

Seated Posture Estimation Based on Monocular Camera Images

Hitoshi Shimomae, Tsubasa Esumi, Noriko Takemura

Kyushu Institute of Technology, 680-4 Kawazu, Izuka City, Fukuoka, 820-0067, Japan

Email: shimomae.hitoshi210@mail.kyutech.jp, esumi.tsubasa157@mail.kyutech.jp, takemura@ai.kyutech.ac.jp

Abstract

Poor seated posture significantly strains the body, leading to symptoms like shoulder stiffness and back pain. While research on seated posture estimation using images has been active, many studies focus on extreme postures not typically seen in daily desk work. This study aims to estimate common postures, such as slouching, which are often experienced in everyday settings. Due to the lack of medical quantitative metrics for evaluating posture quality, we manually annotated some of our collected posture data and used semi-automatic annotation by SVM to build a dataset. Using this dataset, we trained deep learning models for posture estimation with different input data types: RGB images, silhouette images, and posture key points, and compared their performance.

Keywords: Posture Analysis, CNN, GNN

1. Introduction

The average sedentary time in Japan is approximately seven hours, ranking among the highest in the world [1]. Prolonged sedentary periods and poor seated posture can impose significant physical strain, leading to issues such as shoulder stiffness and lower back pain. Recently, development efforts have intensified on systems that promote posture improvement, such as smart chairs [6], [11], necessitating accurate estimation of seated postures as a foundational requirement. Many conventional studies on seated posture estimation target extreme postures, such as heavy lateral tilts or excessive backward bends, which are rarely seen in regular desk work [2], [7], [9], making them unsuitable for real-world application. Therefore, we aim to estimate everyday seated postures, including rounded shoulders and lordosis, using readily deployable and cost-effective RGB cameras.

In this study, we collect posture data while working at a desk using an RGB camera and motion capture to construct a dataset. For some of the collected data, we manually annotate three levels (good posture/slightly bad posture/bad posture) by multiple people, and then train a label estimation model using the motion capture data as input, and perform automatic annotation for the remaining data. Using the created dataset, we perform 3-class classification (good posture/slightly bad posture/bad posture) of sitting postures using a deep learning model that takes RGB images and silhouette images, as well as low-dimensional posture feature points extracted using YOLO [11], as inputs, and compare and discuss the accuracy of the different features of the input data.

2. Dataset Construction

This study targets the estimation of natural seated postures that can be commonly assumed in scenarios like desk work, as shown in Fig.1. Postures like slouching are

often unconsciously assumed and noticed only when pointed out by others. Therefore, we collected posture data from 30 subjects aged 18 to 24 (25 males and 5 females) during desk work tasks without instructing them to assume specific postures consciously. Additionally, we manually annotated a portion of the data and employed an SVM trained on this annotated data to automate the annotation of the remaining data, thus constructing a large-scale dataset.

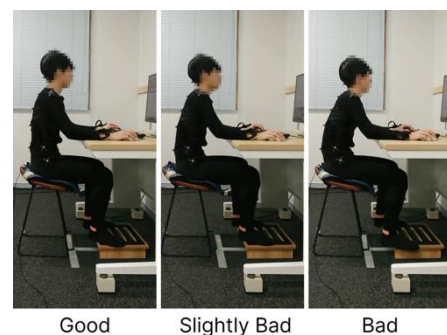


Fig.1 Examples of postures targeted in this study (collected posture data).

2.1. Data Collection Experiment

As a desk task, we set a task to create a report (Word, PowerPoint, etc.) while searching the Internet for a given theme. Each session was 15 minutes long, and we collected data from three sessions per subject, for a total of 45 minutes. RGB images (for posture estimation) were captured at 30 fps from three directions in front of, behind, and to the right of the subject using a Logitech C920n. Motion capture data (for annotation) was collected at 100 fps using NaturalPoint's OptiTrack Flex 13. The motion capture marker positions were set as shown in Fig.2 so that the degree of curvature of the spine and the tilt of the pelvis could be seen.

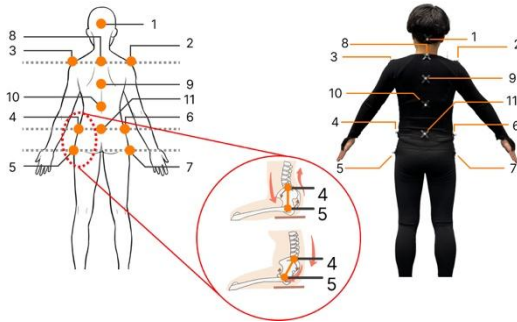


Fig.2 Marker positions for motion capture.

2.2. Annotation

Due to slow postural changes during desk work, we down-sampled the collected image data to 15 fps for posture estimation. Thus, the number of images for posture estimation was approximately 1,215,000 (= 30 subjects × 45 minutes × 60 seconds/minute × 15 fps). For supervised learning, it is necessary to assign true posture labels to each image data. However, since no medical indices exist to evaluate posture quality, a quantitative annotation based on rules is challenging. Additionally, manually annotating all data is time-consuming and labor-intensive, making it impractical. Hence, this study manually annotated part of the data and trained an estimation model for annotation, using the trained model to perform automated annotation on the remaining data.

2.2.1. Manual Annotation

Sampled image data every 25 seconds was used for manual annotation. Due to missed frames in some motion capture data, 2,955 images were extracted. Using images captured from the side view, where the spinal and pelvic tilt is most visually apparent, annotators labeled the data in three stages: good, slightly poor, and poor. Five annotators labeled each data piece to account for potential subjective judgment variability, and an integrated posture label was determined from these labels. Specifically, scores of good, slightly poor, and poor posture, denoted as S , were set to 1, 0, and -1, respectively. The average score \bar{S} from five annotators determined the posture labels L as in Equation 1. In addition, the variance, σ^2 , was calculated, and data with a variance value of less than 0.5 was treated as valid annotation data.

$$L = \begin{cases} \text{Good Posture} & \text{if } \frac{1}{3} \leq \bar{S} \leq 1, \\ \text{Slightly Poor Posture} & \text{if } -\frac{1}{3} < \bar{S} < \frac{1}{3}, \\ \text{Poor Posture} & \text{if } -1 \leq \bar{S} \leq -\frac{1}{3}. \end{cases} \quad (1)$$

2.2.2. Learning Estimation Model and Automatic Annotation

Motion capture data is measured with less than 1 mm error for each marker's 3D coordinates, containing precise information on spinal and pelvic tilt, thus the input for the estimation model comprised 33 dimensions of features, being 3D coordinates (x, y, z) of each marker point (11 points). Support Vector Machine (SVM) [4] was employed for the 3-class classification of good posture, slightly poor posture, and poor posture. RBF kernel was used, and hyperparameters C and γ were optimized via grid search, resulting in $C = 20$ and $\gamma = 0.1$.

Performance evaluation results upon dividing manually annotated data into training and test data are shown in Table 1. Table 1 indicates that although the estimation accuracy of slightly poor posture is somewhat inferior compared to other postures, it is overall accurately estimated. Using the trained SVM model, automated annotation was performed on the remaining data. Class-wise data distribution is shown in Table 2. Through these procedures, all image data were labeled as truth data, allowing them to be utilized for learning and evaluating posture estimation models.

Table 1: Annotation Estimation Performance.

	Precision	Recall	F1-score
Good Posture	0.98	0.94	0.96
Slightly Poor Posture	0.86	0.94	0.89
Poor Posture	0.96	0.91	0.94

Table 2: Breakdown of Annotation Labels.

	Manual Annotation	Automatic Annotation
Good Posture	1,181	487,579
Slightly Poor Posture	1,039	388,479
Poor Posture	704	355,884

3. Posture Estimation Method and Evaluation Experiment

Utilizing input data such as RGB images from the side, silhouette images, and Yolo's posture estimation features, we conducted a 3-class classification of seated postures. Table 3 summarizes the experiments.

Each model utilized the Cross Entropy Loss as the loss function and Adam for optimization. The epoch number was set to 1000, and training was terminated through EarlyStopping if the F1 score decreased for 10 consecutive epochs.

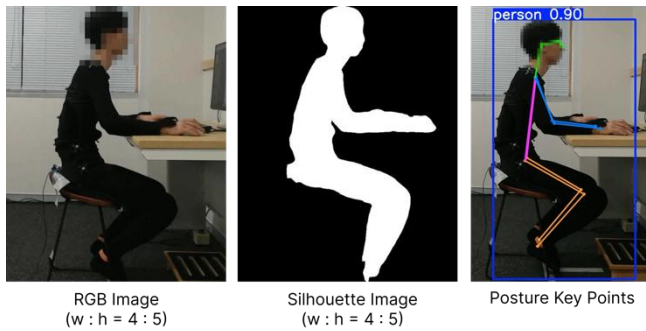


Fig.3 Image Preprocessing.

【Silhouette Image】 : Extracted using DeepLabV3 with ratio adjustments to 5:4 height-to-width based on the subject's height.

【RGB Image】 : Similarly extracted subject regions using DeepLabV3, adjusting the width to a 5:4 aspect ratio.

【 Posture Key Points 】 : Acquired 17 key points' 2D coordinates with YOLO's pose estimation results.

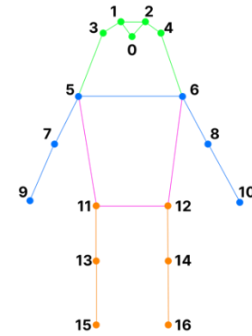


Fig.4 YOLO Pose Estimation Key Points.

0: Nose, 1: Left Eye, 2: Right Eye, 3: Left Ear, 4: Right Ear 5: Left Shoulder, 6: Right Shoulder, 7: Left Elbow, 8: Right Elbow 9: Left Wrist, 10: Right Wrist, 11: Left Hip, 12: Right Hip 13: Left Knee, 14: Right Knee, 15: Left Ankle, 16: Right Ankle

Table 3: Experiments

	Method	Input Data
Experiment 1	ResNet50	RGB Images
Experiment 2	ResNet50	Silhouette Images
Experiment 3	GCN	YOLO Pose Estimation Results

3.1. RGB Images

Using RGB images as input, ResNet50 [5], was utilized for 3-class classification of good posture, slightly bad posture, and bad posture. The RGB image of the subject, segmented using DeepLabV3 [3] as shown in Fig.3, was used as input. For model training, the hyperparameters were as follows: learning rate of 0.001, batch size of 1024, and a weight decay parameter of 0.01. Adam's parameters β_1 , β_2 , and ϵ were set to 0.95, 0.995, and $1e-06$, respectively. For the final layer of ResNet50, the existing fully connected layers were replaced to correspond to the 3-class classification. During training, the cross-entropy loss was used, and the softmax function was used to classify the outputs into 3 classes.

3.2. Silhouette Images

Silhouette images were used as input with ResNet50 [5] for 3-class classification of good posture, slightly bad posture, and bad posture. Silhouette images were generated using DeepLabV3 [3] as outlined in Fig.3. The training hyperparameters were set to the same values as the RGB images. The input layer for ResNet50 was modified to have a channel count of 1 to accommodate silhouette images, and the final layer was replaced for a 3-class classification. During training, cross-entropy loss

and the softmax function were employed to distinguish the outputs into 3 classes.

3.3. Posture Key Points

Using the posture feature points from YOLO's pose estimation as input, GCN [8] was employed for 3-class classification of good posture, slightly bad posture, and bad posture. The results of YOLO's pose estimation consisted of 17 key point 2D coordinates, as illustrated in Fig.4. The hyperparameters for training were configured as follows: learning rate of 0.01, batch size of 512, and weight decay parameter of 0.0. Adam's parameters β_1 , β_2 , and ϵ were set to 0.9, 0.999, and $1e-08$, respectively. The input layer of GCN was set to have a channel count of 17 based on YOLO's pose estimation results. A fully connected layer for 3-class classification was added at the end, and cross-entropy loss was employed for training, classifying the output into 3 classes through a softmax function.

3.4. Evaluation Method

For performance evaluation, subjects were divided into six groups of five subjects each, performing a 6-fold cross-validation. Each group was set as test data once, with another group set as validation data, and the remaining four as training data, continuing the cycle until all groups served as test data once. While automated annotation data was used for training and validation, only manually annotated data was used in testing to account for potential estimation errors. Random undersampling ensured a balanced class distribution for training data across groups. Models were evaluated using those from epochs exhibiting the highest validation data performance, with accuracy and macro-F1 score as evaluation metrics.

3.5. Results and Discussion

The results of the sitting posture estimation are shown in Table 4. When using RGB images as input, when using silhouette images as input, and when using the results of YOLO's pose estimation as input, the average accuracy was 0.729, 0.696, and 0.528, respectively, and the average F1 score was 0.591, 0.589, and 0.459, respectively. The highest Accuracy and F1 score were achieved when the input was an RGB image. The reason for the higher Accuracy and F1 score than for the silhouette images is thought to be that they have the same background and clothing, and there is noise due to false positives in the semantic segmentation of the silhouette images. On the other hand, when the results of Yolo's pose estimation were used as the input, the accuracy and F1 score were the lowest. This is thought to be because, as shown in Fig. 5, the results of YOLO's pose estimation do not express the posture features necessary for estimating sitting posture, such as the curvature of the back. However, the silhouette images contain noise due to false positives in the semantic segmentation, and the RGB images contain noise due to background, color information, clothing, etc. Therefore, it is thought that it would be ideal to use data with less noise and more detailed posture features, rather than sparse posture feature points like YOLO.



Fig. 5 Example of Poor Representation of Posture Features by YOLO Pose Estimation Results.

Table 4: Seated Posture Estimation Results

	Ave-Acc	Ave F1
RGB Image	0.729	0.591
Silhouette Image	0.696	0.589
Posture Key Points	0.528	0.459

4. Conclusion

In this study, we constructed a dataset for estimating natural sitting postures using monocular RGB images. Using the dataset we created, we performed 3-class classification of sitting postures using a deep learning model with RGB images and silhouette images, and low-dimensional posture feature points extracted using Yolo as inputs, and compared the accuracy of each model according to the differences in the characteristics of the

input data. As a result, the model using silhouette images showed the highest accuracy, while the model using posture feature points showed the lowest accuracy. However, since RGB images and silhouette images contain information that is noise and not necessary for posture estimation, it is thought that using posture feature points with appropriate features is effective. Therefore, in the future, the challenge will be to propose a method for extracting posture feature points that are optimal for sitting posture. It is thought that posture feature points that are optimal for sitting posture can be extracted using 3D human body models and 3D human skeletal models.

References

- Bauman, A., Ainsworth, B. E., Sallis, J. F., Hagströmer, M., Craig, C. L., Bull, F. C., Pratt, M., Venugopal, K., Chau, J. and Sjörström, M.: The Descriptive Epidemiology of Sitting: A 20-Country Comparison Using the International Physical Activity Questionnaire (IPAQ), *American Journal of Preventive Medicine*, Vol. 41, No. 2, pp. 228–235 (2011).
- Chen, K.: Sitting Posture Recognition Based on Open-Pose, *IOP Conference Series: Materials Science and Engineering*, Vol. 677, No. 3, p. 032057 (2019).
- Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation (2017).
- Cortes, C. and Vapnik, V.: Support-vector networks, *Machine Learning*, Vol. 20, No. 3, pp. 273–297 (1995).
- He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016).
- Journal Editor, P., Iqbal, M. J., Megha, Vyas, N., Kansal, H. and Kulshreshth, D.: A study of Smart Chair for Monitoring of Sitting Behaviour, *PRATI- BODH*, No. RACON 2023 (2023).
- Kapoor, R., Jaiswal, A. and Makedon, F.: Light-Weight Seated Posture Guidance System with Machine Learning and Computer Vision, *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments, PETRA '22*, New York, NY, USA, Association for Computing Machinery, p. 595–600 (2022).
- Kipf, T. N. and Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks, *International Conference on Learning Representations* (2017).
- Li, L., Yang, G., Li, Y., Zhu, D. and He, L.: Abnormal sitting posture recognition based on multi-scale spatiotemporal features of skeleton graph, *Engineering Applications of Artificial Intelligence*, Vol. 123, p. 106374 (2023).
- Otoda, Y., Mizumoto, T., Arakawa, Y., Nakajima, C., Kohana, M., Uenishi, M. and Yasumoto, K.: Census: Continuous posture sensing chair for office workers, *2018 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, pp. 1–2 (2018).
- Ultralytics inc.: Ultralytics YOLOv8, <https://docs.ultralytics.com/ja/models/yolov8>. (Accessed on 12/12/2024).

Authors Introduction

Mr. Hitoshi Shimomae



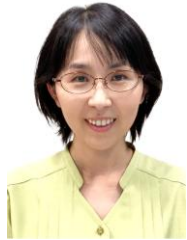
He received his Bachelor's degree in Computer Science and Systems Engineering in 2024 from the Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology in Japan. He is currently a master student at Kyushu Institute of Technology, Japan

Mr. Tsubasa Esumi



He received his Bachelor's degree in Computer Science and Systems Engineering in 2023 from the Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology in Japan. He is currently a master student at Kyushu Institute of Technology, Japan

Dr. Noriko Takemura



She received B.S., M.S., and Ph.D. degrees in engineering from Osaka University, Japan, in 2006, 2007, and 2010, respectively. She is currently an associate professor at Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology. Her current research interests include human-centric intelligent systems