Extended Head Pose Estimation on Synthesized Avatars for Determining the Severity of Cervical Dystonia

Roland Stenger[1][a], Sebastian Löns[2], Feline Hamami[2], Nele Brügge[3], Tobias Bäumer[2]
and Sebastian Fudickar[1][b]

[1]*MOVE Junior Research Group, Institute of Medical Informatics, University of Lübeck, 23562 Lübeck, Germany*
[2]*Institute for Systems Motor Science, University of Lübeck, 23562 Lübeck, Germany*
[3]*German Research Center for Artificial Intelligence, 23562 Lübeck, Germany*

Keywords: Domain Randomization, Deep Learning, Dystonia, Head Pose Estimation, Synthesized Avatars.

Abstract: We present an extended head pose estimation algorithm, which is trained exclusively on synthesized human avatars. Having five degrees of freedom to describe such head poses, this task can be regarded as being more complex than predicting the absolute rotation only with three degrees of freedom, which is commonly known as head pose estimation. Due to the lack of labeled data sets containing such complex head poses, we created a data set, consisting of renderings of avatars. With this extension, we take a step towards an algorithm that can make a qualitative assessment of cervical dystonia. Its symptomatic consists of an involuntary twisted head posture, which can be described by those five degrees of freedom. We trained an EfficientNetB2 and evaluated the results with the mean absolute error (MAE). Such estimation is possible, but the performance works differently well for the five degrees of freedom, with an MAE between 1.71° and 6.55°. By visually randomizing the domain of the avatars, the gap between real subject photos and the simulated ones might tend to be smaller and enables our algorithm being used on real photos in the future, while being trained on renderings only.

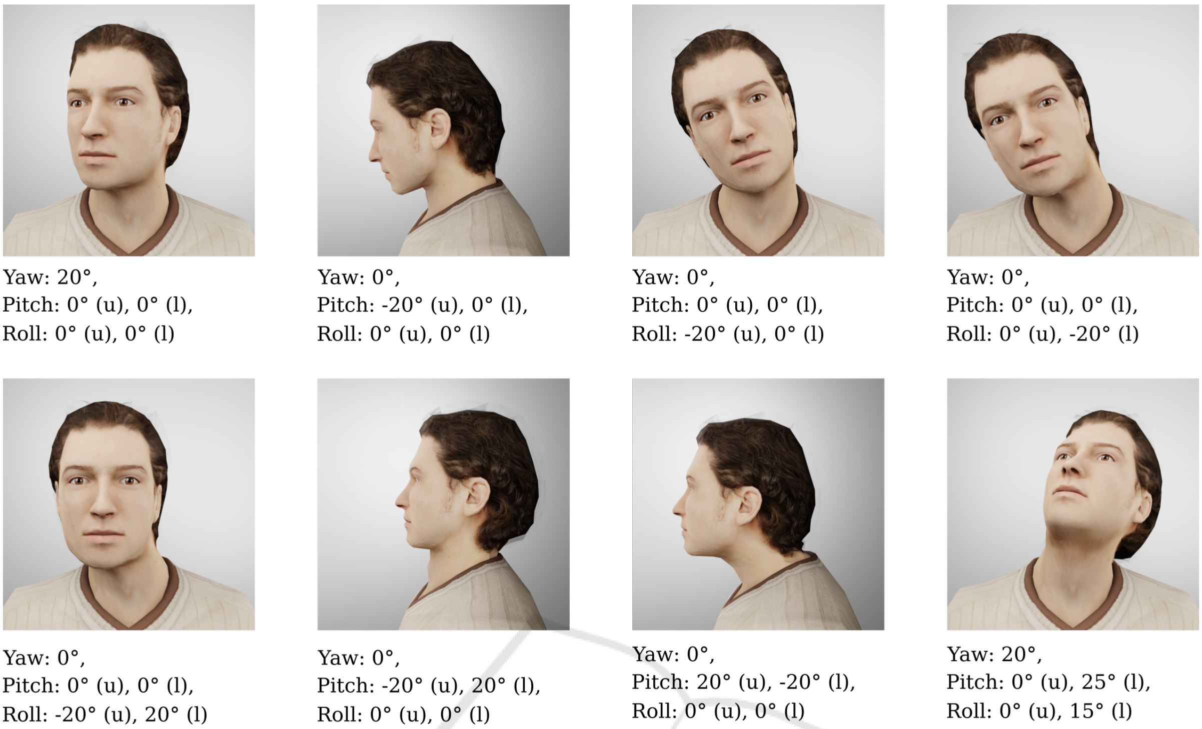
1 INTRODUCTION

Dystonia is a movement disorder, characterized by sustained or intermittent twisting postures. It can be focal or generalized, while the most common form is focal cervical dystonia. The symptomatology of this form of dystonia predominantly comes with twisting and shifting of the head (Albanese et al., 2013).

There are several scores that quantify this altered posture and present it in a severity score. An established score for assessing the severity of dystonia is the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS) (Boyce et al., 2012), which can be collected by clinicians in a direct or video based examination (Zhang et al., 2022). Another score, primarily used for clinical purposes, is the Global Dystonia Severity Rating Scale (GDS) (Comella et al., 2003). In the case of the TWSTRS, the head position is determined by five degrees of freedom. In addition to the rotation around three axes, the severity of a lateral and saggital shift is also determined here. Another description of the head posture can be made according to the caput and collis concept (Finsterer et al., 2015). Here, the head posture is determined with the rotation around an upper and lower rotation center (Figure 2). Following this concept, a lateral or saggital shift can be described as a superposition of an opposite rotation of the roll or pitch, around two rotation centers. Clinicians can determine the severity of a dystonia disorder, using such rating scale. There are already approaches to address this process algorithmically such as (Ansari et al., 2021). However, the authors in this paper present an algorithm which decides on the basis of a video only whether a dystonia disease is present or not. It does not determine the severity of the disease, as we are aiming for. Another publication (Nakamura et al., 2019) deals with the image based determination of the TWSTRS. Here, in contrast to our approach, a depth imaging camera is used. However, such an approach is not as accessible to subjects in the private environment who do not have access to a depth imaging camera.

[a] https://orcid.org/0000-0002-7590-7286
[b] https://orcid.org/0000-0002-3553-5131

Stenger, R., Löns, S., Hamami, F., Brügge, N., Bäumer, T. and Fudickar, S.
Extended Head Pose Estimation on Synthesized Avatars for Determining the Severity of Cervical Dystonia.
DOI: 10.5220/0011677600003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, pages 354-359
ISBN: 978-989-758-631-6; ISSN: 2184-4305



Yaw: 20°,
Pitch: 0° (u), 0° (l),
Roll: 0° (u), 0° (l)

Yaw: 0°,
Pitch: -20° (u), 0° (l),
Roll: 0° (u), 0° (l)

Yaw: 0°,
Pitch: 0° (u), 0° (l),
Roll: -20° (u), 0° (l)

Yaw: 0°,
Pitch: 0° (u), 0° (l),
Roll: 0° (u), -20° (l)

Yaw: 0°,
Pitch: 0° (u), 0° (l),
Roll: -20° (u), 20° (l)

Yaw: 0°,
Pitch: -20° (u), 20° (l),
Roll: 0° (u), 0° (l)

Yaw: 0°,
Pitch: 20° (u), -20° (l),
Roll: 0° (u), 0° (l)

Yaw: 20°,
Pitch: 0° (u), 25° (l),
Roll: 0° (u), 15° (l)

Figure 1: Different types of head postures. Shown are examples of rotations around a single rotation center around one axis, as well as combinations of such rotations.

From the caput and collis concept, the severity of cervical dystonia may be derived. We want to follow this description of the head pose by looking at the rotation around the two rotation centers. This gives us five angles that must be predicted for a complete description of the head posture, which are two pitch angles, two roll angles and one yaw angle. Due to the difficulty of distinguishing a rotation of the yaw around the two centers of rotation (see Figure 1) from each other, we dispense with this subdivision, and summarize these rotations in a single angle that expresses the yaw. In this context, we consider this description to be reasonably accurate.

We want to train an artificial neural network (ANN) for predicting the five degrees. However, there is a lack of publicly available datasets, taking into account the necessary of the head posture with these five degrees of freedom. This limits the trainability of data-driven algorithms. There are public data sets available, which are suitable for training a head-pose-estimation algorithm, such as AFLW2000 (Zhu et al., 2015) or BIWI (Fanelli et al., 2013). Although, in these data sets, the labels are not complete for our purpose, since they only specify the absolute head position in space with three Euler angles. Existing head pose estimation algorithms that predict the head pose in space, based on an image such as (Hempel et al., 2022), (Zhou and Gregson, 2020), (Valle et al., 2021) are also insufficient since their predicted pale head pose cannot fully capture the head pose for our purposes. The head pose must be determined relative to the body and not absolute.

However, the collection of a new data set for our purposes, consisting of annotated images of subjects seems unsuitable for several reasons. First, raters must be trained to qualify for the task. Systematic errors of individual raters cannot be excluded, which is why multiple annotations on the same data would be necessary to estimate the inter-rater reliability for quality assurance. Furthermore, the consent to use the subjects image data must also be obtained.

Due to these disadvantages, we created a synthetic data set consisting of images of avatars with a complex head rotation, based on rotations with the five degrees of freedom. By using a simulated environment we are able to generate a data set with thousands of labeled images (samples) and we have exact information about the head posture which would not be possible with human annotations in this accuracy.

Given the synthesized data set, we address the following research questions (RQ), which we evaluate in Section 3:

- RQ1: How good is the prediction of the ANN for each of the five angle predictions?

- RQ2: What is the influence of the size of the dataset?
- RQ3: How good is the performance as a function of the actual angle size?

We test the performance of the network on the synthesized data only.

2 METHODS

In 2.1 we describe how the data set was generated by using the open source library for human avatars Rocketbox (Gonzalez Franco et al., 2020). In 2.2 we introduce the used ANN architecture and the training process.

2.1 Generation of the Data Set

Using the Rocketbox library, we have a selection of 115 rigged human avatars which we can use for our data set generation. Among the 115 avatars, 40 are adult models and 73 models are related to a profession, such as a firefighter, nurse, or athlete. The avatars are rendered with the 3D program Blender, which allows automatic script based posture manipulation. For our purposes, avatars whose neck or head are completely or partially covered by clothing or headgear were excluded, since it makes an algorithmic prediction of the head posture unreasonable. Finally, we consider a selection of 94 avatars, from which a data set consisting of 9239 renderings (samples) of these models is generated, resulting in around 98 images per avatar.

In each rendering, the head was rotated around the two specified centers, where each of the five rotation angles comes from a normal distribution with $\mu = 0$ and $\sigma = 20$. The two rotation centers are marked in Figure 2. The rotation always follows the ZYX Euler angles, while rotating around the upper center first, followed by a rotation around the lower center. Thus, the indications of the rotation angles represent ZYX Euler angles, insofar as one builds the sum of the two angles each for pitch and roll. From an anatomical standpoint, we consider a differentiation of the rotation center (upper or lower rotation center) around the yaw rotation as unreasonable. Therefore, only one angle is used for the rotation of the yaw. Examples from the data set are visualized in Figure 3.

There is always a difference between the virtual world and the real world. Although, we want to train a network that is only based on virtual data, it should still able to generalize on ”real” images. Domain randomization (Tobin et al., 2017) can reduce the visual inconsistency between the rendered images and real photos of subjects. With this technique, some visual properties of the image are varied. As a result, the trained network should be able to generalize regardless of these variations. It is thus more likely that such a robust network will also be able to generalize to real world data. In generating the data set described here, the varied parameters are background color, lighting conditions, and camera position. A transfer of the trained network to real image data remains a research question to be tested in the future. In this virtual-only test, however, we can already investigate how robust the network is in the face of these randomizations.

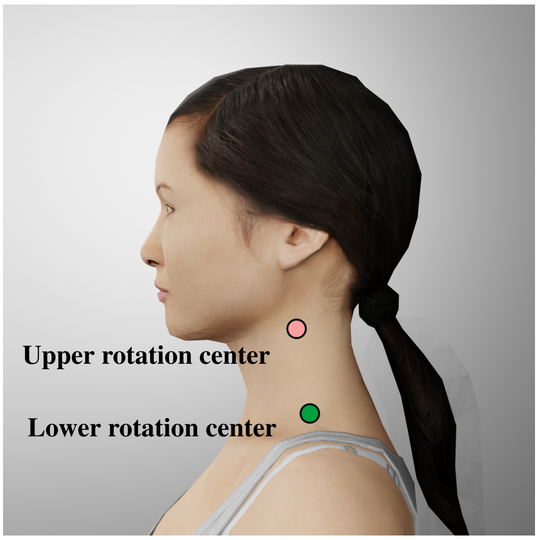


Figure 2: Locations of the two centres of rotation.

2.2 Artificial Neural Network and Training Process

The ANN for our purpose is an EfficientNetB2 (Tan and Le, 2019), pretrained on ImageNet (Deng et al., 2009), implemented in PyTorch. For a more meaningful evaluation, several training runs were performed in which the images of a single avatar were withheld from training for test purposes and trained with all the others. The mean value of this leave-one-out cross evaluation represents the result. Excluding one avatar at a time ensures that the training and test data are not too similar to each other.

For training, a ReduceLROnPlateau scheduler with an initial learning rate of 0.001 and the general-purpose layer-wise adaptive large batch optimize (LAMB) (You et al., 2019) for the Adam-Optimizer were used. For the loss function, the mean squared error was chosen. During training time, the data

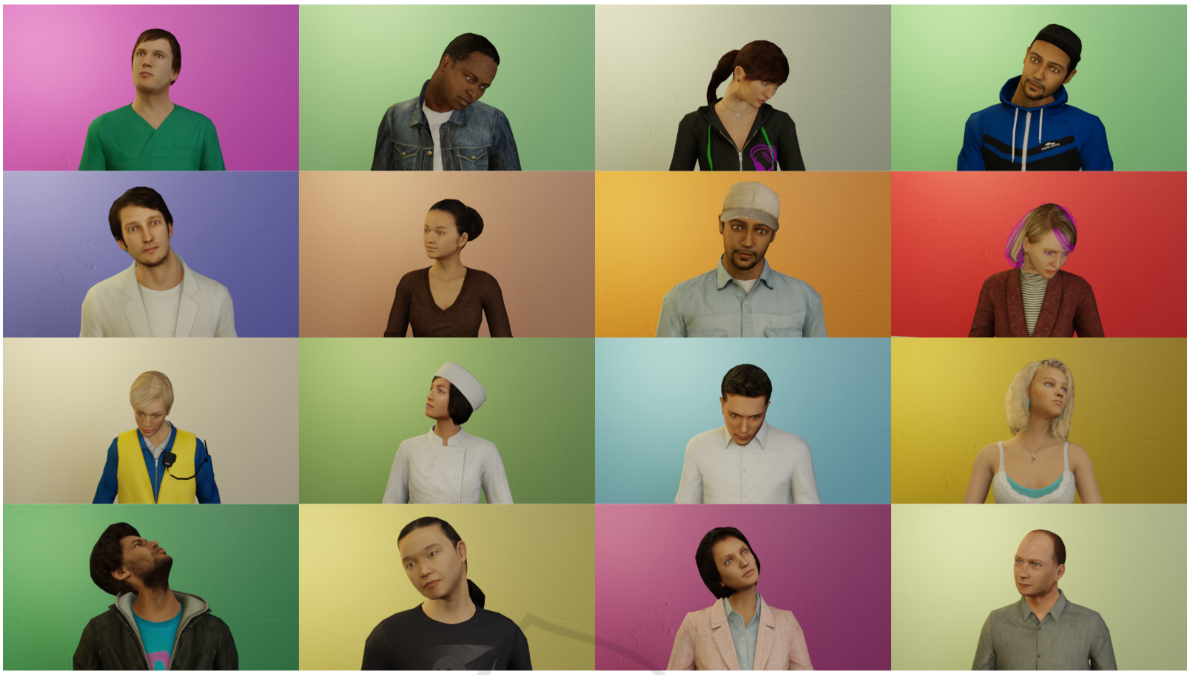


Figure 3: Avatars with random head poses, based on rotations, defined by the five angles, in a visually randomized setting (background color, lightning conditions, camera position).

was randomly augmented by adjusting the brightness, hue, and contrast and a Gaussian Blur was applied with a chance of 50% on each image. Furthermore, the RGB-values were normalized using the means and standard deviations from the ImageNet data set. Since the learning rate is adaptive according to the ReduceLROnPlateau rule, it depends on the course of the MSE. The MSE in this concern is computed on a validation set, which consists of the renderings of three randomly selected avatars that do not appear in the training data set.

To evaluate the predictions, we calculate the mean absolute error (MAE) of the angle predictions. This is calculated independently for each of the five angles. Since the rotation angles are taken from a normal distribution $\mathcal{N}$ with

$$\mathcal{N}(\mu,\sigma) = \mathcal{N}(0,20) \tag{1}$$

we assume that it is sufficient to know about the absolute error without considering large angles of more than 180°, for each of the angles. The absolute error is calculated as

$$\mathrm{AE} = ||\theta_{\mathrm{pred}} - \theta_{\mathrm{true}}||, \quad \text{and} \tag{2}$$

$$\mathrm{MAE} = \sum_{i=1}^{N} ||\theta_{\mathrm{pred}} - \theta_{\mathrm{true}}||/N. \tag{3}$$

3 RESULTS

The average MAE over all leave one out cross predictions calculates to the following values in Table 1.

Table 1: Summary of results, average MAE in degree [°] of the Leave-One-Out cross evaluation.

Yaw	Pitch	Roll	Pitch (neck)	Roll (neck)
1.79	6.50	4.06	6.55	3.95

Since the six rotation angles are taken from a probability distribution with $\mu = 0$ and $\sigma = 20$, it would be trivial to always assume an angle of 0°, which would result in the error being 20° in average. We can show that the ANN on the data set with 8842 training samples performs much better, but differently well for the respective angles. The prediction of the pitch around the upper and lower rotation center turned out to be the most error-prone. This can be explained by the fact that a rotation around the pitch rotation plane shows up visually mainly as a movement of depth, which might be more difficult to detect than roll or yaw movements due to the frontal view. Predictions of the yaw, on the other hand, turn out to be most accurate. One reason could be that the two yaw rotations (around the upper and lower rotation center) are added together and the network only needs to predict the sum. The resulting decrease

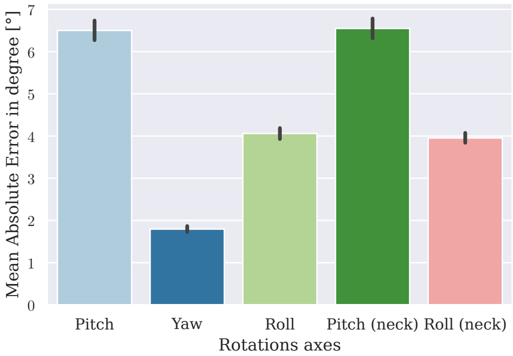


Figure 4: Mean absolute error per rotation.

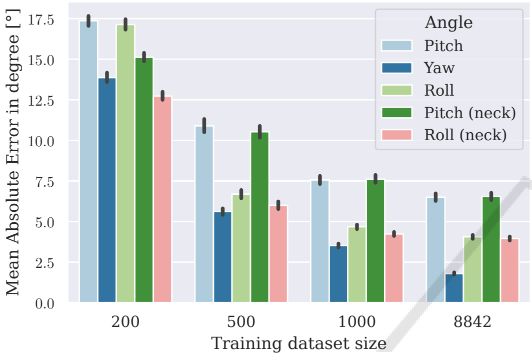


Figure 5: Mean absolute error in dependence of the training data set size.

in the complexity of the rotation possibly might lead to the better result.

According to RQ2 (see section 1), we want to explore the influence of the amount of training samples, regarding the MAE on the test data set. Figure 5 visualizes the MAE in dependency of the size of subsets of the whole data set, where we trained with 200, 500, 1000 and the full 8842 samples. A clear trend is noticeable, that more training data leads to better results, but with diminishing gain. The trend underlines the non-triviality of the problem.

Encountering the third research question (RQ3), we further want to investigate how the error develops in dependence of the values of the true rotation angles. As expected, the trend shows that larger angles lead to a larger absolute error. However, the course of the error in dependence of the true rotation angle does not behave the same for the five angles. The error of the pitch rotation around the lower rotation center increases considerably more in comparison to the other angles. We explain this behavior (besides statistical uncertainties) with the inability of the 3D models of the avatars to represent a strong rotation of the pitch (around the lower rotation angle), which is to be distinguished from the pitch rotation around the upper center of rotation. In addition, it could also be due to the fact that we only use a picture from the front as training data. In contrast to the poor trend of the error in pitch rotations at large angles, the roll (around both centers of rotation) is comparably well detectable, which possibly indicates that this rotation is very well representable by the 3D models.

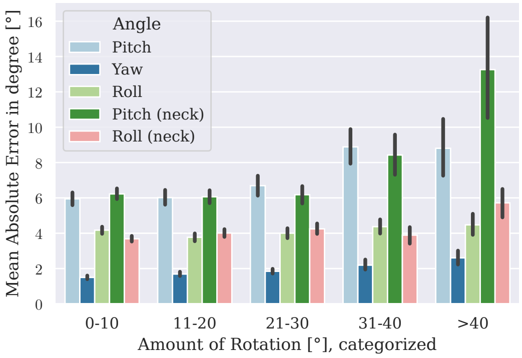


Figure 6: Mean absolute error in dependence of the absolute true rotation angle.

4 CONCLUSIONS

In this work, we can show that an ANN, in our case the EfficientNetB2, can make non-trivial predictions on the data set with synthetically generated avatars with complex head poses. Having only a front view towards the avatar, we show, that the error predicting the respective angles varies considerably strong. This difference can on the one hand be due to the fact that the perspective of the renderings plays a role, since the camera only faces directly at the front of the avatars. On the other hand, limitations of the 3D models of the avatars can also be the cause, which maybe can not visualize rotations of the yaw, pitch and roll equally well. One way to investigate this would be to generate a new data set containing renderings from three perspectives. This could be used to investigate whether the image perspective causes the errors. At the same time, it would also be an insight that could play a role for clinicians in the evaluation of dystonia, that multiple perspectives are important for being accurate. Such insight may be considered, when we develop a dataset, consisting of real subject images to be annotated by experts.

By synthesizing the data set of avatars, we can show to what extent a ANN can generalize with respect to our randomization's in the data set, which are the camera position, lighting conditions, and background color. In further research, we would also like to test the ANN on photographs of real subjects.

We see the use cases of such an algorithm in the assessment of the severity of cervical dystonia, where the head posture is not only determined by three degrees of freedom. Previous head pose estimation algorithms do not address the necessary complexity of head poses that are symptoms of cervical dystonia. By generating a large data set to train the neural network, we see the possibility to address the lack of large data sets from real subjects with good quality annotations regarding head posture. While the network may have achieved good results on the synthesized avatar data, this may not necessarily translate to real-world situations where the input data may be more varied and complex. Testing the network on real images will allow us to assess how well it can handle these variations. However, the extent to which this generalization can also be applied to real subject images is a question that we want to address on the basis of this work in the future.

ACKNOWLEDGEMENTS

We acknowledge financial support by the BMBF (01ZZ2007). Additionally we acknowledge the AI-Lab Lübeck, that has provided us with computational resources for training the neural networks.

REFERENCES

Albanese, A., Bhatia, K., Bressman, S. B., DeLong, M. R., Fahn, S., Fung, V. S. C., Hallett, M., Jankovic, J., Jinnah, H. A., Klein, C., Lang, A. E., Mink, J. W., and Teller, J. K. (2013). Phenomenology and classification of dystonia: A consensus update. *Mov. Disord.*, 28(7):863–873.

Ansari, S. A., Nijhawan, R., Bansal, I., and Mohanty, S. (2021). Cervical dystonia detection using facial and eye feature. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 43–48.

Boyce, M. J., Canning, C. G., Mahant, N., Morris, J., Latimer, J., and Fung, V. S. C. (2012). The toronto western spasmodic torticollis rating scale: reliability in neurologists and physiotherapists. 18(5):635–637.

Comella, C. L., Leurgans, S., Wuu, J., Stebbins, G. T., Chmura, T., and and The Dystonia Study Group (2003). Rating scales for dystonia: a multicenter assessment. *Mov. Disord.*, 18(3):303–312.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Fanelli, G., Dantone, M., Gall, J., Fossati, A., and Van Gool, L. (2013). Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458.

Finsterer, J., Maeztu, C., Revuelta, G. J., Reichel, G., and Truong, D. (2015). Collum-caput (COL-CAP) concept for conceptual anterocollis, anterocaput, and forward sagittal shift. *J. Neurol. Sci.*, 355(1-2):37–43.

Gonzalez Franco, M., Ofek, E., Pan, Y., Antley, A., Steed, A., Spanlang, B., Maselli, A., Banakou, D., Pelechano, N., Orts-Escolano, S., Orvalho, V., Trutoiu, L., Wojcik, M., Sanchez-Vives, M. V., Bailenson, J., Slater, M., and Lanier, J. (2020). The rocketbox library and the utility of freely available rigged avatars. *Frontiers in Virtual Reality*. TECHNOLOGY AND CODE ARTICLE Front. Virtual Real. — frvir.2020.561558.

Hempel, T., Abdelrahman, A. A., and Al-Hamadi, A. (2022). 6D rotation representation for unconstrained head pose estimation.

Nakamura, T., Sekimoto, S., Oyama, G., Shimo, Y., Hattori, N., and Kajimoto, H. (2019). Pilot feasibility study of a semi-automated three-dimensional scoring system for cervical dystonia. *PLOS ONE*, 14:e0219758.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world.

Valle, R., Buenaposada, J. M., and Baumela, L. (2021). Multi-Task head pose estimation in-the-wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(8):2874–2881.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2019). Large batch optimization for deep learning: Training BERT in 76 minutes.

Zhang, Z., Cisneros, E., Lee, H. Y., Vu, J. P., Chen, Q., Benadof, C. N., Whitehill, J., Rouzbehani, R., Sy, D. T., Huang, J. S., Sejnowski, T. J., Jankovic, J., Factor, S., Goetz, C. G., Barbano, R. L., Perlmutter, J. S., Jinnah, H. A., Berman, B. D., Richardson, S. P., Stebbins, G. T., Comella, C. L., and Peterson, D. A. (2022). Hold that pose: capturing cervical dystonia's head deviation severity from video. *Ann. Clin. Transl. Neurol.*, 9(5):684–694.

Zhou, Y. and Gregson, J. (2020). WHENet: Real-time fine-grained estimation for wide range head pose.

Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2015). Face alignment across large poses: A 3D solution.