Camper’s Plane Localization and Head Pose Estimation Based on Multi-View RGBD Sensors

HUAQIANG WANG\textsuperscript{1,2,3}, LU HUANG\textsuperscript{1,2,3}, KANG YU\textsuperscript{1,3}, TINGTING SONG\textsuperscript{1,3}, FENGEN YUAN\textsuperscript{1,2,3}, HAO YANG\textsuperscript{1,2,3}, AND HAIYING ZHANG\textsuperscript{1,2,3}

\textsuperscript{1}Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China
\textsuperscript{2}School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China
\textsuperscript{3}Beijing Key Laboratory for Next Generation RF Communication Chip Technology, Beijing 100029, China

Corresponding author: Hao Yang (yanghao@ime.ac.cn)

This work was supported in part by the Key Research Program of the Chinese Academy of Sciences Foundation under Grant ZDRW-ZS-2021-1.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Xiyuan Hospital, China Academy of Chinese Medical Science.

\section*{ABSTRACT}
Head pose estimation (HPE) is a key step in computation and quantification of 3D facial features and has a significant impact on the precision and accuracy of measurements. High-precision HPE is the basis for standardized facial data collection and analysis. The Camper’s plane is the standard (baseline) plane commonly used by anthropologists for head and face research, but there is no research on automatic positioning of the Camper’s plane using color and depth cameras. This paper presents a high-accuracy method for Camper’s plane localization and HPE based on multi-view RGBD depth sensors. The 3D facial point clouds acquired by the multi-view RGBD depth sensors are aligned to obtain a complete 3D face. Keypoint RCNN is used for facial keypoint detection to obtain facial landmarks. A method is proposed to build a general face datum model based on a self-built dataset. The head pose is estimated by applying rigid body transformation to an individual 3D face and the general 3D face model. In order to verify the accuracy of Camper’s plane localization and HPE, 102 cases of 3D facial data and experiments were collected and conducted. The tragus and nasal alar points are localized to within 7 pixels (about 0.83 cm) and the average accuracy of the three dimensions of Camper’s plane identified is 0.87°, 0.64° and 0.47° respectively. The average accuracies of the three dimensions of HPE were 1.17°, 0.90° and 0.97. The experiment results demonstrate the effectiveness of the method for Camper’s plane localization and HPE.

\section*{INDEX TERMS}
Head pose estimation, multi-view, camper’s plane, depth sensor, 3-D point cloud.

\section*{I. INTRODUCTION}
The human face records important characteristics of the human. Face features are widely used in face recognition, expression recognition, entertainment, e-commerce, beauty and healthcare \cite{1, 2, 3, 4, 5}. The human face has a complex and fine geometric structure, and face geometric feature attributes are important properties of the human face. Facial geometric features have been extensively studied in the fields of genetics, genetic disease screening and anthropology, and have the advantages of being fast and contactless in use \cite{6, 7, 8, 9}. 2019 nature medicine reported facial features to screen for genetic diseases \cite{10}. It can also be used for the detection and measurement of facial swelling, gender classification and face detection and classification \cite{11, 12, 13}. Facial features contain rich information about diseases and people have a history of assessing health status through the face since ancient times \cite{14}. Therefore, facial feature recognition techniques are useful in a wide range of fields such as biometric identification, healthcare, business and public safety.

Head pose estimation is a key step in the computational and quantitative study of 3D facial features, where head pose can be determined relatively easily by humans, but
remains challenging for computers. In computer vision, Head Pose Estimation is the process of inferring the position and orientation of the human head from digital images, requiring the conversion of pixel-based representations of the head into a more advanced parametric form of head pose. In head pose estimation, the left-right symmetry information of the head can be well resolved in the roll and yaw dimensions of pose angle estimation, while the pitch dimension of head pose estimation requires the use of the relative geometric position relationships of salient regions of the face. Facial feature recognition head pose estimation methods are currently available based on traditional geometric model methods [15], [16], [17], [18], [19] and deep learning [20], [21], [22], [23], [24], [25], [26], [27] based methods. The head pose estimation methods described above are geared towards a wide range of pose estimation, and the accuracy of pose estimation is still relatively coarse. Most of the previous head posture estimation researches evaluate the head posture estimation using the mean error and pay less attention to the maximum angular error of the head relative to the absolute horizontal plane after posture correction. Research on head pose estimation for biology, psychology and medical related fields requires high accuracy, certainty (note: deep learning methods have some uncertainty) and robustness of the computational results. Due to the large differences in facial features between different people, determining the standard head pose is a difficult matter [15], and Camper’s plane and Frankfurt plane have been proposed and used by anthropologists and others [28], [29], [30], and the research in this paper focuses on Camper’s plane.

Although head pose estimation has been extensively studied, it is still difficult to achieve stable and high-accuracy head pose estimation. This could come from two limitations. On the one hand, it is difficult to obtain a large number of continuous, accurate and complete 3D faces, which limits the development of research work. On the other hand, there is a lack of automatic feature point localization methods for complete 3D faces. To overcome these limitations, we propose a novel multi-view depth sensor-based method for acquiring full 3D faces, Camper’s plane localization, and head pose estimation.

Specifically, the significant contributions of this research paper are as follows:

1) We propose a continuous full face acquisition method based on multi-view depth sensors, which can acquire full 3D faces continuously and rapidly.
2) We construct a multi-view 3D face dataset CAS-MVS-3D-FACE-DATASETS, and propose a 3D Camper’s plane localization method. Pixel error and spatial distance error results for Camper’s plane localization show that our method can effectively localize Camper’s plane.
3) We extract a method for constructing a 3D facial model with 18 key points including left and right alar points and tragus points.
4) Include the estimated maximum error and errors less than 2° and less than 5° into the error evaluation.
5) We propose a head pose estimation method based on rigid body transformation estimation and conduct extensive experiments to verify the effectiveness of our method. Experimental results with mean error, maximum error, error less than 2°, and error less than 5° of head pose estimation demonstrate that our head pose estimation method achieves state-of-the-art performance and can effectively estimate head pose.

The remainder of this paper consists of 4 sections; Section II presents related work, Section III describes the proposed method, followed by experiments and results in Section IV, and in Section V presents the conclusions and some future research directions.

II. RELATED WORK
Head pose estimation has been a highly studied topic for the past 50 years. Head pose estimation based on vision sensor is to find the mapping of head information in the image to the head pose. According to the research papers studied for this work, the review work is divided into two parts, methods based on traditional geometric models and methods based on deep learning.

A. TRADITIONAL GEOMETRIC MODEL METHODS
The POSIT algorithm was proposed in [16]. Detecting and matching four or more non-coplanar feature points of an object in an image, with their relative geometry on the object is known, which just eliminates the need to start with an initial guess and computes pose by using orders of magnitude fewer floating-point computations. The POSIT algorithm can only obtain an approximate solution of the pose transformation matrix with a certain accuracy at last. A non-iterative PnP algorithm was proposed in [17]. The PnP algorithm is widely used in head pose estimation scenarios. The PnP problem is to study how to solve the camera pose from 3D-2D matching pairs (representing 3D points as four weighted sum of two virtual control points, and solve according to their coordinates - very general). In [18], according to practical application requirements, the classical PnP problem is extended to the case of two images under the translational motion of the camera, and a linear solution method of this kind of PnP problem is proposed. The pose estimation problem is formulated in [19] as minimizing an error metric based on object (relative to image) spatial collinearity. Using the object space collinearity error, a globally convergent iterative algorithm for directly computing orthogonal rotation matrices is derived. Methods based on traditional geometric models mainly use the correspondence between two-dimensional image features and three-dimensional head models to estimate the head pose. The estimation results are affected by the large translation of the head and camera pose, so the accuracy is low.
B. DEEP LEARNING METHODS

Real-time, six degrees of freedom (6DoF) 3D face pose estimation was proposed in [20] without face detection or landmark localization. They propose a 6DoF face pose estimation and alignment method that does not rely on first running a face detector or locating facial landmarks. A new pose transformation algorithm is proposed to maintain the consistency of pose estimation for the same face in different image crops. Extensive experiments are used to demonstrate the effectiveness of this method for face pose estimation and face detection. In [21], a deep learning-based method for estimating 3D facial expression coefficients was proposed based on the fact that a CNN can be trained as a 3D deformation model with head pose information in the regression accuracy and discrimination directly from image intensity. In [22], it was shown how to directly estimate the face shape, viewpoint, and expression of the face containing the head pose from the image, without using facial markers. Furthermore, during deep 3D face modeling, facial landmarks can be obtained. The head pose estimation problem is solved in [23] by reducing the problem’s complexity. A deep task reduction-guided image regularization module is integrated with an anchor-guided pose estimation module, and the HPE problem is formulated as a unified end-to-end learning framework. ElasticNet and DCNN were used for HPE respectively in [24], and the obtained results were then combined with the proposed coordinate pair angle method (CPAM) method to provide results with high accuracy. A multi-stream multi-task deep neural network was proposed in [25] for human detection and head pose estimation in RGB-D videos. Three streams (RGB, depth, and flow data) are fed to a deep neural network, scale-invariant proposals for depth data are proposed, and a multi-stream deep network is constructed with using CRP. In [26], a deterministic conditional GAN model was used to convert depth images to gray-level images, and a complete end-to-end framework for monitoring driver body pose was proposed for estimating head and shoulders based only on depth images geometry. There is no need to precompute specific facial features. A head pose estimation method for various facial conditions such as occlusion and challenging viewpoints was proposed in [27]. A combination of coarse and fine feature map classification was proposed to train a multi-loss deep convolutional neural network in order to get the exact Euler angles for the head position. Based on the premise that there is a relationship between the 3D head pose and certain features of the head image, deep learning based methods recovering this relationship while determining head pose by training a large number of head images with known poses. However, because of the unclear mapping from input to output, deep learning based methods would have a large bias in the pose estimation results for some inputs. The head pose estimation methods mentioned above cannot meet the needs of applications such as 3D facial measurement.

This paper presents a complete framework that fuses several modern aspects of computer vision and investigates how to infer precise head poses from images. We design a method to automatically acquire full 3D face based on multi-view depth sensors. Then we propose a method for automatic localization of 3D Camper’s plane keypoints and Camper’s plane based on multi-view RGBD depth sensors. We also propose an 18-point 3D face reference model construction method, and then propose a head pose estimation method based on multi-view depth sensors with higher accuracy than previous methods.

III. METHODS

In this section, we focus on the main steps of Camper’s plane localization and head pose estimation based on multi-view RGBD depth sensors. The system mainly includes multi-view 3D facial data acquisition, multi-view facial feature point localization, Camper’s plane localization and head pose estimation. The overall block diagram of the system solution is shown in Figure 1.

A. FACIAL KEYPOINT LOCALIZATION MODEL

Facial keypoint localization is a key step in Camper’s plane localization and head pose estimation. Camper’s plane localization and head pose estimation based on multi-view depth sensors require the localization of facial tragus points and nasal alar points.

As is the nasal alar point, the tragus point is also unclearly distinguishable from the area around, and it is difficult to achieve automatic localization of the tragus point and nasal alar point by using traditional methods. Compared to traditional methods, deep learning has stronger abstraction and expressiveness. The Keypoint RCNN deep neural network model is a model formed by adding the keypoint detection branch and deleting the mask branch to Mask RCNN [31], which mainly consists of a feature extractor module, a region proposal network and a region of interest head network. The structure is shown in Figure 2. This model is a two-stage model that can perform three tasks simultaneously: target detection, target classification and keypoint detection. The key sub-modules of the model will be introduced separately.
Keypoint RCNN neural network structure mainly consists of: feature extractor, region candidate network and region of interest head network.

1) FEATURE EXTRACTOR

The Keypoint RCNN uses the residual network [32] (ResNet) as the backbone network for feature extraction. To make the model capable of multi-scale detection, the Keypoint RCNN adopts a feature pyramid network [33] (FPN) structure, as shown in Figure 3. The residual network is divided into five stages: stage1, stage2, stage3, stage4 and stage5, each with a different feature map size, and the feature map in the latter stage is 0.5 times as long as the previous one. The feature map is then up-sampled and up-sampled by a factor of 2, summed with the feature map from the previous stage by $1 \times 1$ convolution, and then convolved by $3 \times 3$. Since the stage1 feature map is not used, the above operation produces new feature maps for stage2, stage3, stage4 and stage5, as well as the feature map at the starting point of upsampling, for a total of five feature maps. The feature pyramid network allows features to have both strong semantic and strong spatial information.

2) REGION PROPOSAL NETWORK

The role of the region proposal network (RPN) is to generate a series of anchor frames that are classified as positive examples and to correct the coordinates of the anchor frames for subsequent networks to make more accurate predictions of detection frames and masks, etc. The structure is shown in Figure 4.

3) REGION OF INTEREST HEADS

The input to the network is the five feature maps output by the feature extractor, and the region proposal network generates anchor frames at each point on the feature map. The Softmax layer classifies the anchor frames and determines whether the anchor frames contain objects, i.e. whether they are positive examples. The Softmax layer classifies the anchor frames and determines whether they contain objects, in other words, they are positive or not. After obtaining a large number of positive anchor frames, in order to reduce the amount of data and speed up the training, the anchor frames are sorted by classification confidence and the anchor frames with high confidence are selected. Then the anchor frames that are beyond the boundary of the feature map are cropped and those with sizes smaller than the threshold are removed. As the network may generate multiple anchor frames on the same object, non-maximal suppression is also performed.

The detection box branch outputs the classification confidence of the object and the regression value of the detection box coordinates through multiple fully connected layers. Using the cross entropy loss to calculate the classification loss, denoted as $\text{Loss}_{\text{cls}}$, and using Smooth L1 Loss to calculate the detection box coordinate regression loss,
denoted as \( \text{Loss}_{\text{box, reg}} \).

\[
\text{Smooth}_L_1\text{Loss} = \begin{cases} 
0.5x^2 & |x| < 1 \\
|x| - 0.5 & |x| \geq 1 
\end{cases} 
\] (1)

The keypoint branch obtains an \( N \times 28 \times 28 \times C \) feature map through \( 3 \times 3 \) convolution and transposed convolution, where \( C \) is the number of key points. The feature map is enlarged by bilinear interpolation, and the annotation is converted into a heat map. Then calculate the cross-entropy loss with the feature map and heat map, denoted as \( \text{Loss}_{\text{kp}} \).

The loss function of Keypoint RCNN is

\[
\text{Loss} = \text{Loss}_{\text{obj}} + \text{Loss}_{\text{rpn, box, reg}} + \text{Loss}_{\text{cls}} + \text{Loss}_{\text{box, reg}} + \text{Loss}_{\text{kp}} 
\] (2)

In the formula: \( \text{Loss}_{\text{obj}} \) is the positive and negative case classification loss of RPN and \( \text{Loss}_{\text{rpn, box, reg}} \) is the detection box regression loss of RPN.

The Keypoint RCNN deep neural network model contains 3 types of network layers: convolutional (Conv) layer, Batch Normalization (BN) layer, and fully connected (FC) layer. The number of various network module layers and the number of parameters contained in Feature Extractor, RPN, and ROI Heads were counted. This deep neural network contains a total of 131 network layers with a total of 59125743 parameters. The number of network layers contained in each key sub-module is shown in Table 1. The number of parameters of each network layer contained in each key sub-module is shown in Table 2.

TABLE 1. Number of Conv, BN, and FC network layers contained in Feature Extractor, RPN, and ROI Heads respectively.

|                     | Conv | BN   | FC   |
|---------------------|------|------|------|
| Feature Extractor   | 61   | 54   | 0    |
| RPN                 | 3    | 0    | 0    |
| ROI Heads           | 9    | 0    | 4    |

TABLE 2. Number of parameters of the Conv, BN, and FC network layers contained in Feature Extractor, RPN, and ROI Heads, respectively.

|                     | Conv          | BN   | FC      |
|---------------------|---------------|------|---------|
| Feature Extractor   | 26799296      | 106240 | 0       |
| RPN                 | 599075        | 0    | 0       |
| ROI Heads           | 13905930      | 0    | 17715202|

This network has the characteristics of simple structure, few parameters, and strong robustness, and conforms to Occam’s razor principle of “Do not multiply entities beyond necessity”. This RCNN network can effectively extract image features and achieve facial keypoint localization. Compared with one-stage networks such as YOLO and SSD, RCNN series networks have better performance and robustness, which are widely used in scenarios requiring robustness and high precision.

B. CAMPER’S PLANE POSITIONING

The positioning of the 3D Camper’s plane requires the acquisition of the spatial positions of 2 tragus points and 2 nasal alar points, which cannot be acquired simultaneously for a single view camera. There are relative rotations and translations between the point clouds acquired by different view cameras, and the complete 3D face needs to be aligned with the 3D facial point clouds acquired by multiple view cameras. Multi-view point cloud alignment is the basis for the 3D Camper’s plane localization.

1) MULTI-VIEW FACIAL POINT CLOUD REGISTRATION

How to accurately acquire the relative poses between point clouds from multiple views is a difficult and worthwhile problem. Fortunately, the scene in this paper can be considered to have fixed relative poses between point clouds acquired from different viewpoints. The chessboard shown in Figure 6 is relatively flat and has significant texture, allowing the relative poses between the collected multi-view 3D tessellated camera calibration plate point clouds to be obtained relatively accurately. The relative poses obtained from the tessellation plate point cloud alignment are used to adjust the spatial position and pose of the face point cloud so that the three views of the face point cloud can together form a complete 3D face.

The current mainstream point cloud alignment algorithm is the Iterative Closest Point (ICP) algorithm and its variant [34], [35]. The ICP algorithm relies on a relatively high initial value, and a good initial value can yield better alignment results and faster alignment speed. The checkerboard calibration board is placed at the multi-view face acquisition location to collect multi-view point clouds. The ICP algorithm is used to align the multi-view tessellation point cloud, the results of which provide an ideal parameter value for the multi-view face point cloud alignment.

A chessboard is used to calibrate the relative poses of multiple cameras to obtain the camera extrinsic parameters for the multi-view camera system. By using the acquisition software we developed, the multiple camera system could capture the chessboard simultaneously. The color images of the multiple viewpoints are located using OpenCV’s findChessCorner function to locate all corner points of the checkerboard grid, find the data in the depth images at the corresponding locations, and transform them into 3D spatial point cloud data.
based on the depth camera’s internal parameters, and then estimate the relative positional R, t parameters between the multiple viewpoints using the ICP algorithm. The point cloud is then aligned by rotation and translation of the multi-view 3D spatial point cloud using the multi-view camera system camera external parameters obtained from the calibration.

2) CAMPER’S PLANE POSITIONING

Camper’s plane positioning is completed by two tragus points and two nasal alar points. A plane is determined by the left tragus point, the right tragus point and the midpoint of the two alar points, which is Camper’s plane.

The depth sensor camera used by the system allows both color and depth images to be acquired and aligned. The locations of the tragus point and the nasal alar point in the color image are located using the Keypoint RCNN algorithm described earlier. The data from the corresponding positions in the depth map is then used in conjunction with the camera’s internal parameters to calculate the 3D spatial coordinates of the four 3D points.

The detailed positioning steps of Camper’s plane are as shown in Figure 7. A schematic diagram of the Camper’s plane is shown in Figure 8.

\[
\mathbf{\hat{n}}_i \times \mathbf{\hat{n}}_j \times \mathbf{\hat{n}}_k = \mathbf{0} \\
\mathbf{\hat{n}}_i \times \mathbf{\hat{n}}_j \times \mathbf{\hat{n}}_k \times \mathbf{\hat{n}}_k = K
\]

C. HEAD POSE ESTIMATION

Head pose estimation can be performed through the Camper’s plane 4 keypoints obtained in Section II.A and Medi-apipe [36] gets 14 facial feature points: four brow bone points, four eye points, two nose points, two mouth points, lower center points of mouth and chin points. The 18 facial feature points are shown in Figure 9.

The Camper’s plane has a clear anthropological and anatomical definition. Rotating the Camper’s plane around the line connecting the left and right tragus points can be approximately rotated to the true horizontal plane. Therefore, the Camper’s plane can be easily positioned and used for head pose estimation. Head pose estimation based on Camper’s plane mainly uses 4 key points of Camper’s plane and other 14 facial feature points. The steps mainly include: constructing a 3D face reference model, extracting individual face 3D key points and solving pose parameters.

1) BUILD FACE MODEL

Building a general 3D face model from a self-built dataset consists of 5 steps. The construction processes are as follows and are shown in Figure 11.
(1) Extract the three-dimensional coordinates of 18 facial feature points: two tragus points, two alar points, four brow bone points, four eye points, two nose points, two mouth points and two chin points.

(2) Move the origin of the coordinates to the geometric centers of the 18 3D points.

(3) Calculate the scale factor: the sum of the distances between the four points in pairs.

(4) The 3D data points are normalized according to the scale factor of (3).

(5) The mean values of the 3D coordinates of each of the 18 feature points were calculated for the normalized data. The 18 mean values are used as a generic 3D face reference model.

2) HEAD POSTURE ANGLE CALCULATION
First of all, 18 facial feature points of the individual 3D face are extracted. Then, the relative pose parameters (rotation matrix R and translation t) between the extracted individual facial keypoint data and the constructed 3D face reference model are calculated using the 3D rigid body transformation estimation method. Figure 12 shows a schematic diagram of the relative pose parameters R and t calculated by the rigid body transformation estimation. Finally, the Euler angle is calculated by the rotation matrix, and the head posture pitch, roll and yaw are obtained.

IV. EXPERIMENTAL RESULTS AND DISCUSSION
In this section, the preliminaries of the data acquisition hardware system, dataset creation and data pre-processing are described, followed by experiments and analysis of the results for Camper plane positioning and head pose estimation.

A. PRELIMINARY PREPARATION
This section describes the construction of the multi-view 3D facial acquisition hardware system, the creation of the experimental data set and the pre-processing of the experimental data.

1) MULTI-VIEW 3D FACE ACQUISITION SYSTEM
In conventional visual head pose estimation systems, a single camera is mainly used. A single camera cannot acquire the face completely at once. Areas such as both ears cannot be acquired simultaneously or are severely distorted, limiting the continuous acquisition and use of the Camper’s plane. In order to acquire complete 3D facial area data, a spatial modelling approach was used to design the acquisition scheme, and after experimentation and analysis, a multi-view simultaneous acquisition scheme was chosen. In order to determine the appropriate number of cameras and spatial location layout, a multi-degree-of-freedom hardware mechanical structure was designed for experimental testing. The spatial layout of the cameras as shown in Figure 13 was obtained after testing, and Figure 14 shows the final produced hardware structure. The face image acquisition distance is approximately 50cm to 55cm, with one camera for the frontal view, one for the left deflection at 30° and one for the right deflection at 30°.

For the choice of depth sensor, based on the reliability and consistency of RealSense measurements reported in previous studies, the Realsense D415 can be considered a viable option for performing objective 3D anthropomorphic measurements of the face in the middle of the frontal face where a low-cost and portable camera is required [37].

FIGURE 13. Multi-view 3D face acquisition hardware system structure design.

FIGURE 14. Front view (a) and side view (b) of the multi-view 3D face acquisition hardware system.

2) EXPERIMENTAL DATA
The multi-view 3D face acquisition system equipment described in the previous section allows for simultaneous acquisition of multi-view color images and depth point cloud data. The multi-view color and depth images are captured as shown in Figure 15.

Data set creation: Images captured from both the left and right views can be added to the tragus point and nasal alar
point localization data sets and are unnecessarily differentiated, with a total of 792 sample data from 102 people. Informed consent was obtained from all subjects involved in the study. The collected multi-view image data was labelled with tragus points, nasal alar points and facial areas according to anthropological measures using the software Labelme. A total of 792 images were labelled. The collection of the acquired images and the annotated data was named CAS-MVS-3D-FACE-DATASETS, or CAS-M3FD for short. 582 images were used as the training set and 204 as the test set.

### TABLE 3. Experimental data.

|       | Sum | Train sample | Test sample |
|-------|-----|--------------|-------------|
|       | 792 | 582          | 204         |

3) DATA PREPROCESSING

In order to facilitate the subsequent algorithm design, the data is preprocessed. Filling the depth map holes makes the depth map face region free of invalid data. The RGB images have a similar data distribution by color correction [38].

Depth map preprocessing: The depth map of the face area may have individual pixel data with an invalid value of 0. A neighborhood non-zero value mean filling method is proposed to fill holes in the face region data of the depth map, which can effectively eliminate invalid values. The formula is as follows:

\[
depth_{\text{filled}}_{i,j} = \frac{1}{9} \sum_{i=1}^{3} \sum_{j=1}^{3} depth_{i,j}, \quad (depth_{i,j} \neq 0) \tag{3}
\]

Color image color correction: Color correction can make the image data have a close data distribution range, which is conducive to more accurate recognition and positioning of facial features. The color correction is performed through the color correction algorithm in the MCC module in OpenCV_Contrib. The calibration color used is the most popular Macbeth ColorChecker. The Macbeth ColorChecker provides the reference color for correction. It includes 4 × 6 color blocks, among which, the color block in the last row is a gray patch that can be used for grayscale linearization or white balance.

### B. CAMPER’S PLANE LOCALIZATION

In this section, the preprocessed data is used for the positioning of Camper’s plane 3D key points, Camper’s plane positioning and head pose estimation experiments.

1) KEYPONT POSITIONING

Experimental results: where Table 4 shows the statistics of the minimum, maximum and mean values of the pixel errors between the auto detection and the label of the tragus point and nasal alar points. Table 5 shows the statistics of the three-dimensional Euclidean distance errors between the auto-positioning and the label of the tragus point and the nasal alar, and the maximum error are 8.3 mm and 6.6 mm, and the mean values are 3.2 mm and 2.5 mm, achieving accurate 3D spatial position positioning.

### TABLE 4. Pixel error between automatic positioning and labeling of tragus points and alar points.

|                  | Min (pixel) | Max (pixel) | Mean (pixel) |
|------------------|-------------|-------------|--------------|
| Tragus row       | 0           | 6           | 1.17         |
| Tragus column    | 0           | 4           | 2.28         |
| Alar row         | 0           | 7           | 2.02         |
| Alar column      | 0           | 5           | 1.62         |

### TABLE 5. Three-dimensional Euclidean distance error between automatic positioning and labeling of tragus and alar points.

|                  | Min (mm) | Max (mm) | Mean (mm) |
|------------------|----------|----------|-----------|
| Tragus           | 0        | 8.3      | 3.2       |
| Alar             | 0        | 6.6      | 2.5       |

2) CAMPER’S PLANE LOCALIZATION

Table 6 shows the maximum and average values of the errors in Pitch, Roll and Yaw between the automatically positioned camper’s plane and the calibrated camper’s plane. pitch, roll and yaw had average errors of 0.87°, 0.64° and 0.48° respectively, achieving a low cost, automated and accurate method of positioning Camper’s plane.

### TABLE 6. Euclidean distance error in three-dimensional space between automatic positioning and labeling of tragus and alar points.

|                | Pitch (°) | Roll (°) | Yaw (°) |
|----------------|-----------|----------|---------|
| Max error      | 4.36      | 3.87     | 1.43    |
| Mean error     | 0.87      | 0.64     | 0.48    |

### C. HEAD POSE ESTIMATION

Head pose estimation was performed using the collected data based on the constructed 3D face reference model. The results were compared with the EPnP [17] and Img2pose [20] methods for head pose estimation.
It is shown in Figure 16 the comparison of the Euler angle pitch, roll, and yaw of the test sample using EPnP, Img2pose, and our algorithm to estimate the head pose results. The label (blue line) in the figure is the Euler angle pitch, roll, and yaw of the head pose of each sample manually labeled. The epnp (orange line) in the figure represents the Euler angle pitch, roll, and yaw of the head pose estimated by the EPnP algorithm. The img2pose (green line) in the figure represents the Euler angle pitch, roll, and yaw of the head pose estimated by the Img2pose algorithm. The ours (red line) in the figure represents the Euler angle pitch, roll, and yaw of the head pose estimated by the algorithm we extracted.

We show in Figure 17 the Euler angle pitch, roll, and yaw error comparisons between the test samples using EPnP, Img2pose, and our algorithm to estimate the head pose results and the manually annotated head pose labels. The epnp (blue line) in the figure represents the head pose estimation error curve of the EPnP method. The head pose estimation error
curve of the Img2pose method is represented by img2pose (orange line) in the figure. Ours (green line) in the figure represents the head pose estimation error curve using our extracted method. It can be seen from the error curve in the figure that our method is robust compared to EPnP and Img2pose.

We show in Figure 18 the side-view visualization of some test samples after pose correction using the head pose estimation results from EPnP, Img2pose, and our algorithm, respectively. Among them, Raw is the original head pose side view display captured by the camera. GT is a side-view visualization of pose correction using human-annotated pose parameters. EPnP, Img2pose, and Ours in the figure are side view visualizations obtained by correcting the head pose according to the head pose parameters estimated by EPnP, Img2pose, and our algorithm, respectively.

We show in Figure 19 some examples of low accuracy head pose estimation using our proposed algorithm. Raw in
The method we proposed is compared with the EPnP [17] and Img2pose [20] methods for head pose estimation. The experimental results are shown in Table 7. Compared with the previously reported EPnP method and Img2pose method, using the method proposed in this paper, pitch and roll have higher accuracy, the pitch accuracy is slightly lower than the EPnP method, and the pitch accuracy of the three methods is high. The proportions of pitch, roll and yaw angle errors of head pose estimation less than 2° and 5° are counted, as shown in Table 8. The pitch, roll and yaw errors of head pose estimation are 87.25%, 91.18% and 100.0%, respectively, and the errors are less than 5° are 96.08%, 100.0% and 100.0%, which are significantly higher than the other two method. There is a clear advantage to the estimation method based on the full 3D face.

The EPnP method for head pose estimation uses the Perspective-n-Point (a method for solving 3D to 2D point-to-point motion) principle, which is strongly influenced by the translation of the person with respect to the camera. The deep learning method can learn more features related to the head pose, and map it to the pose angle in the output part of the neural network, which makes it robust compared to traditional methods when estimating head pose over a wide angular ranges. However, due to the characteristics of deep neural networks, it is difficult to achieve particularly accurate and stable deterministic results for head pose estimation.

All in all, this method has clear advantages in high precision facial acquisition and analysis application scenarios.

| Methods    | Pitch (%) | Roll (%) | Yaw (%) | Pitch (%) | Roll (%) | Yaw (%) |
|------------|-----------|----------|---------|-----------|----------|---------|
| EPnP[17]   | 2.97      | 1.06     | 0.97    | 14.54     | 22.18    | 2.31    |
| Img2pose[20]| 1.17      | 0.90     | 0.97    | 7.08      | 4.56     | 1.74    |
| Ours       | 6.25      | 91.18    | 100.0   | 96.08     | 100.0    | 100.0   |

**Table 7. Comparison of the head pose estimation results of the method proposed in this paper, the EPnP method and the Img2pose method.**

| Methods    | ≤ 2° (%)  | ≤ 5° (%)  |
|------------|-----------|-----------|
| Pitch (%)  | Roll (%)  | Yaw (%)   |
| EPnP[17]   | 24.51     | 98.04     |
| Img2pose[20]| 52.00    | 98.04     |
| Ours       | 87.25     | 100.0     |

**Table 8. Proportion of pitch, roll and yaw errors of head pose estimation less than 2 degrees and 5 degrees.**

**V. CONCLUSION AND FUTURE WORK**

Nowadays, the key supporting technology of automatic measurement of 3D facial features will play a role in promoting research in the fields of genetics, genetic disease screening, and anthropology related to facial features. The system solution proposed in this paper effectively acquires the complete face and achieves tragus and nasal alar points localization, and further enables localization in the Camper’s plane of the 3D face, achieving better convenience and higher accuracy and robustness. This system can be applied to automated,
interpretable head pose estimation, which has broad research implications. The proposed system scheme cannot achieve particularly ideal results for the head pose estimation accuracy of a small number of people. At the same time, the person whose face image is collected needs to be within a range of the distance from the camera and the person needs to face the front camera directly. In order to continuously improve the applicability of the algorithm. In the future, aiming at the problem of large deviation of head pose estimation caused by a large facial gap between individuals and the public, the individual maximum error of head pose estimation test error can be incorporated into the evaluation system. At the same time, hardware devices should be made more portable.

REFERENCES

[1] J. Zhao, S. Yan, and J. Feng, “Towards age-invariant face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 474–487, Jan. 2022, doi: 10.1109/TPAMI.2020.3011426.

[2] L. Jiang, J. Zhang, and B. Deng, “Robust RGB-D face recognition using attribute-aware loss,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 10, pp. 2552–2566, Oct. 2020, doi: 10.1109/TPAMI.2020.2919284.

[3] S. Dwiyanti, M. Iqbal, and B. Y. Suprapto, “Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network,” IEEE Access, vol. 10, pp. 89786–89886, 2022, doi: 10.1109/ACCESS.2022.3290762.

[4] L. Kong, Y. Wu, Y. Zhao, L. Dong, M. Hui, M. Liu, and X. Liu, “Robust imaging photoplethysmography in long-distance motion,” IEEE Photon. J., vol. 12, no. 3, pp. 1–12, Jun. 2020, doi: 10.1109/JPHOT.2020.2995397.

[5] J. Axellson, T. Sundelin, M. J. Olsson, K. Sorjonen, C. Axellson, J. Lasselin, and M. Lekander, “Identification of acutely sick people and facial cues of sickness,” Proc. Roy. Soc. B. Biol. Sci., vol. 285, no. 1870, Jan. 2018, Art. no. 20172430, doi: 10.1098/rspb.2017.2430.

[6] K. Grifantini, “Detecting faces, saving lives,” IEEE Pulse, vol. 11, no. 2, pp. 2–7, Mar. 2020, doi: 10.1109/MPULS.2020.2984288.

[7] A. R. Porras, M. Summar, and M. G. Linguraru, “Objective differential diagnosis of noonan and Williams–Beuren syndromes in diverse populations using quantitative facial phenotyping,” Mol. Genet. Genomic Med., vol. 9, no. 5, May 2021, Art. no. e1636, doi: 10.1002/mgg3.1636.

[8] Q. Zhao, N. Werghi, K. Okada, K. Rosenbaum, M. Summar, and M. G. Linguraru, “Ensemble learning for the detection of facial dysmorphology,” in Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., Aug. 2014, pp. 754–757, doi: 10.1109/IEMBS.2014.6943700.

[9] D. Sero, A. Zaidi, J. Li, J. D. White, T. B. G. Zarzar, M. L. Marazita, S. M. Weinberg, P. Suetens, D. Vandermeulen, J. K. Wagner, M. D. Shriver, and P. Claes, “Facial recognition from DNA using face-to-DNA classifiers,” Nature, vol. 529, no. 1, p. 2557, Dec. 2019, doi: 10.1038/s41467-019-10617-y.

[10] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Bascl-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, L. M. Bird, and K. W. Gripp, “Identifying facial phenotypes of genetic disorders using deep learning,” Nature Med., vol. 25, pp. 60–64, Jan. 2019, doi: 10.1038/s41591-018-0279-0.

[11] A. Paviotti, N. Brusco, and G. M. Cortelazzo, “3D detection and measurement of facial swellings,” in Proc. 14th Eur. Signal Process. Conf., Sep. 2006, pp. 1–5.

[12] X. Han, H. Ugail, and I. Palmer, “Gender classification based on 3D face geometry features using SVM,” in Proc. Int. Conf. CyberWorlds, Sep. 2009, pp. 114–118, doi: 10.1109/CW.2009.41.

[13] N. Sabri, J. Henry, Z. Ibrahim, N. Ghazali, N. A. Angshor, N. F. M. Johari, and S. Ibrahim, “A comparison of face detection classifier using facial geometry distance measure,” in Proc. 9th IEEE Control Syst. Graduate Res. Colloq. (ICSGRC), Aug. 2018, pp. 116–120, doi: 10.1109/ICSGRC.2018.8657592.

[14] X. Li, D. Yang, Y. Wang, W. Zhang, F. Li, and W. Zhang, “TCMInet: Face parsing for traditional Chinese medicine inspection via a hybrid neural network with context aggregation,” IEEE Access, vol. 8, pp. 93069–93082, 2020, doi: 10.1109/ACCESS.2020.2995202.

[15] E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation in computer vision: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 4, pp. 607–626, Apr. 2009, doi: 10.1109/TPAMI.2008.106.

[16] D. F. Dementhon and L. S. Davis, “Model-based object pose in 25 lines of code,” Int. J. Comput. Vis., vol. 15, no. 1, pp. 123–141, Jul. 1995, doi: 10.1007/bf01450585.

[17] V. Lepeit, F. Moreno-Noguer, and P. Fua, “EPoP: An accurate O(n) solution to the PnP problem,” Int. J. Comput. Vis., vol. 81, no. 2, pp. 155–166, 2009, doi: 10.1007/s11263-008-0152-6.

[18] W. Fu-Chao and H. Zhan-Yi, “A linear method for the PnP problem,” J. Softw., vol. 3, pp. 682–688, Jan. 2003.

[19] C.-P. Lu, G. D. Hager, and E. Mjolsness, “Fast and globally convergent pose estimation from video images,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 9, pp. 1010–1023, Sep. 2000, doi: 10.1109/34.862199.

[20] V. Alibiero, X. Chen, X. Yin, G. Pang, and T. Hassner, “‘inn:2pose: Face alignment and detection via 6DoF face pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 7613–7623, doi: 10.1109/CVPR46437.2021.00753.

[21] F.-J. Chang, A. Tuan Tran, T. Hassner, I. Masri, R. Nevatia, and G. Medioni, “ExpNet: Landmark-free, deep, 3D facial expressions,” in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2018, pp. 122–129, doi: 10.1109/FG.2018.00027.

[22] F.-J. Chang, A. T. Tran, T. Hassner, I. Masri, R. Nevatia, and G. Medioni, “Deep, landmark-free FAME: Face alignment, modeling, and expression estimation,” Int. J. Comput. Vis., vol. 127, nos. 6–7, pp. 930–956, Jun. 2019, doi: 10.1007/s11263-019-01151-x.

[23] E. Murphy-Chutorian and M. M. Trivedi, “A comparison of face detection classification methods for small number of people,” in Proc. 9th IEEE Control Syst. Graduate Res. Colloq. (ICSGRC), Aug. 2018, pp. 116–120, doi: 10.1109/ICSGRC.2018.8657592.

[24] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[25] P. J. Besl and N. D. McKay, “A method for registration of 3-D shapes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 14, no. 2, pp. 239–256, Apr. 1992, doi: 10.1109/34.121791.
[35] J. Yang, H. Li, D. Campbell, and Y. Jia, “Go-ICP: A globally optimal solution to 3D ICP point-set registration,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 11, pp. 2241–2254, Nov. 2016, doi: 10.1109/TPAMI.2015.2513405.

[36] K. Sokal. MediaPipe 3D Face Transform, Google. Accessed: Sep. 26, 2022. [Online]. Available: https://github.com/google/mediapipe

[37] T. C. ten Harkel, S. Vinayahalingam, K. J. A. O. Ingels, S. J. Berge, T. J. J. Maal, and C. M. Speksnijder, “Reliability and agreement of 3D anthropometric measurements in facial palsy patients using a low-cost 4D imaging system,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 28, no. 8, pp. 1817–1824, Aug. 2020, doi: 10.1109/TNSRE.2020.3007532.

[38] S. Bianco, A. R. Bruna, F. Naccari, and R. Schettini, “Color correction pipeline optimization for digital cameras,” J. Electron. Imag., vol. 22, no. 2, May 2013, Art. no. 023014, doi: 10.1117/1.JEI.22.2.023014.

**HUAQIANG WANG** is currently pursuing the Ph.D. degree with the Institute of Microelectronics, University of Chinese Academy of Sciences, Beijing, China. His current research interests include instance segmentation, expression recognition, machine learning, and deep learning.

**LU HUANG** is currently pursuing the M.S. degree with the Institute of Microelectronics, University of Chinese Academy of Sciences, Beijing, China. His current research interests include instance segmentation, point cloud processing, visual measurement, machine learning, and deep learning.

**KANG YU** received the bachelor’s degree in automation from China Jiliang University, Hangzhou, China, in 2014, and the M.S. degree from the Armour College of Engineering, Illinois Institute of Technology, USA, in 2016, with a focus on image and signal processing. He is currently working as an Assistant Professor with the Institute of Microelectronics of Chinese Academy of Sciences (IMECAS), Beijing, China. His current research interests include computer vision and intelligent health data collection terminals.

**TINGTING SONG** received the B.S. degree from the Beijing University of Chemical Technology, Beijing, China, in 2014. She is currently an Research Assistant with the Institute of Microelectronics of Chinese Academy of Sciences. She is mainly concerned about modernized technology in traditional Chinese medicine, including novel diagnosis instruments development, medical image processing, disease severity assessment, and medical data analytics. Her research interests include computer vision, image processing, machine learning, and deep learning.

**FENGEN YUAN** received the bachelor’s degree from the Department of Microelectronics Science and Engineering, North University of China, in 2020. He is currently pursuing the M.S. degree with the Institute of Microelectronics, University of Chinese Academy of Sciences, Beijing, China. His study is mainly focused on CV and CNN accelerators.

**HAO YANG** received the B.S. and M.S. degrees in electronic engineering from Jilin University, Changchun, China, in 2001 and 2004, respectively, and the Ph.D. degree in microelectronics and solid state electronics from the Institute of Microelectronics of Chinese Academy of Sciences, Beijing, China, in 2007. He is currently a Professor with the Institute of Microelectronics of Chinese Academy of Sciences. His current research interests include human signal acquisition and health electronics technology.

**HAIYING ZHANG** was born in China, in 1964. She received the B.S. degree in semiconductor physics and technology from the Harbin Institute of Technology, Harbin, China, in 1986, the M.S. degree in microelectronics from Jilin University, Changchun, China, in 1989, and the Ph.D. degree in electrical and electronic engineering from Saint-Petersburg State Electrotechnical University, Saint Petersburg, Russia, in 1993. From 1994 to 1997, she was an Engineer with Croena Semiconductor Ltd., Moscow, Russia, where she was involved in clean-room fabrication and technology process setup. From 1997 to 1999, she worked at the Department of Microwave Circuits and Compound Semiconductor Devices, Institute of Microelectronics of Chinese Academy of Sciences (IMECAS), Beijing, China. In 1999, she became a Professor and the Deputy Director of the Department of Microwave Circuits and Compound Semiconductor Devices, IMECAS. In 2010, she established the RF Integrated Circuit Department, IMECAS. Since 2000, she has been involved with research of RF circuit design technology and device technology for years. Her current research interests include 3G/4G wireless transmitter circuit design technology, embedded multimode, multi-frequency wireless transceiver IP design methodologies, medical electronics integration technology, ultra-wideband wireless personal area network standard, and the Internet of Things-related technologies.

***