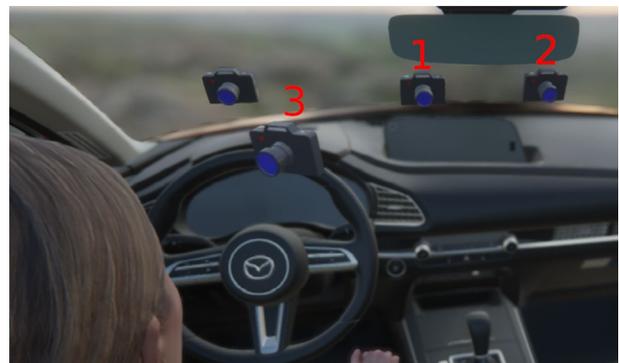


Fig. 5 Visualization of the virtual camera positions for the synthetic camera poses experiments. Camera numbers are used for reference within [Table 6](#)

distances between the camera and the observed human face may require special consideration.

Impact of Head Model Divergence

Our camera pose estimation approach relies on iterative Perspective-n-Point (PnP) [54]. Hence, our method assumes a static pre-defined head model, as described in “Camera Pose Estimation” section. Refer to Fig. 2 for a visualization of the correspondence between the 2D facial landmarks and the 3D head model. In this subsection, we explore to what degree the static head model impacts the estimation accuracy and how different facial expressions, especially in the form of grimaces, affect the performance of camera pose estimation. We investigate the impact of the static pre-defined head model by evaluating the performance using synthetically rendered 3D data. We can access the true 3D head model via the 3D mesh of the human we use for rendering the scene. Hence, we can directly measure the impact of diverging 3D head models. In the synthetic experiments, we see improvements for the translational error of 1 cm, and 0.72 degrees for the rotational error, if we use the exact head model instead of the generic default head model. We consider the observed slightly negative impact of diverging head models an acceptable trade-off for our camera pose estimation technique.

Further, related to head model divergence, we compare recordings of grimaces with recordings where the participants show a neutral facial expression while turning their head 90 degrees to their side. Refer to Fig. 6 for a selection of facial expressions contained within the grimacing experiment. Table 7 contains the evaluation results of this experiment, and the estimation accuracy of several modalities relating to facial expressions are compared. We observe that facial expressions impact the mean Euler difference significantly with a difference relative to the neutral baseline of 1.2 degrees for the stationary case and a mean Euler angle difference of over 1.8 degrees for the non-stationary turning case. The translation is similar when comparing the baseline with the grimace recordings. For our approach, it seems to be more difficult to estimate the translation if participants make grimaces while also turning their heads. In such cases, we observe a translational error of 5 cm. Future research could elaborate on that observation and customize the assumed

Table 7 Comparison of head model divergence experiments

Experiment	Mean	Mean	Std.	Std.
	Dist.	Euler	Dist.	Euler
	[m]	[deg]	[m]	[deg]
Synthetic default head model	0.09	4.14	0.08	4.48
Synthetic exact head model	0.08	3.42	0.05	2.65
Stationary Grimacing	0.10	4.01	0.04	3.10
Grimacing and Turning	0.15	4.67	0.56	6.80
Baseline (Neutral)	0.10	2.81	0.05	2.12

head model or use camera pose estimation techniques that do not rely on such a model. Nonetheless, we interpret these results as acceptable for the previously proposed use cases of our camera pose estimation method.

Impact of Face Occlusions

In this subsection, we evaluate how our camera pose estimation method handles face occlusions. For this purpose, we created three validation datasets containing real NIR images. The first validation dataset of this experiment consists of recordings where the test subjects’ mouths are occluded by FFP2 masks. The second validation dataset contains recordings where the participants wear eye-occluding sunglasses. Finally, we perform several experiments where we deliberately ignore the detection of certain facial regions, to identify the facial regions that are more significant for accurate camera pose estimation than others. In particular, we selectively exclude the following facial regions: left and right ear, left and right eye, left and right side of the face, mouth and nose. The experiment contains recordings of all six participants from three different camera angles (90°, 45° and 0°) relative to the test subject. Refer to Table 8 for detailed experiment results. Masks are a considerable challenge for our approach, as the mean distance increases to over 0.5 m and the mean Euler angle is also relatively high. Occlusions of the entire left or right side of the faces also affect the camera pose estimation negatively. Contrarily, occlusions of the eyes, ears, mouth and nose seem to have a relatively low impact on the estimation accuracy. These experiments indicate that our approach is stable against the majority of

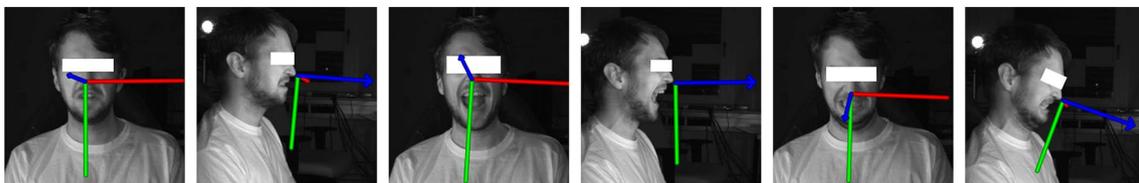


Fig. 6 Various face expressions of the grimacing experiment

Table 8 Comparison of the impact of several occlusion modalities

Occlusion Experiment	Mean	Mean	Std	Std.
	Dist	Euler	Dist	Euler
	[m]	[deg]	[m]	[deg]
Mask	0.52	11.85	2.58	21.86
Sun Glasses	0.14	4.61	0.21	7.12
Left Ear	0.16	5.60	0.18	7.94
Right Ear	0.19	6.16	0.23	7.03
Left Eye	0.15	5.32	0.21	7.24
Right Eye	0.15	5.34	0.21	8.08
Left Side	0.33	11.45	0.26	11.31
Right Side	0.26	8.07	0.29	9.68
Mouth	0.15	5.39	0.28	8.70
Nose	0.15	5.52	0.18	6.79
Core experiment	0.18	5.17	0.97	9.96

tested occlusion types but there is room for improvement regarding occlusion handling in certain cases.

Bias Towards Skin Color and Gender

Deep neural networks are known to have certain biases [63]. In this subsection, we explore the potential presence of bias towards skin color and gender. We compare the estimation accuracy of all recordings of participants that identify as female and male. Additionally, we perform the same analysis for the synthetic data. Figure 7 visualizes the 3D models we use for rendering the synthetic data. Within the real dataset, there is only a single type of skin color present. However, the synthetic validation dataset contains recordings of humans with different skin colors. We do not observe any significant bias towards gender or skin color in Table 9. When comparing the *Real female* and *Real male* recordings, the difference in translational error is 1 cm, while the rotational error measured in mean Euler difference is 0.08 degrees. The skin color experiments *Synthetic lighter skin color* and *Synthetic darker skin color* do not demonstrate a noticeable bias. In this case, the translational error difference is 1 cm and the

Table 9 Comparison of potential bias towards gender or skin-color

Bias Experiment	Mean	Mean	Std	Std.
	Dist	Euler	Dist	Euler
	[m]	[deg]	[m]	[deg]
Real female	0.12	4.73	0.06	6.47
Real male	0.11	4.65	0.10	7.29
Synthetic female	0.10	4.49	0.07	4.18
Synthetic male	0.07	3.59	0.09	4.66
Synthetic lighter skin color	0.09	4.78	0.07	4.54
Synthetic darker skin color	0.10	4.26	0.08	4.67

rotational error difference is approximately 0.5 degrees. For the synthetic cases *Synthetic female* and *Synthetic male*, we observe a more noticeable difference of 3 cm for the translational error and around 1 degree difference in rotation. These error differences are put into perspective by the respective standard deviations of up to 9 cm for the translational error and up to 4.66 degrees for the rotational error.

Variance Among Individual Participants

Another important factor we want to investigate is the variance of estimation accuracy across different people. We explore this question by analyzing the estimation results for each recorded participant. The setup was the same for all participants as in the other experiments. During the recording, the participants were instructed to perform several actions: turning their heads, grimacing, wearing masks and wearing sunglasses. Table 10 shows that, in general, our method is relatively stable across different participants. The translational error ranges usually from 10 cm to 14 cm, the rotational error from 3.52° to 5.76° . However, participant 3 and, to a certain degree, participant 6 seem to be outliers. After analyzing their recordings, we conclude that this observation is most likely explained by their performed grimaces. The recordings show exceptionally pronounced facial expressions, which lead to a significant decrease in estimation accuracy. Nevertheless, for almost all participants the

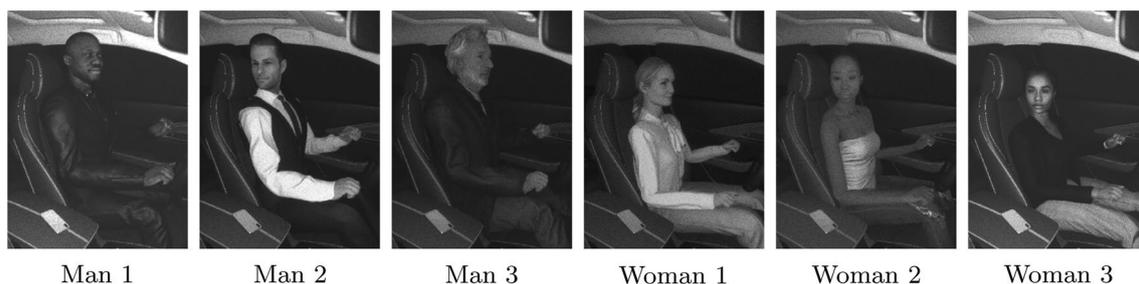


Fig. 7 Rendering of the six different 3D models we use for data generation within our synthetic data generation pipeline. We try to cover a broad range of different appearances of humans and their facial landmarks with these models. (From [11])

Table 10 Comparison of our camera pose estimation method across different participants

Participant Experiment	Mean	Mean	Std	Std.
	Dist [m]	Euler [deg]	Dist [m]	Euler [deg]
Participant 1	0.14	4.64	0.18	6.70
Participant 2	0.14	5.76	0.09	4.68
Participant 3	0.25	5.98	0.70	12.59
Participant 4	0.10	3.52	0.05	2.63
Participant 5	0.11	4.65	0.10	7.29
Participant 6	0.19	6.09	0.45	13.52

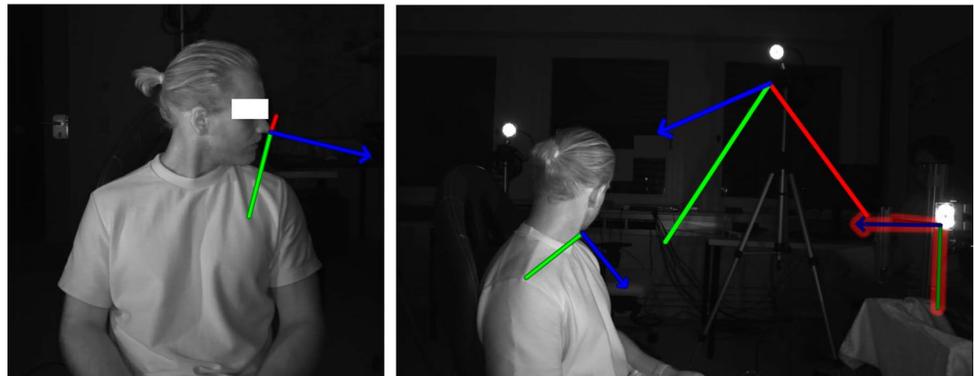
thresholds we defined in “Experiment Setups” section are not exceeded.

Qualitative Evaluation

In this subsection, we explore the accuracy of our camera pose estimation technique qualitatively. We visually selected two result sets from the core experiment outlined in “Core Experiment” section. One set contains examples of accurately estimated camera poses and the other set contains

inaccurate examples. We evaluate our method based on the visually perceived overlapping of the estimated and ground truth camera pose transformation. The more the 2D projected transformation axes visibly overlap the ground truth 2D projected transformation axes, the better we consider the estimated pose to be. Figure 8 and Fig. 9 visualize the pose estimation results of cameras from 0° and 90°, respectively. We notice in Figs. 9a and 9b that even uncommon head rotations can result in accurate camera pose estimations. This observation matches our quantitative results from “Impact of Relative Camera Poses” section, where we also did not notice any significant degradation of estimation accuracy as a consequence of different relative camera angles. Figure 9c shows that our camera pose estimation method handles face occlusions in certain situations relatively well. Contrarily, there are also situations where occlusions of the face can lead to inaccurate camera pose estimations. In Fig. 8a, the head pose estimation fails due to face occlusions. In Fig. 8b, we see how face masks may negatively impact camera pose estimation accuracy. These observations also match the insights of “Impact of Face Occlusions” section. In conclusion, our camera pose estimation method has shown to provide qualitatively satisfactory results and the observations match with our quantitative results of previous sections.

Fig. 8 Examples of inaccurate camera pose estimations. XYZ-Axes located in front of the participants’ noses represent their estimated head pose. Highlighted XYZ-axes correspond to the ground truth camera pose. Remaining XYZ-axes show the estimated camera pose

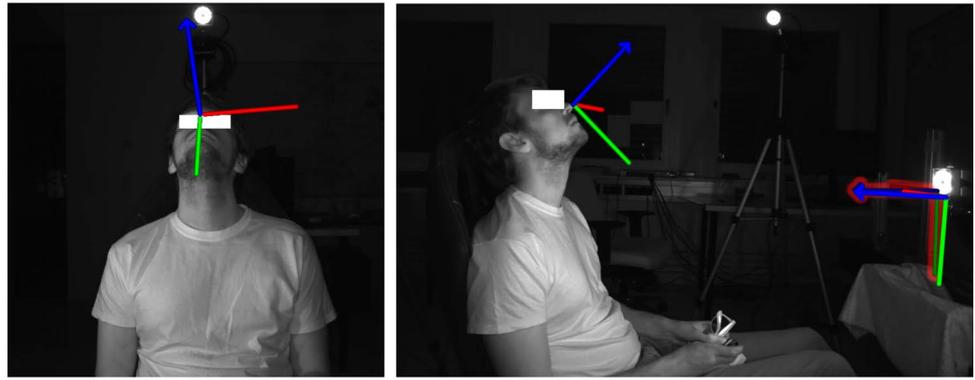


(a) Translational Error: 80.27cm, Rotational Error: 36.14°

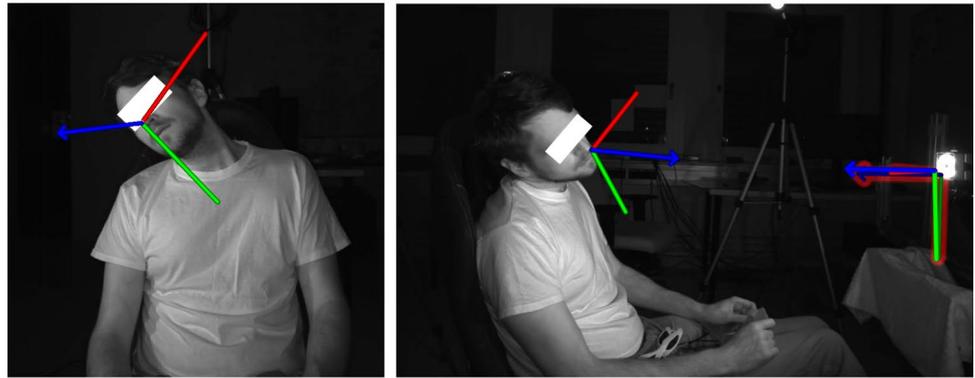


(b) Translational Error: 38.25cm, Rotational Error: 9.04°

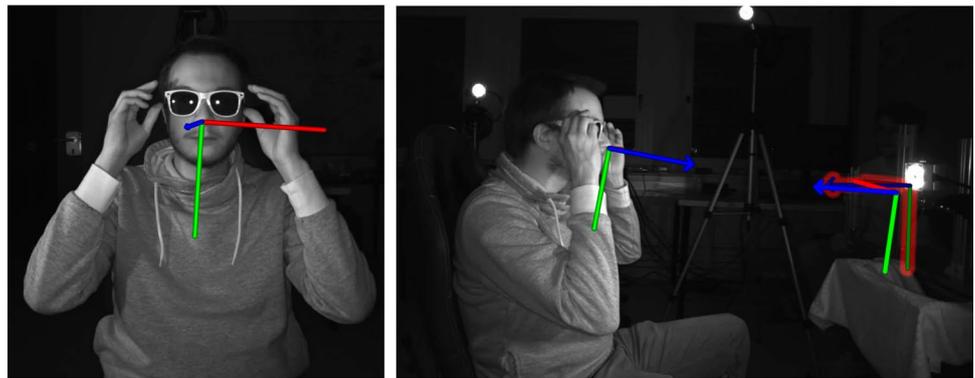
Fig. 9 Examples of accurate camera pose estimations. XYZ-Axes located in front of the participants' noses represent their estimated head pose. Highlighted XYZ-axes correspond to the ground truth camera pose. Further XYZ-axes show the estimated camera pose



(a) Translational Error: 4.46cm, Rotational Error: 3.35°



(b) Translational Error: 5.97cm, Rotational Error: 1.73°



(c) Translational Error: 3.20cm, Rotational Error: 3.33°

Conclusion and Future Work

In our paper, we have established and evaluated the usefulness of a novel application of human head pose estimation for single- and multi-camera pose estimation. We determine the accuracy of our calibration approach by performing a broad range of experiments on real multi-camera recordings and synthetic renderings. We present the results of using various state-of-the-art 2D facial landmark detectors, and a comparison with a conventional

checkerboard-based method. We determine the impact of aggregation over time, the robustness against different input parameters, such as varying relative camera positions, variations of head models, and face occlusions. We further explore potential biases and variance among people. Our analysis on real multi-camera data shows an average aggregated translational and rotational error of around 3 cm and less than 1.33 degrees, respectively. Our proposed approach is tailored to scenarios where reduced estimation accuracy in comparison to more traditional

approaches, is acceptable in favor of easy, natural and flexible handling of head-pose-based camera pose estimation. Use cases where such a trade-off may be desirable include human analysis (e.g. attention monitoring) within the cabin of a vehicle, plane or train, where the passengers are the focus of the cameras. Other situations where our method is applicable include customer interest monitoring in automated stores, robot attention tracking or rapid prototyping.

Subsequent research could generalize our method to also estimate camera intrinsics. The 2D facial landmarks and symmetries typically present in human faces can be potentially leveraged to estimate the camera intrinsics. Combining both the extrinsics from the camera pose estimation and the intrinsics would result in full camera calibration from human faces. At the moment, our method estimates the head pose by detecting 2D facial landmarks. Future work might investigate solutions for settings in which the human face is significantly occluded due to more extreme viewing conditions or when no face is detectable. Our approach to head pose estimation aggregation over time might be further explored. To improve the handling of head pose estimation outliers, the camera pose estimation could be modified by weighting each head pose with the inverse of the mean projection error.

Acknowledgements This work was partly supported by the SyntheticCabin project (no. 884336), which is funded through the Austrian Research Promotion Agency (FFG) on behalf of the Austrian Ministry of Climate Action (BMK) via its Mobility of the Future funding program, and the European Union's H2020 Fast Track to Innovation project SmartRCS (no. 971619).

Author Contributions All authors contributed to the conception of the paper, the design of the evaluation study and the interpretation of the results. The concept of the calibration method was developed by MH and RF. RF performed the implementation of the camera pose estimation algorithm and experiments. The first draft of the manuscript was written by RF, and all authors substantially revised the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by TU Wien (TUW).

Availability of Data and Materials Not available.

Declarations

Conflict of interest Not applicable.

Ethical Approval Not applicable.

Consent to Participate Participants provided consent for their data to be used for academic research (and hence publication).

Consent for Publication Every participant explicitly and knowingly consented to publication of their recorded images. Additionally, we anonymized human faces in the images.

Code Availability Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Pajdla T, Hlavác V. Camera calibration and euclidean reconstruction from known observer translations. In: Proc. of CVPR. 1998; pp. 421–426
- Xu Y, Li Y-J, Weng X, Kitani K. Wide-baseline multi-camera calibration using person re-identification. In: Proc. of CVPR. 2021; pp. 13134–13143
- Fuhrmann A, Schmalstieg D, Purgathofer W. Practical calibration procedures for augmented reality. In: Proc. of virtual environments. 2000; pp. 3–12
- Lamia A, Moshiri HM. Vision-based driver's attention monitoring system for smart vehicles. In: Intelligent computing & optimization. Cham: Springer; 2019. p. 196–209.
- Mefenza M, Yonga F, Saldanha LB, Bobda C, Velipassalar S. A framework for rapid prototyping of embedded vision applications. In: Conference on design and architectures for signal and image processing. 2014; pp. 1–8
- Stiefelshagen R, Yang J, Waibel A. Tracking focus of attention for human-robot communication. In: Proc. of IEEE-RAS international conference on humanoid robots, p. (2001)
- Gua J, Deboeverie F, Mn Slembrouck, Dv Haerenborgh, Dv Cauwelaert, Veelaert P, Philips W. Extrinsic calibration of camera networks using a sphere. *Sensors*. 2015;15(8):18985–9005.
- Zhang Z. A flexible new technique for camera calibration. *TPAMI*. 2000;22(11):1330–4.
- Abad F, Camahort E, Vivó R. Camera calibration using two concentric circles. In: Proc. of ICIAR. 2004; pp. 688–696
- Gross R. How the Amazon go store's AI works. *Towards data science* (<https://bit.ly/3tVHXi2>). 2021
- Fischer R, Hödlmoser M, Gelautz M. Camera pose estimation using human head pose estimation. In: Proc. of International joint conference on computer vision, imaging and computer graphics theory and applications. 2022; pp. 877–886
- NaturalPoint Inc.: OptiTrack motive camera system. <https://optitrack.com/software/motive/>. Accessed: 2021-07-26 (2021)
- Newell A, Yang K, Deng J.: Stacked hourglass networks for human pose estimation. In: Proc. of ECCV, pp. 483–499 (2016)
- Zhang Z. Flexible camera calibration by viewing a plane from unknown orientations. In: Proc. of ICCV. 1999; vol. 1, pp. 666–673
- Ansar A, Daniilidis K. Linear pose estimation from points or lines. In: Proc. of ECCV. 2002; pp. 282–296
- Manolis L, Xenophon Z. Model-based pose estimation for rigid objects. In: Computer vision systems. Berlin: Springer; 2013. p. 83–92.
- Zhang Z. Camera calibration with one-dimensional objects. *TPAMI*. 2004;26(7):892–9.
- Svoboda T, Martinec D, Pajdla T. A convenient multicamera self-calibration for virtual environments. *Presence*. 2005;14(4):407–22.

19. Camposeco F, Cohen A, Pollefeys M, Sattler T. Hybrid camera pose estimation. In: Proc. of CVPR. 2018; pp. 136–144
20. Nöll T, Pagani A, Stricker D. Markerless camera pose estimation - an overview. In: Proc. of VLUDS. 2010; vol. 19, pp. 45–54
21. Shao X, Qiang Z, Lin H, Dong Y, Wang X. A survey of head pose estimation methods. In: Proc. of iThings, GreenCom, CPS Com, SmartData, Cybermatics. 2020; pp. 787–796
22. Chen L, Zhang L, Hu Y, Li M, Zhang H. Head pose estimation using fisher manifold learning. In: Proc. of AMFG. 2003; pp. 203–207
23. Balasubramanian V, Nallure Ye J, Panchanathan S. Biased manifold embedding: A framework for person-independent head pose estimation. In: Proc. of CVPR. 2007; pp. 1–7
24. Raychev B, Yoda I, Sakaue K. Head pose estimation by nonlinear manifold learning. In: Proc. of ICPR. 2004; vol. 4, pp. 462–466
25. Qiao T-z, Dai S. Fast head pose estimation using depth data. In: Proc. of CISP. 2013; pp. 664–669
26. Fanelli G, Gall J, Gool Lv. Real time head pose estimation with random regression forests. In: Proc. of CVPR. 2011; pp. 617–624
27. Fanelli G, Dantone M, Gall J, Fossati A, Gool Lv. Random forests for real time 3D face analysis. IJCV. 2013;101(3):437–58.
28. Huang C, Ding X, Fang C. Head pose estimation based on random forests for multiclass classification. In: Proc. of ICPR. 2010; pp. 934–937
29. Valle R, Buenaposada J, Valdés A, Baumela L. Head-pose estimation in-the-wild using a random forest. In: Proc. of AMDO. 2016; pp. 24–33
30. Li Y, Wang S, Ding X. Person-independent head pose estimation based on random forest regression. In: Proc. of ICIP. 2010; pp. 1521–1524
31. Venturelli M, Borghi G, Vezzani R, Cucchiara R. Deep head pose estimation from depth data for in-car automotive applications. In: Proc. of ICPRW. 2016; pp. 74–85
32. Ruiz N, Chong E, Rehg JM. Fine-grained head pose estimation without keypoints. In: Proc. of CVPRW. 2018; pp. 2155–215509
33. Wu H, Zhang K, Tian G. Simultaneous face detection and pose estimation using convolutional neural network cascade. IEEE Access. 2018;6:49563–75.
34. Liu X, Liang W, Wang Y, Li S, Pei M. 3D head pose estimation with convolutional neural network trained on synthetic images. In: Proc. of ICIP. 2016; pp. 1289–1293
35. Patacchiola M, Cangelosi A. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recog. 2017;71:132–43.
36. Xu Y, Jung C, Chang Y. Head pose estimation using deep neural networks and 3d point clouds. Pattern Recog. 2021;121: 108210.
37. Liu H, Liu T, Zhang Z, Yang AKSB, Li Y. Arhpe: Asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction. IEEE Trans Ind Inform. 2022;18(10):7107–17.
38. Li W, Liu H, Ding R, Liu M, Wang P, Yang W. Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multimedia. 2021
39. Xia J, Zhang H, Wen S, Yang S, Xu M. An efficient multitask neural network for face alignment, head pose estimation and face tracking. Expert Syst Appl. 2022;205: 117368.
40. Ling X, Wang D, Yang J. A new head pose estimation method using vision transformer model. In: Proc. of ICCAI. 2021; pp. 154–159
41. Hu Z, Zhang Y, Xing Y, Zhao Y, Cao D, Lv C. Toward human-centered automated driving: a novel spatiotemporal vision transformer-enabled head tracker. IEEE Vehicular Technol Magazine. 2022;17(4):57–64.
42. Dhingra N. Lwposr: Lightweight efficient fine grained head pose estimation. In: Proc. of WACV. 2022; pp. 1495–1505
43. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: Proc. of CVPR. 2019; pp. 5686–5696
44. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. In: Proc. of CVPR. 2018; pp. 4510–4520
45. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of CVPR. 2016; pp. 770–778
46. Liu J-J, Hou Q, Cheng M-M, Wang C, Feng J. Improving convolutional networks with self-calibrated convolutions. In: Proc. of CVPR. 2020; pp. 10096–10105
47. Bleser G, Wuest H, Stricker D. Online camera pose estimation in partially known and dynamic scenes. In: Proc. of ISMAR. 2006; pp. 56–65
48. Rodrigues R, Barreto J, Nunes U. Camera pose estimation using images of planar mirror reflections. In: Proc. of ECCV. 2010; pp. 382–395
49. Hödlmoser M, Micusik B, Kampel M. Camera auto-calibration using pedestrians and zebra-crossings. In: Proc. of ICCVW. 2011; pp. 1697–1704
50. Puwein J, Ballan L, Ziegler R, Pollefeys M. Joint camera pose estimation and 3D human pose estimation in a multi-camera setup. In: Proc. of ACCV. 2014; pp. 473–487
51. Kosuke T, Dan M, Mariko I, Hideaki K. Human pose as calibration pattern: 3D human pose estimation with multiple unsynchronized and uncalibrated cameras. In: Proc. of CVPRW. 2018; pp. 1856–18567
52. Moliner O, Huang S, Åström K. Better prior knowledge improves human-pose-based extrinsic camera calibration. In: Proc. of ICPR. 2020; pp. 4758–4765
53. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P. Microsoft COCO: Common objects in context. In: Proc. of ECCV. 2014; pp. 740–755
54. Lu C-P, Hager GD, Mjolsness E. Fast and globally convergent pose estimation from video images. TPAMI. 2000;22(6):610–22.
55. Hartley RI, Trunpf J, Dai Y, Li H. Rotation averaging. Int J Comput Vis. 2012;103:267–305.
56. Murphy-Chutorian E, Trivedi MM. Head pose estimation in computer vision: a survey. TPAMI. 2009;31(4):607–26.
57. Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 faces in-the-wild challenge: database and results. Image Vis Comput. 2016;47:3–18.
58. Koestinger M, Wohlhart P, Roth PM, Bischof H. Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: Proc. of ICCV-W. 2011; pp. 2144–2151
59. Jin S, Xu L, Xu J, Wang C, Liu W, Qian C, Ouyang W, Luo P. Whole-body human pose estimation in the wild. In: Proc. of ECCV. 2020; pp. 196–214
60. Wu W, Qian C, Yang S, Wang Q, Cai Y, Zhou Q. Look at boundary: a boundary-aware face alignment algorithm. In: Proc. of CVPR. 2018; pp. 2129–2138
61. Zhang F, Zhu X, Dai H, Ye M, Zhu C. Distribution-aware coordinate representation for human pose estimation. In: Proc. of CVPR. 2020; pp. 7093–7102
62. Wang X, Bo L, Fuxin L. Adaptive wing loss for robust face alignment via heatmap regression. In: Proc. of ICCV. 2019; pp. 6971–6981
63. Mehrabi N, Morstatter F, Saxena NA, Lerman K, Galstyan AG. A survey on bias and fairness in machine learning. ACM Comput Survys. 2021;54:1–35.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.