ENHANCING HAND POSE ESTIMATION FOR GLOVED HANDS THROUGH COLOR DISTRIBUTION MATCHING AND U-NET SEGMENTATION

HIDETO MIURA[†], MASAKAZU MORIMOTO[†]

[†]Graduate School of Engineering, University of Hyogo, 2167 Shosha, Himeji, Hyogo 671-2280, Japan
E-MAIL: ei25p023@guh.u-hyogo.ac.jp, morimoto@eng.u-hyogo.ac.jp

Abstract:

Hand pose estimation accuracy significantly degrades in glove-wearing environments due to shifts in color distribution. This study investigates a lightweight preprocessing method that transforms glove images to a bare-hand appearance, enabling skeletal extraction using existing pretrained models. The approach combines glove region segmentation using the U-Net architecture with color space correction, specifically hue-saturation-value (HSV) and CIE-Lab histogram matching. Evaluation using three common glove types showed that detection rates improved from a maximum of 27% to over 90%, with Lab matching demonstrating the most stable performance. This method provides a practical strategy for automatic annotation and adapting models to glove-wearing conditions without the computational cost of retraining or complex generative models.

Keywords:

Hand pose estimation; Glove-wearing environment; U-Net; Color space correction; MediaPipe Hands; MMPose

1. Introduction

In medical and manufacturing settings, there is an increasing demand for technologies that quantitatively evaluate workers' hand and finger movements, with expected applications in workflow improvement, skill assessment, and the construction of hands-on training environments. In recent years, advancements in hand pose estimation techniques, represented by MediaPipe Hands and MMPose, have enabled highly accurate hand landmark estimation from monocular RGB images.

On the other hand, in medical and manufacturing fields, workers typically wear gloves for safety and hygiene management. However, existing hand pose estimation models are primarily trained on images of bare hands, and it has been reported that their detection accuracy degrades substantially when gloves are worn [1]. One of the main causes of this issue is that gloves alter the texture information and color distribution of the skin, resulting in a distribution shift from the training data. The presence of multi-colored structures and low-contrast regions further increases the likelihood of missed detections and false detections.

To address this problem, additional training using glove-wearing images is considered effective. However, such additional training requires skeletal landmark annotations for glove images. Given the diversity of glove types, colors, and lighting conditions, large-scale data collection and manual annotation demand considerable effort. Therefore, the primary objective of this study is to first transform glove images into an appearance closer to bare hands, automatically extract skeletal information using existing pose estimation models to generate pseudo-labels, and then use these pseudo-labeled data for additional training on glove-wearing images.

As the first step toward this goal, this study investigates the extent to which preprocessing that converts glove images into hand-like images can improve skeletal information extraction accuracy in existing models. Specifically, glove regions are extracted using U-Net and subsequently subjected to color space transformation. By applying color correction to the extracted glove regions to approximate the color distribution of bare hands, the processed images are then input into existing pose estimation models. Detection rate and landmark localization error are quantitatively evaluated using MediaPipe Hands and MMPose to verify the effectiveness of the proposed method. Unlike conventional approaches that require additional training with manually annotated glove images, this study demonstrates that appearance-level domain adaptation through simple color distribution correction can substantially recover detection performance without modifying pretrained pose estimation models. This finding provides practical insights into deploying hand pose estimation systems in real-world glove-wearing environments.

2. Dataset

In this study, three types of gloves were evaluated in order to assess skeletal detection performance under glove-wearing conditions. The selected gloves were blue vinyl gloves, black–green mixed work gloves, and white-gray work gloves. These types were chosen to represent gloves typically used in medical and manufacturing environments.

Five participants were recruited, and recording was primarily conducted from a frontal viewpoint. For three of the participants, additional recordings were performed from a lateral viewpoint. Furthermore, for one participant, additional images were intentionally captured with the back of the hand facing the camera. This configuration enabled evaluation of robustness against variations in viewpoint and hand orientation.

The videos were recorded using an Apple iPhone 16 camera at a resolution of 1920 × 1080 and a frame rate of 30 fps. In the recorded videos, participants continuously performed a sequence of hand gestures, including a fist (rock), a two-finger sign (scissors), an open hand (paper), a thumbs-up sign, and an OK sign. The recorded videos were divided into individual frames, and each frame was used as a still image in the dataset. A dataset comprising 1,000–1,200 frame images was prepared for each glove type and used for evaluating skeletal detection rates.

FIGURE 1. Evaluated glove types: (a) black–green mixed, (b) blue vinyl, and (c) white-gray

3. Glove Region Extraction Using U-Net

3.1. Training

In this study, semantic segmentation using U-Net was adopted for glove region extraction. A total of 60 images were used for training: five frontal-view images each from three participants, five lateral-view images each from two participants, and five additional images from one participant showing the back of the hand. This dataset configuration incorporated variations in viewpoint and hand orientation.

During training, geometric data augmentation was applied, including horizontal flipping, translation, scaling, and rotation. A U-Net model with a ResNet-34 encoder was employed, and the encoder was initialized with ImageNet-pretrained weights.

3.2. Accuracy Evaluation Using Intersection over Union (IoU)

For evaluation, a total of 15 images from participants not included in the training dataset were used to assess generalization performance. Segmentation accuracy was evaluated using the mean Intersection over Union (IoU).

The mean IoU was 0.9156 for blue gloves, 0.8703 for black–green mixed gloves, and 0.7496 for gray gloves. The relatively lower IoU for gray gloves was mainly attributed to occasional misclassification of the participant's face mask, whose color was visually similar to that of the glove material. In such cases, parts of the mask region were incorrectly segmented as glove regions.

It should be noted that this misclassification occurred outside the hand area and therefore did not influence the subsequent pose estimation experiments, in which color correction and landmark detection were applied specifically to the hand region. For reference, when these mask-induced misclassification cases were excluded, the mean IoU for gray gloves increased to 0.8498.

Overall, the segmentation accuracy was sufficient to provide reliable glove-region extraction for the purpose of color distribution correction, as evidenced by the consistent improvements in pose estimation performance reported in Section 5.

4. Color Space Correction Methods

One of the primary reasons for the degradation in skeletal detection accuracy under glove-wearing conditions is the shift in skin color distribution, which causes a deviation from the feature distribution assumed by existing models. Therefore, in this study, color space correction was applied to the glove regions extracted by U-Net in order to transform them into a color distribution closer to that of bare hands.

The proposed histogram-matching approach was selected for its computational efficiency and minimal data requirements, offering a practical alternative to more complex generative approaches that require extensive training datasets.

This section describes three methods: hue (H channel) modification, HSV histogram matching, and Lab histogram matching. In this study, these methods are denoted as follows: Original (unprocessed), H (hue modification), HSV (HSV histogram matching), and Lab (Lab histogram matching).

4.1. Hue (H Channel) Modification (H)

As the simplest correction approach, we investigated a method in which the image is first converted to the HSV color space, and only the hue (H) component is modified while preserving the saturation (S) and value (V) components.

An analysis of the hue distribution was conducted using hand regions extracted from five pre-acquired bare-hand reference images captured from a single specific participant under identical lighting conditions. By utilizing a universal reference distribution derived from one individual, the proposed method eliminates the need for per-user calibration, which significantly enhances its practical utility for deployment across diverse users. While this study focuses on identical lighting conditions for reference capture, the integration of automated color constancy algorithms represents a promising direction for enhancing robustness in dynamic environments. The pixel values from these five images were aggregated to analyze the overall hue distribution. The results revealed that the H values tend to concentrate within a specific range.

Therefore, the H values of the extracted glove regions were constrained to this range, thereby transforming the color distribution to approximate that of human skin.

4.2. HSV Histogram Matching (HSV)

As a more statistically grounded correction approach, histogram matching was applied in the HSV color space. First, hand regions were extracted from five pre-acquired bare-hand reference images captured from a single specific participant under identical lighting conditions. Using a single-user reference to construct the baseline histograms eliminates the need for per-user calibration, allowing the system to be used immediately by new users without prior data collection.

The pixel values of the H, S, and V components from these images were aggregated by concatenating all pixel samples to construct a single reference histogram for each channel. This aggregation increased the statistical robustness of the reference distributions.

The histogram of the extracted glove region was then transformed to match these reference histograms, thereby correcting the color distribution to approximate that of bare hands.

4.3. Lab Histogram Matching (Lab)

Furthermore, histogram matching was performed in the perceptually uniform CIE-Lab color space. After converting the RGB image into the Lab space, color correction was conducted by preserving the luminance component (L) while matching the chromatic components (a and b) to the reference distributions.

The Lab color space is designed based on human visual perception, and thus is expected to provide more stable and perceptually consistent color distribution correction.

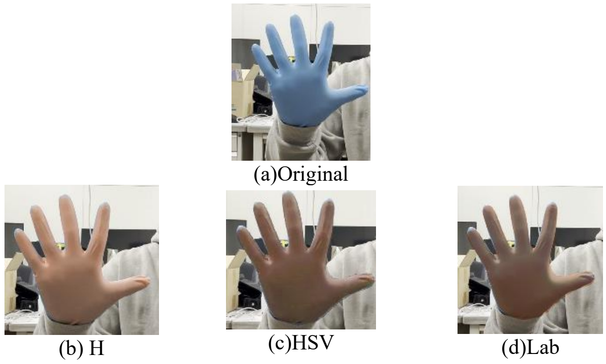

FIGURE 2. Appearance Transformation Example: (a) unprocessed image, (b) H: hue modification, (c) HSV: HSV histogram matching, and (d) Lab: Lab histogram matching.

5. Performance Evaluation

Under glove-wearing conditions, it was frequently observed that existing pose estimation methods failed to recognize the hand as a hand. In other words, a primary cause of performance degradation is that the hand region itself is not detected prior to landmark estimation. In such cases, evaluating only landmark localization error does not adequately reflect the impact of missed detections.

Therefore, this study adopts a two-stage evaluation protocol. First, the detection rate is evaluated to determine whether the hand is correctly detected. Subsequently, landmark localization error is assessed only for images in which detection is successful. This approach enables comprehensive verification of the proposed method from both detection performance and localization accuracy perspectives.

5.1. Detection Rate Evaluation

For the detection rate evaluation, frame images extracted from the recorded videos were used. The evaluation dataset comprised all five participants. Notably, this assessment included individuals whose data were not used for U-Net training (Section 3.1), confirming the method's ability to generalize to unseen users. The evaluation dataset consisted of 1,208 images of black–green mixed gloves, 1,065 images of blue gloves, and 1,134 images of gray gloves, for a total of 3,407 images.

MediaPipe Hands and MMPose were employed for evaluation. In MediaPipe Hands, a detection was defined as successful when landmark estimation results were returned. In MMPose, a detection was defined as successful when a bounding box corresponding to the hand region was detected. The detection rate was calculated as the ratio of successfully detected frames to the total number of frames.

TABLE 1 presents the detection rates (%) for each method and each glove type.

TABLE 1. Detection Rates of Each Method (MediaPipe / MMPose) [%]

Method	*Black and Green*	*Blue*	*Gray*
Original	27 / 2	14 / 4	18 / 9
H	70 / 45	91 / 91	48 / 43
HSV	87 / 93	86 / 88	89 / 96
Lab	90 / 93	89 / 88	90 / 97

For unprocessed images, detection rates remained low across all glove types. In particular, MMPose achieved only 2% detection for the black–green mixed gloves, indicating that, under glove-wearing conditions, the hand was frequently not recognized as a hand.

In contrast, when color space correction was applied, detection rates improved markedly. For example, in the case of black–green mixed gloves, the detection rate of MMPose increased from 2% in the unprocessed condition to 93% after applying Lab histogram matching, corresponding to an improvement of approximately 91%. Similarly, for blue gloves, the detection rate of MediaPipe Hands improved from 14% to 91%.

Although hue (H channel) modification alone led to some improvement, higher detection rates were achieved with HSV and Lab histogram matching. In particular, Lab histogram matching yielded detection rates of approximately 90% across all three glove types and demonstrated stable improvement for both models.

These results quantitatively demonstrate that color space correction is effective in mitigating the missed-detection problem that arises under glove-wearing conditions.

5.2. Landmark Localization Error Evaluation

Detection rate alone does not allow evaluation of the positional accuracy of the detected landmarks. Therefore, in this study, landmark localization error was evaluated using the Normalized Mean Error (NME).

A total of 180 images were used for evaluation. These consisted of 20 frontal-view images for each of five participants (100 images in total), 20 lateral-view images for each of three participants (60 images in total), and 20 additional images from one participant specifically showing the back of the hand. Manual annotation was performed on these images to obtain ground-truth coordinates for each landmark. The error for each landmark was defined as the Euclidean distance between the predicted coordinates and the ground-truth coordinates. To eliminate the influence of hand size variation, the error was normalized by the diagonal length of the hand-region bounding box computed from the ground-truth landmark coordinates.

In addition, to prevent missed detections or severe misdetections from being excluded from the evaluation, a penalty mechanism was introduced. If no prediction was returned, or if the estimated hand center position deviated significantly from the ground truth, an error equal to twice the bounding box diagonal length was assigned to all landmarks of the corresponding hand. This design corresponds to an NME value of 2.0.

The penalty factor was selected as a statistically sufficiently large value to represent a critical detection failure. Since the bounding box diagonal reflects the spatial extent of the hand region, assigning twice this length ensures that such cases are treated as severe errors beyond the hand scale while preserving scale invariance across different hand sizes. At the same time, a finite value was used to allow stable computation of the aggregated NME.

TABLE 2 presents the NME results computed over the entire dataset (each cell represents MediaPipe / MMPose).

TABLE 2. NME (ALL data) (MediaPipe / MMPose)

Method	*Black and Green*	*Blue*	*Gray*
H	0.668 / 1.180	0.286 / 0.317	1.145 / 1.025
HSV	0.355 / 0.299	0.332 / 0.400	0.299 / 0.264
Lab	0.283 / 0.287	0.305 / 0.379	0.276 / 0.253

From the TABLE, it can be confirmed that the application of color space correction led to an overall reduction in error. However, even after correction, the best-performing condition (Lab for gray gloves) still exhibited a relative error of approximately 25% with respect to the bounding box diagonal length, indicating that highly accurate localization was not fully achieved.

One possible reason for this is that, in images captured from a lateral viewpoint, finger overlap and self-occlusion frequently occur, making landmark estimation more difficult. The increased error observed in lateral-view images can be attributed to the inherent geometric challenge of self-occlusion, where invisible joint positions follow a multimodal distribution with multiple plausible solutions, as reported in prior work [7]. This suggests that the remaining errors are primarily caused by geometric ambiguity rather than residual appearance-based domain shift.

In such situations, conventional regression-based methods, which estimate a single deterministic solution, are prone to larger localization errors. This interpretation is consistent with the increased errors observed in our lateral-view experiments.

Therefore, the increased error observed under lateral-view conditions in this study is considered to stem not primarily from color distribution issues, but rather from structural difficulties caused by viewpoint changes and geometric self-occlusion. Typical examples of these landmark localization shifts in lateral-view images are shown in FIGURE 4.

The main objective of this study is to mitigate the impact of appearance-based domain shift—specifically, changes in color distribution due to glove wearing—on existing skeletal detection methods. Accordingly, in order to evaluate the pure effect of the color correction methods, a re-evaluation was conducted using 120 images (NO_SIDE), excluding the 60 lateral-view images. The results are shown in TABLE 3.

TABLE 3. NME (NO_SIDE data) (MediaPipe / MMPose)

Method	*Black and Green*	*Blue*	*Gray*
H	0.578 / 1.120	0.142 / 0.059	1.067 / 0.910
HSV	0.215 / 0.118	0.136 / 0.089	0.145 / 0.098
Lab	0.127 / 0.062	0.168 / 0.090	0.144 / 0.081

(a) front (b) side

FIGURE 3. Difference in Viewpoint

(a) Black and Green (b) Blue (c) Gray

FIGURE 4. Examples of landmark localization shifts in lateral-view images corrected by Lab histogram matching: (a) black–green mixed, (b) blue vinyl, and (c) white-gray gloves.

From TABLE 3, it can be confirmed that excluding the lateral-view images resulted in a substantial reduction in NME. In particular, a marked improvement in accuracy was observed for MMPose. When HSV or Lab histogram matching was applied, the NME for black–green mixed gloves, blue gloves, and gray gloves all fell below 0.10, achieving a relative error of less than 10% with respect to the bounding box diagonal length. These results indicate that, under frontal-view conditions, combining the proposed method with MMPose enables highly accurate landmark estimation with a relative error below 10%.

In addition to NME, we evaluated landmark localization accuracy using the Percentage of Correct Keypoints (PCK) under the NO_SIDE condition. A landmark was regarded as correctly localized when its normalized error was less than 0.05 relative to the diagonal length of the ground-truth hand bounding box, corresponding to a strict 5% hand-scale tolerance.

TABLE 4 summarizes the PCK@0.05 results for frontal-view images.

TABLE 4. PCK@0.05 (NO_SIDE data) (MediaPipe / MMPose)

Method	*Black and Green*	*Blue*	*Gray*
H	0.468 / 0.186	0.645 / 0.687	0.313 / 0.334
HSV	0.595 / 0.550	0.643 / 0.635	0.663 / 0.652
Lab	0.664 / 0.631	0.632 / 0.652	0.670 / 0.646

As shown in TABLE 4, under frontal-view conditions, PCK@0.05 exceeded 0.60 for all glove types when HSV or Lab histogram matching was applied to MMPose. This indicates that more than 60% of landmarks were localized within a strict 5% tolerance of the hand scale.

Although the NO_SIDE subset reduces geometric difficulty by excluding lateral-view images, the evaluation still represents a domain-shift scenario in which the pretrained models were not optimized for glove-wearing conditions. Achieving PCK values above 0.60 under such conditions demonstrates practically meaningful localization accuracy rather than trivial performance.

These findings suggest that distribution-level color normalization is sufficient to recover stable landmark localization under frontal-view settings, even without model retraining.

Based on the above evaluations, it can be concluded that distribution-level color normalization effectively mitigates the appearance-based domain shift caused by glove wearing in existing skeletal detection methods.

In the detection rate analysis, conditions that exhibited low detection rates in unprocessed images were substantially recovered through the application of hue adjustment and histogram-based matching. Furthermore, the NME and PCK evaluations demonstrated that, under frontal-view conditions, reliable landmark localization can be achieved without retraining the underlying pose estimation models.

These results suggest that introducing color space normalization as a preprocessing step provides a practical and lightweight strategy for stabilizing hand pose estimation in glove-wearing environments.

6. Conclusions

In this study, we proposed a preprocessing method that transforms glove images to resemble bare-hand images in order to address the degradation in accuracy of existing hand pose estimation methods under glove-wearing conditions. The primary objective of this research is to accurately extract skeletal information from the transformed images using existing pretrained models and to utilize the obtained results for automatic annotation and additional training of glove-wearing datasets.

By combining glove region extraction using U-Net with HSV and Lab histogram matching, the detection rate, which remained at a maximum of approximately 27% for unprocessed images, improved to over 90%. In particular, Lab histogram matching demonstrated stable performance improvements across all three types of gloves.

In the landmark localization error evaluation, self-occlusion in lateral-view images was identified as a major source of error. However, when lateral-view data were excluded, MMPose achieved a relative error of less than 10%.

These results indicate that the proposed method serves as an effective preprocessing approach for adapting existing skeletal detection models to glove-wearing environments without modifying the models themselves, and offers a practical solution toward realizing automatic annotation. To ensure the quality of generated pseudo-labels, future implementations will incorporate a filtering mechanism based on confidence scores to select only high-reliability frames for model retraining. Future work will focus on validating the effectiveness of additional training using pseudo-labels generated by the proposed method, addressing self-occlusion, and extending the approach to a wider range of glove materials and lighting conditions.

References

[1] A.W. Kiefer, D. Willoughby, R. P. MacPherson, R. Hubal, and S. F. Eckel, "Enhanced 2D Hand Pose Estimation for Gloved Medical Applications: A Preliminary Model," *Sensors,* Vol. 24, No. 18, Article 6005, 2024.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Proceedings of MICCAI*, pp. 234–241, 2015.

[3] F. Zhang, V. Bazarevsky, A. Vakunov, G. Sung, C. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," arXiv preprint arXiv:2006.10214, 2020.

[4] MMPose Contributors, "MMPose: OpenMMLab Pose Estimation Toolbox and Benchmark," arXiv preprint arXiv:2009.12328, 2020.

[5] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Pearson, 2018.

[6] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color Transfer between Images," IEEE Computer Graphics and Applications, Vol. 21, No. 5, pp. 34–41, 2001.

[7] Q. Ye and T.-K. Kim, "Occlusion-aware Hand Pose Estimation Using Hierarchical Mixture Density Network," in Proc. ECCV, pp. 103–118, 2018.