Fast and accurate creation of annotated head pose image test beds as prerequisite for training neural networks

Danny Kowerko,[1] Robert Manthey [1] Marcel Heinz,[2] Thomas Kronfeld,[2] Guido Brunnett[2]

Abstract: In this paper we present an experimental setup consisting of 36 cameras on 4 height levels covering more than half space around a centrally sitting person. The synchronous image release allows to build a 3D model of the human torso in this position. Using this so-called body scanner we recorded 36 different positions giving in total 1296 images in several minutes obtaining tens to hundreds of different pitch-roll-yaw head pose combinations with very high precision of less than ±5°. From annotation of 7 facial keypoints (ears, eyes, nose, corners of the mouth) in the 36 calculated 3D models of a human head/upper body, we automatically get 1296 × 7 2D facial landmark points saving a factor 36 in annotation time. The projection of the 3D model to the camera provides a foreground/background separation mask of the person in each image usable for data set augmentation e.g. by inserting different backgrounds (required for training convolutional neural networks, CNNs). Moreover, we utilize our 3D model in combination with textures to create realistic images of the pitch-roll-yaw range not assessed in experiments. This interpolation is ad hoc applicable to a subset of 10 central out of 36 total camera views where fine-grained interpolation of head poses is possible. Using interpolation and background masks for background exchange enables us to augment the data set easily by a factor of 1000 or more knowing precisely pitch, roll, yaw and the 7 annotated facial keypoints in each image.

Keywords: human pose; head pose; 3D body scanner; image annotation; test bed; pitch; roll; yaw

1 Introduction

Nowadays photography technically allows for taking hundreds to thousands of images of the same event within minutes. Photography-intense scenarios e.g. are red carpet, sports, ceremonial or press conference events. Unique to these images might be the fact that persons are appearing with only marginal differences in terms of perspective and angle of view relative to the camera. Systematic organization or selection of images by a person's head spatial posture is still limited using camera software or external image managing programs. Recent advances using convolutional neural networks allow for determination of human full body keypoints [Ca17, We16]. While in [We16] only head and neck coordinates are available, Cao et. al already provide the 2D coordinates of eyes, ears and nose [Ca17].

[1] Technische Universität Chemnitz, Junior Professorship Media Computing, Straße der Nationen 62, 09111 Chemnitz, Germany firstname.lastname@informatik.tu-chemnitz.de

[2] Technische Universität Chemnitz, Chair GDV, Straße der Nationen 62, 09111 Chemnitz, Germany firstname.lastname@informatik.tu-chemnitz.de

doi:10.18420/in2017_221

Both are technically usable to extract information about a human head's position. Typical benchmark data sets from this community mostly contain only the information of some keypoints [BM09, JE10, An14], but not the quantitative orientation of head in terms of pitch, roll and yaw. More sophisticated outcomes have recently been reported by [Fa11, FGVG11, Fa13, Ve17] which are supposed to provide high resolution in term of pitch, roll and yaw in depth images. They archieve a mean error and deviation of 3.6° ± 6° in pitch, 5.5° ± 6.2° and 4° ± 7° in yaw after training with 50000 synthetically generated image faces and a testing set of 10000 real images.

In the research of [Ni12] sequences of images and the optical flow parameters correlating them being used to estimate the position and pose of the head. With this reconstructions of nearly frontal faces being performed to analyze facial expressions. [MB11] used the texture to learn the mapping between frontal and non-frontal facial points, as shown in Fig. 1 for instance, to estimate the pose and the expressions of the face. And [RPP13] used a Coupled Scaled Gaussian Process Regression (CSGPR) to recognize pose-invariant facial expressions. With the use of Kinect sensors the approach of [SAHG15] combine Gabor filter-based features (GAB), Local binary pattern features (LBP) and Histogram of oriented gradients (HOG) to recognize the surface of a head to infer the position and pose. With this a high detection rate of the head pose be achieved within a range of ±30° at pitch, ±20° at roll and ±40° at yaw.

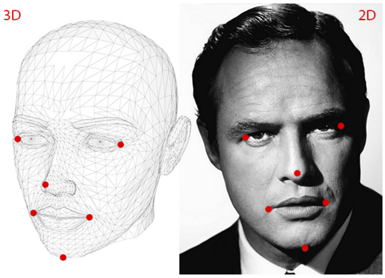

Fig. 1: Facial landmarks needed to calculate the pose of the head.[MA16]

2 Body scanner setup and 3D-Model calculation

A DSLR full body 3D scanner is based on photogrammetry to create a 3D scan. Photogrammetry works by finding shared features between multiple photographs to calculate depth. The body scanner consists of 36 DSLR cameras, which are aimed at the head and upper body of the subject. Each camera is equipped with a 35mm lens. The necessary brightness is provided by four studio flashes. An overview of the construction is shown in figure 2. The recording is triggered by an Arduino® board [AR17] connected to the remote cord of the cameras. This ensures almost simultaneous triggering of all cameras. At the beginning of a recording session, the subject is first focused by all cameras. Then the cameras switch to manual mode in order to prevent further focusing which leads to an asynchronous behavior. In order to ensure a sufficient depth of field for the subsequent reconstruction, an f-number of $f/10$ is used. After each shot, the pictures are successively transferred to the computer which takes about $20 - 30s$. After that the system is ready for the next shot.

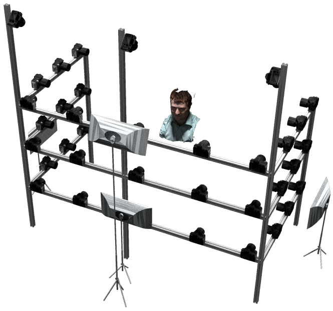

Fig. 2: Sceme of the current body scanner setup.

Agisoft Photoscan [Ag17] is used to reconstruct the 3D model. The software can be automatically controlled using a Python script, which allows the batch-processing of many independent recordings. The schematic sequence of the individual reconstruction steps is shown in figure 3. After loading the images of a recording, a background subtraction is performed. The software then calculates the relative position of the cameras. Therewith, the photogrammetry software will create a point cloud by locating features. In the last step, the 3D point cloud is triangulated and texturized. It takes two to four hours to process a 3D model using Agisoft photoscan depending on the desired resolution. These results were obtained on an Intel Xeon E5-1620 v2 CPU with 32GB memory and an Amd FirePro W8000 graphics card.

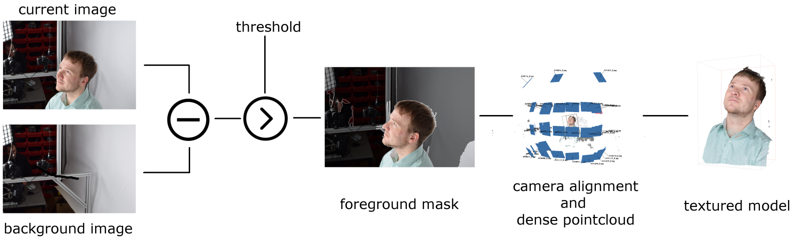

Fig. 3: Processing steps of the 3D model reconstruction.

3 Facial keypoint based calculation of pitch, roll and yaw

We implemented a simple annotation tool based on OpenGL in C++. Based on the reconstructed, triangulated 3D mesh, the 3D model is visualized and can be viewed from arbitrary angles and distances. We use the Möller-Trumbore ray-triangle intersection test [MT05] so that the user can mark arbitrary points on the surface of the mesh by clicking with the mouse. The software manages a set of 7 feature points:

- left and right ear (top/apex of the ear)
- left and right eye (center)
- tip of the nose
- left and right corner of the mouth

The user's task is to annotate these in all 36 3D models as precisely as possible using the built-in zoom and rotation functionality.

The 3D reconstruction done with Agisoft automatically provides the the position and orientation of each camera (extrinsic camera parameters). The instrinsic camera parameters are defined by the focal lenght and the sensor size. The data can be exportet to XML files, which we can use in the annotation tool. The extrinsic parameters can be represeneted by a transformation matrix composed from a rotation matrix and a scale matrix:

$$M_c = T_c \cdot R_c \qquad \forall c \in \{0, \ldots, 35\} \tag{1}$$

Let r be the index of the reference camera - the one camera the test subject was directly looking into. Then, the orientation of the head in all other cameras can be derived as

$$Q_c = R_r^{-1} \cdot R_c \qquad \forall c \in \{0, \ldots, 35\} \tag{2}$$

where Q_c denotes the rotation the test subject's head would have to carry out so that the reference camera would observe the image which was observed by camera c. We decompose this rotatation matrix into a standard rotation sequence with Yaw-Pitch-Roll angles, which allows us to annotate the head orientation to all images of a scan.

Using both the extrinsic and intrinsic camera matrices, we can derive standard view and projection matrices for the 3D rendering, which enables us to render the model - and the annotated viewpoints - from the exact viewpoint, and using the exact field of view, of each camera in the set. This can be used to create binary masks describing the converage of the actual model for each pixel in each camera image. Furthermore, the annotated 3D feature points can be projected into each camera image, yielding the exact 2D locations of these features - even if they are not visibile or obscured by other parts in the particular images.

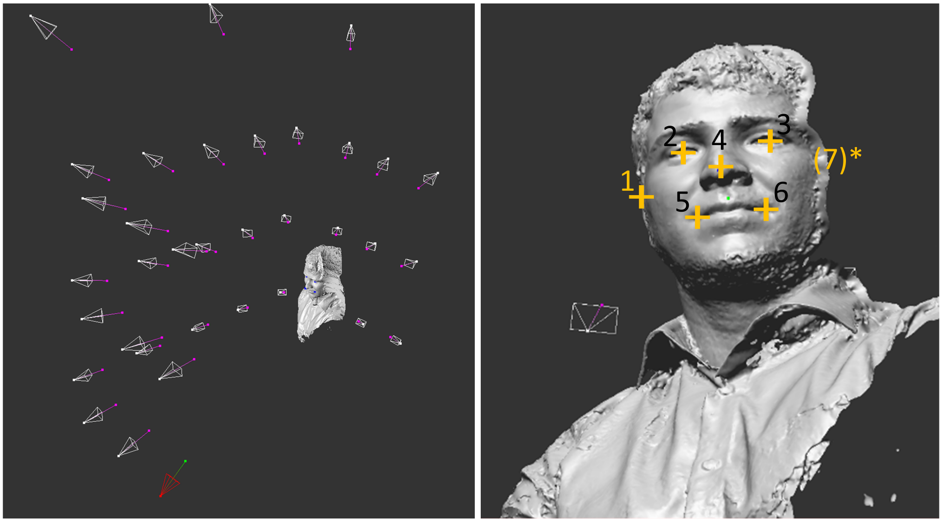

Fig. 4: 3D model of person: (left) Scheme representing the geometic relations relative to the surrounding cameras (red sensor marks the camera with frontal orientation relative to head) and (right) scheme of seven facial keypoints marked manually in 3D space. We use the following convention: 1 - right ear, 2 right eye, 3- left eye, 4 - nose, 5 - right corner of mouth, 6 - left corner of mouth, 7 - left ear. Note, that the left ear (7) is not captured by the body scanner in this view.

4 Data set description

An example of a single shot with 36 images together with the obtained calculated yaw, pitch and roll angles is shown in Figure 5. Watching frontally into the camera yaw, pitch and roll are set to zero. Using the keypoints of eyes and nose to define the plane relative to the camera sensor plane our test person showed the following systematic variations: yaw $=(-2.2 \pm 2.4)^\circ$, pitch $=(39.1 \pm 6.0)^\circ$, roll $=(-2.4 \pm 2.3)^\circ$. This is the insecurity given by the person. The insecurity given by annotation of the respective keypoints was not determined to this end. Yaw was defined to be negative if the right part of face points towards the camera, and to be positive vice versa. Pitch obtains negative values moving the head down relative to the camera and roll is negative when the chin is closer to the camera than the vertex. Within cameras of equal height, only the one at coarsely the height of the head, the pitch remains roughly the same. In cameras recording at worm's or bird's eye view systematic shifts not only of yaw, but also of pitch and roll occur.

The assessed range of pitch, roll and yaw and their interrelations are illustrated in the 3D scatter plot of figure 6(a). The projections help identifying parameter space gaps that have not been measured directly in the 36 shots. Since roll was not actively changed relative to the camera gazed into, for yaw = $0\circ$, we have no roll variation for any pitch and the roll range is

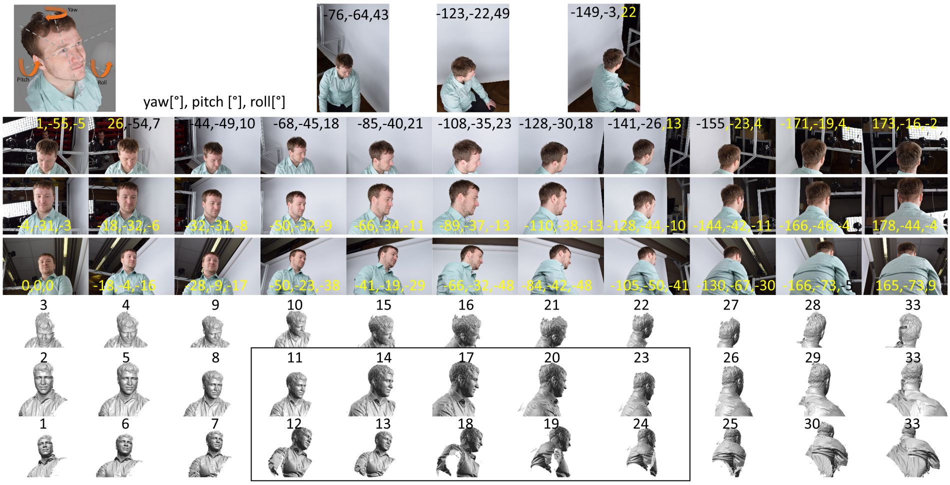

Fig. 5: Example of a single shot taken simultaneously by each camera and the obtained yaw, pitch and roll angles. Here, the person is heading towards the most left camera in the row of the category worm's eye view (#1). The black box marks 3D-models images where the head region is not distinctly affected by gaps.

limited to below ±45◦. Such gaps in the parameter combinations have to be assessed using other head poses or by interpolation of the 3D head model. The yaw angle range obtained is coarsely even until ±90◦. Due to the chosen head poses, we underrepresent large yaw angles (back view of the head).

5 Data set augmentation

For training neural networks data set augmentation strategies are versatile, e.g. the addition of noise, horizontal flip and arbitrary crop. In our dataset, we have additional methods at hand usable for augmentation. The approach presented here allows for discrimination between foreground (person/3D model) and background (everything but the person). Each image taken is provided with a (binary) mask as illustrated in the top of figure 7. The background mask is henceforth used to create arbitrary backgrounds around the person, e.g. indoor, outdoor, textures and colors. Additionally the size of the head/body maybe rescaled and placed to defined or random positions within the background image. The coordinates of the annotated head keypoints and the known orientation help scaling a bounding box for automatically cropping images to the head or head/upper body part as visualized in the right part of figure 7. The second approach for augmentation is the capability to interpolate camera positions, e.g. between each real camera. The obtained 3D model images from the respective views are subsequently calculated and added with the known textures giving a photo-realistic pseudo-image of known head orientation. Technically, all gaps in between the point cloud depicted in the left of figure 6(a) can be filled, which is schematized by the black markers stemming from 5 additional linearly interpolated cameras between camera number 13, 14, 17 and 18. For these interpolated images, foreground/background masks are available too, thus augmentation by background exchange and upper body/head resizing is additionally possible. Since the body scanner setup presented herein covers only about a half space around the person in the middle, the 3D model is not fully closed but includes missing parts in the back towards the side of the upper body and head. Only the central cameras indicated by a black rectangular box in figure 5 provides a more or less complete image of the head. The neighboring cameras sights mostly capture some missing parts of the head from their field of view. It is reasonable to use such examples as representatives for occlusions. Finally, an automatic augmentation of the initial 1296 images could easily be more than a factor of 1000 in the given example: (36+18 interpolated) cameras · 36 head poses/directions · 50 backgrounds · 5 positions (within background image) · 3 body sizes = 1,458,000 images per person, recorded in 30 min and calculated in not more than 2 days computation time using high end consumer hardware. Further augmentation could be achieved using different head poses, illuminations and different crops. Which augmentation strategy will prove beneficial for determination of head poses in real life images will be investigated in detail in the future. Moreover, it seems plausible to train neural nets only for an individual person if her/his activity area is fixed and head orientation changes are small in time e.g. flight controllers in front of PC or car drivers.

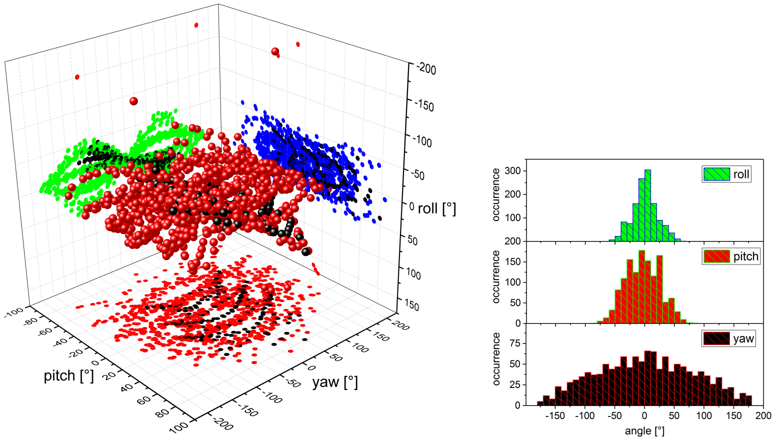

Fig. 6: 3D view of geometric restraints obtained using this 3D body scanner setup and the restricted view always towards a camera's lens. Black spheres and circles indicate interpolated camera positions as depicted in figure XYZ.

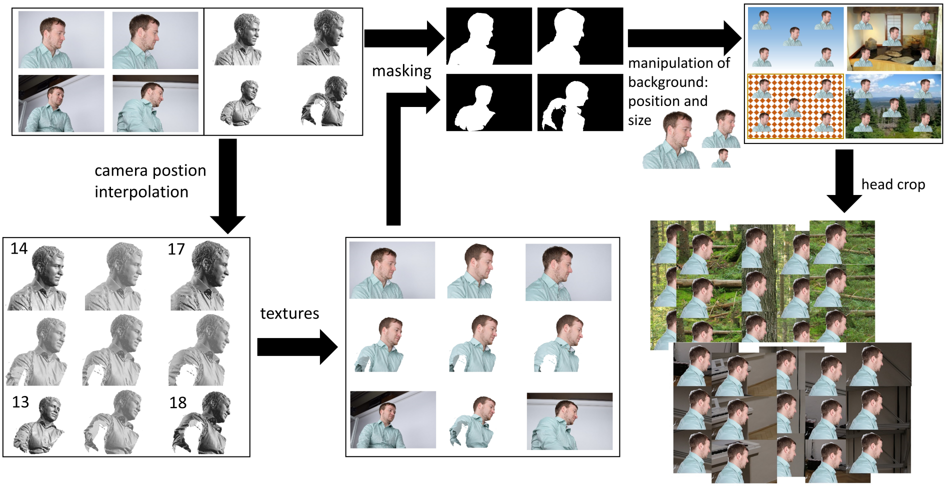
masking
manipulation of background: position and size
camera postion interpolation
head crop
textures
14
17
13
18

Fig. 7: Data set augmentation strategies as prerequisite for deep learning.

6 Conclusion

We have demonstrated the usage of 3D upper body and head scanner setup in the context of head pose analysis. We successfully created a head pose data set containing 1296 real images precisely documented in terms of pitch, roll and yaw with 6°, 2° and 2° accuracy. The obtained 3D model and the automatically obtained foreground/background masks as well as the possibility to create photo-realistic images from interpolated camera positions using the given textures enables a multitude of image augmentation strategies normally not at hand in 2D photography-based image data sets. The data set might be used for testing existing head pose analysis algorithms but provides additionally a good training data base for deep learning based approaches. Moreover, the workflow could be repeated easily to more persons enhancing the head/upper body feature diversity.

Acknowledgments

This work was partially accomplished within the project *localizeIT* (funding code 03IPT608X) funded by the *Federal Ministry of Education and Research* (BMBF, Germany) in the program of *Entrepreneurial Regions InnoProfile-Transfer*.

References

[Ag17] Agisoft LLC.

[An14] Andriluka, Mykhaylo; Pishchulin, Leonid; Gehler, Peter; Schiele, Bernt: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 2014.

[AR17] ARDUINO AG.

[BM09] Bourdev, Lubomir; Malik, Jitendra: Poselets: Body part detectors trained using 3d human pose annotations. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, pp. 1365–1372, 2009.

[Ca17] Cao, Zhe; Simon, Tomas; Wei, Shih-En; Sheikh, Yaser: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: CVPR. 2017.

[Fa11] Fanelli, Gabriele; Weise, Thibaut; Gall, Juergen; Van Gool, Luc: Real time head pose estimation from consumer depth cameras. In: Joint Pattern Recognition Symposium. Springer, pp. 101–110, 2011.

[Fa13] Fanelli, Gabriele; Dantone, Matthias; Gall, Juergen; Fossati, Andrea; Van Gool, Luc: Random Forests for Real Time 3D Face Analysis. Int. J. Comput. Vision, 101(3):437–458, February 2013.

[FGVG11] Fanelli, Gabriele; Gall, Juergen; Van Gool, Luc: Real time head pose estimation with random regression forests. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, pp. 617–624, 2011.

[JE10] Johnson, Sam; Everingham, Mark: Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. British Machine Vision Association, pp. 12.1–12.11, 2010.

[MA16] MALLICK, SATYA: , Head Pose Estimation using OpenCV and Dlib, 2016.

[MB11] Moore, S.; Bowden, R.: Local binary patterns for multi-view facial expression recognition. Computer Vision and Image Understanding, 115(4):541 – 558, 2011.

[MT05] Möller, Tomas; Trumbore, Ben: Fast, minimum storage ray/triangle intersection. In: ACM SIGGRAPH 2005 Courses. ACM, p. 7, 2005.

[Ni12] Niese, R.; Al-Hamadi, A.; Farag, A.; Neumann, H.; Michaelis, B.: Facial expression recognition based on geometric and optical flow features in colour image sequences. IET Computer Vision, 6(2):79–89, March 2012.

[RPP13] Rudovic, O.; Pantic, M.; Patras, I.: Coupled Gaussian processes for pose-invariant facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(6):1357–1369, June 2013.

[SAHG15] Saeed, Anwar; Al-Hamadi, Ayoub; Ghoneim, Ahmed: Head Pose Estimation on Top of Haar-Like Face Detection: A Study Using the Kinect Sensor. Sensors, 15(9):20945–20966, 2015.

[Ve17] Venturelli, Marco; Borghi, Guido; Vezzani, Roberto; Cucchiara, Rita: Deep Head Pose Estimation from Depth Data for In-car Automotive Applications. arXiv preprint arXiv:1703.01883, 2017.

[We16] Wei, Shih-En; Ramakrishna, Varun; Kanade, Takeo; Sheikh, Yaser: Convolutional Pose Machines. CoRR, abs/1602.00134, 2016.