Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169743436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Multiscale Feature Fusion Network for Monocular Complex Hand Pose Estimation

Zhi Zhan[1] and Guang Luo[2]

[1]Guangdong Engineering Polytechic
[2]South China Normal University

October 16, 2023

Abstract

Hand pose estimation based on a single RGB image has low accuracy due to the complexity of the pose, local self-similarity of finger features, and occlusion. A multiscale feature fusion network (MS-FF) for monocular vision gesture pose estimation is proposed to address this problem. The network can take full advantage of different channel information to enhance important gesture information, and it can simultaneously extract features from feature maps of different resolutions to obtain as much detailed feature information and deep semantic information as possible. The feature maps are merged to obtain the hand pose results. The InterHand2.6M dataset and Rendered Handpose Dataset (RHD) are used to train the MS-FF. Compared with the other methods (which can estimate interacting hand poses from a single RGB image), the MS-FF obtains the smallest average error of hand joints on RHD, verifying its effectiveness.

Multiscale Feature Fusion Network for Monocular Complex Hand Pose Estimation

Zhi Zhan , Guang Luo

Hand pose estimation based on a single RGB image has low accuracy due to the complexity of the pose, local self-similarity of finger features, and occlusion. A multiscale feature fusion network (MS-FF) for monocular vision gesture pose estimation is proposed to address this problem. The network can take full advantage of different channel information to enhance important gesture information, and it can simultaneously extract features from feature maps of different resolutions to obtain as much detailed feature information and deep semantic information as possible. The feature maps are merged to obtain the hand pose results. The InterHand2.6M dataset and Rendered Handpose Dataset (RHD) are used to train the MS-FF. Compared with the other methods (which can estimate interacting hand poses from a single RGB image), the MS-FF obtains the smallest average error of hand joints on RHD, verifying its effectiveness.

Introduction: Hand pose estimation aims to identify and localize key points of human hands in images, and it has a wide range of applications in virtual reality (VR) and augmented reality (AR) [1]. Methods based on deep learning have obvious advantages over traditional methods, both in processing speed and prediction accuracy. However, owing to the complexity and diversity of the photographic environment, such as hand shapes and occlusion, the robustness of hand pose estimation methods is low.

Hand pose estimation methods can be categorized as either depth- [2-5,15] and RGB-based [6-14,16]. Most methods rely on depth images, such as Chen et al. [2] extracted effective joint features through the initially estimated hand pose as guiding information, then fused the joint features of the same fingers, and finally regressed the hand pose by fusing the finger features. However, the method of connecting five fingers and the palm at the same time can cause loss in accuracy. According to Zhang et al. [4] made full use of the information between the adjacent joints of the fingers to estimate the depth coordinates. Then, 2D hand

joint estimation and depth estimation of a part of the hand joints were used as the bootstrap information to obtain depth coordinates of all the hand joints.

Deep images are often limited by the application context, so RGB images have been used for hand pose estimation. Simon et al. [6] estimated 2D hand poses from multi-view images and extended them to the 3D space. However, this method could not estimate hand pose from a single RGB image. Spurr et al. [7] used RGB images to train an encoder-decoder model to estimate the complete 3D hand pose with different inputs. However, the method did not make full use of the hand structure. Yang et al. [9] learned the hand pose and hand images by a disentangled variational autoencoder to achieve image synthesis and hand pose estimation, but the disentangled process may lose useful information. Since most datasets only have single hand sequences, estimating complex gestures is relatively difficult. For this reason, Moon et al. [16] constructed a dataset containing single and interacting hand sequences. Additionally, the InterNet model was proposed to estimate hand poses by a single RGB image. Due to the influence of occlusion, the method cannot estimate complex hand pose well. However, the edge information in the hand pose estimation is usually ignored, due to the presence of occlusion, this information is especially important for extracting the information of the occluded part. Simultaneously, because the fingertip is a small object, it is relatively difficult to recognize the joint at the fingertips. To address this, a robust Multi-Scale Feature Fusion Network (MS-FF) is presented in this paper. The main contributions of this method are as follows:

1. MS-FF more accurately estimates hand poses in an RGB image and better copes with complex application scenarios, so as to better deal with difficult-to-recognize joints and inaccurate gesture recognition in occlusion scenes;
2. Channels contain different implicit information. We need to focus on the information that is more important for recognizing gestures. A channel conversion module adjusts the weights of channels to enhance important information;
3. Fingertips occupy a small percentage of an image, and are relatively difficult to identify. A global regression module generates different resolutions with rich semantic information, to better utilize image edge details and deep information, which is important in estimating finger poses;
4. The global regression module may not accurately identify occluded joints. A local optimization module is designed with deeper information in the feature map. It fuses all level feature maps, correcting joints that do not return to the correction position, for better application to the occlusion scene;

Method:

A. Multiscale Feature Fusion Network

Gesture pictures usually contain complex detailed features. A strong correlation between fingers and joints is present. Therefore, the use of a single feature for hand pose estimation tends to ignore diverse feature information, which makes accurate extraction of more gesture information difficult. Fig. 1 shows the proposed MS-FF, whose purpose is to estimate the hand pose through a single RGB image. Feature maps of different resolutions are extracted from RGB images through the ResNet50 module. Feature maps are fed into the channel conversion module to explicitly learn the dependencies between channels, so as to enhance important information and downplay minor information. Because the level of feature information depends on the resolution of a feature map, the global regression module obtains high-resolution feature maps containing more semantic information, and these are separately input in the local optimization module to extract deeper information. The Gaussian heatmap of hand joints () is obtained to improve the spatial generalization ability of the model, and thus obtain more accurate joint locations. We take the feature map with the smallest resolution from the channel conversion module, through which the handedness () and relative depth information between the wrist joints () are obtained. The above results are combined to estimate the hand pose,

, (1)

, (2)

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169743436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

where equation (2) represents the result of gesture estimation, and and are the camera inverse projection and inverse affine transformation, respectively.

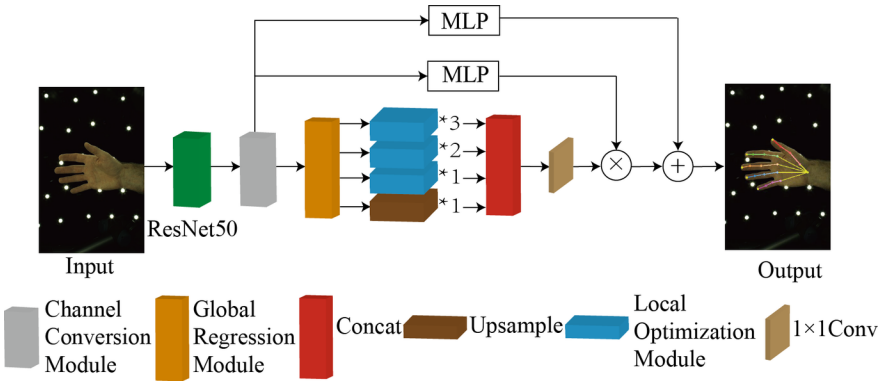

Fig. 1 *Overview of architecture.*

B. Channel Conversion Module

Each channel of the feature map contains different feature information. To make better use of this information, the relationship between channels of the feature map is modeled explicitly. Higher weights are assigned to channels with higher semantic characterization ability, so as to improve the sensitivity of the model to important feature information. The structure of channel conversion module is shown in Fig. 2.

Hosted file

`image6.emf` available at https://authorea.com/users/673583/articles/672325-multiscale-feature-fusion-network-for-monocular-complex-hand-pose-estimation

Fig. 2 *Structure of channel conversion module.*

The channel conversion module has aggregation and excitation stages. Global feature information of spatial dimension is aggregated into a channel descriptor of dimension by average pooling. The c-th element calculation of vector A is

, (3)

where and are the height and width, respectively, of the feature map; and is the pixel in the c -th channel. The average feature of each channel in the feature map is calculated by aggregation.

To fully utilize the aggregated feature information, the excitation operation captures the dependencies between channels. The aggregated information learns the inter-channel dependencies through the fully connected layer. The weight vector with dimension is obtained by the sigmoid function, and can characterize the importance of each channel. The weight vector is multiplied by the original feature map to obtain the reassigned feature map, which can enhance important information and weaken the minor information. The dependency between the channels is

, (4)

where is the calculation of channel weights, and are the weight matrices of the two fully connected layers, is the sigmoid function, and is the ReLU function. The channel information of the feature map is recalibrated as

, (5)

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169744436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

where is the feature map after reassigning channel weights, and is channel-wise multiplication between the weight vector and feature vector .

C. Global Regression Module

The ResNet50 module produces feature maps with different resolutions. High-resolution, low-level feature maps contain less semantic information but rich spatial detail information, while low-resolution, high-level feature maps have rich semantic information and less spatial detail information. To fully exploit the feature information of different dimensions, the low- and high-resolution feature maps are combined by vertical and horizontal paths. The vertical path obtains the high-resolution feature map by upsampling the spatially low-resolution feature map. Then, 1×1 convolution is used to reduce the number of channels in the low-level feature map, so as to obtain a feature map with the same dimension as the corresponding longitudinal path feature. The horizontal path fuses the two feature maps (Fig. 3). This pyramidal structure allows feature maps of different resolutions to contain more semantic information, enabling the network to learn richer feature information.

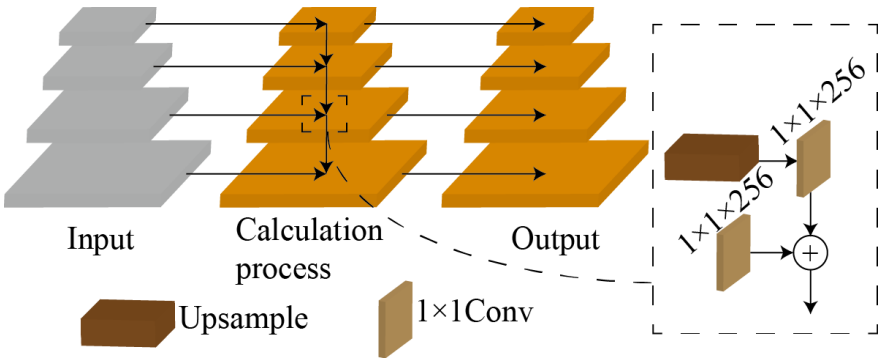

Fig. 3 *Structure of global regression module.*

Feature maps are obtained by the channel conversion module. and have high spatial resolution but low semantic information, while and have more semantic information but low spatial resolution. In addition to obtaining rich hand feature information, the fusion of feature maps can obtain detailed information, such as that of fingertips and masked edges. To fuse the feature information, feature maps in different dimensions are subjected to dimensionality reduction, so that their channels can be unified under the same dimension,

, (6)

, (7)

where V_k is the feature map obtained by dimensionality reduction, U_k is the feature map obtained by upsampling, R_1 is the convolution operation with a 1×1 convolution kernel, is the ReLU function, and$B$ is the upsampling operation of bilinear interpolation, which calculates the corresponding points in the new image by the four adjacent points as

, (8)

, (9)

. (10)

Equations (8) and (9) are linear interpolation operations in the$x$ -direction, and equation (10) is a linear interpolation operation in the y -direction. , , , and are points in the original image with coordinates , , , and , respectively. and are added to fuse feature information of different spatial resolutions. The calculation method is

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169743436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

. (11)

D. Local Optimization Module

To reduce errors generated by the global regression module, a local optimization module addresses the inaccuracy of predicting the joint position under occlusion. This can extract deeper information from feature maps obtained by the global regression module.

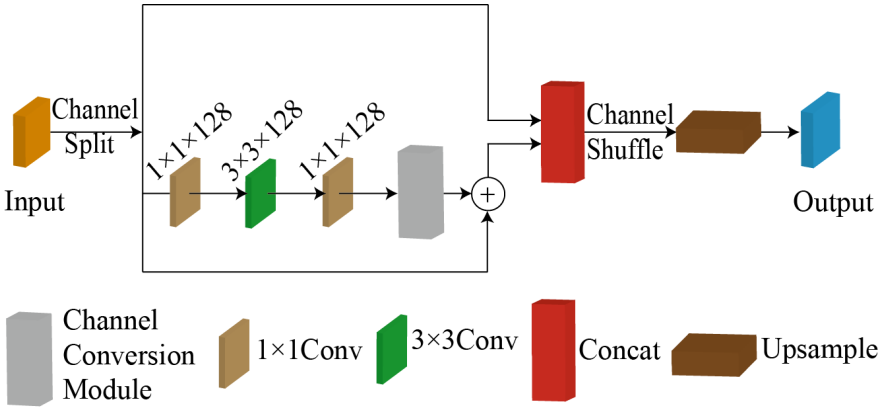

Fig. 4 *Structure of local optimization module.*

The input information is divided into two branches by the "channel split" operation (Fig. 4). The feature maps are processed separately through two paths; one is not processed, and the other has 1 × 1, 3 × 3, and 1 × 1 convolution kernels and can extract deep semantic information. The channel conversion module explicitly models the dependencies between the channels, which can enhance important information. Residual connectivity solves the problem of network degradation and improves representational capability. The outputs of the two paths are spliced to ensure that the channel dimension remains unchanged. The "channel shuffle" operation disrupts the order of the channels to improve the efficiency of information transmission and promote information fusion. Finally, the upsampling operation of bilinear interpolation is used to obtain a high-resolution feature map.

Four feature maps of different resolutions are taken from the global regression module. The same dimensional feature maps are obtained by the local optimization module,

, (12)

, (13)

where is the local optimization module and denotes upsampling. Let . Then , , , and denote the feature maps at the 1/4, 1/8, 1/16, and 1/32 scales, respectively, of the original image. The result in (13) represents the processing times of the above four feature maps by the local optimization module, i.e., , , , and . At this time, the four feature maps have the same dimension, and the "concat" operation is performed as

. (14)

The 2.5D Gaussian heatmap of the joints of the hand obtained by 1 × 1 convolution is

. (15)

Experimental Results and Analysis: Datasets RHD and InterHand2.6M were used to evaluate the performance of the proposed method. The PyTorch framework was used for training. The hand image was resized to 256 × 256 and input to the network. In the experiment, the batch size was set to 16. The network was trained

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169744436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

for 20 epochs with an NVIDIA 3090 GPU. The initial learning rate was set to 0.0001 and reduced by a factor of 10 at the 15th and 17th epochs to optimize the output of the network.

Hosted file

`image64.emf` available at https://authorea.com/users/673583/articles/672325-multiscale-feature-fusion-network-for-monocular-complex-hand-pose-estimation

Fig. 5 Mean per joint position error of interacting hand sequences on various test sets.

Different methods were used to test interacting hand pictures. We averaged the error of the left and right hand joints as the error of each joint point. As seen in Fig. 5, it was more difficult to predict the joint points near the fingertips than those near the palm. For all joints, the average errors of our method were lower than those of the compared methods. Fig. 6 shows the hand pose estimation results of PoseNet, InterNet, and the MS-FF. Since most joints are flexible, occlusions will be present when gestures interact, so it is more complicated to estimate hand poses through a single RGB picture. As seen in Fig. 6, our results are better than the others.

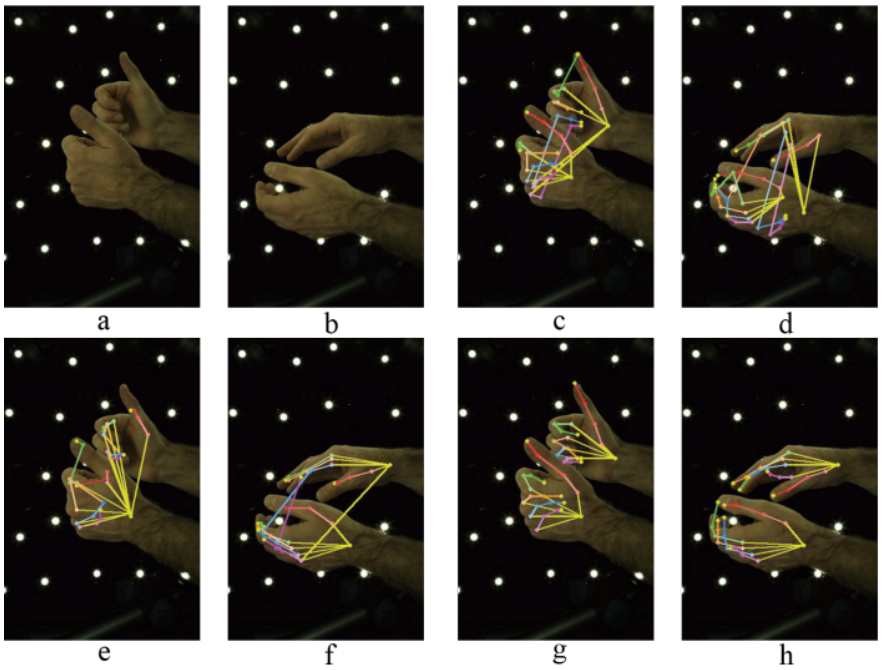

Fig. 6 *Prediction results of different methods on the test set: the orignial RGB images (a-b), the result obtained by the PoseNet method (c-d), the result obtained by the InterNet method (e-f), and the result obtained by our method (g-h), respectively.*

Table 1: Results of Different Methods on RHD.

Methods	GT S	GT H	EPE
PoseNet(2017)	Y	Y	30.42
Chen(2018)	Y	Y	24.20
Yang(2019)	Y	Y	19.95
Spurr(2018)	N	N	22.53
InterNet(2020)	N	N	20.89
Song(2022)	N	N	20.58
MS-FF	N	N	**20.21**

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169743436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169743436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Table 1 compares different methods on RHD, where EPE is the average error of hand joints, and GT H and GT S indicate handedness and scale of the hand, respectively. It can be seen that Spurr et al. [7] and Yang et al. [9] required additional input at test time, achieving lower joint errors, while our method could obtain low errors without ground-truth information during testing.

Conclusion: We proposed an MS-FF for monocular visual hand pose estimation. To effectively process the detailed information of occluded edges and fingertips, the network can extract information of different levels from feature maps of different resolutions to more accurately estimate hand poses. A channel conversion module adjusts the weights of channels. To make full use of both the edge detail characteristics of the images and deep semantic information, a global regression module fuses feature maps of different resolutions. An optimization procedure corrects some joints that are not returned to the correct position. Higher accuracy and robustness were achieved using the proposed method. Experiments verified the effectiveness of the MS-FF.

Acknowledgments: This work was supported by the National Natural Science Foundation of China under Grant 61601213, and Special Innovative Projects of General Universities in Guangdong Province under Grant 022WTSCX210.

Zhi Zhan (Guangdong Engineering Polytechnic,Zhan Zhi, China)

Guang Luo (South China Normal University, Luo Guang, China)

E-mail: luoguang_arts@163.com

References

1. H. Yang, C. Wang, B. Jiang, et al: 'Visual Perception Enabled Industry Intelligence: State of the Art, Challenges and Prospects', IEEE Transactions on Industrial Informatics, 2021, **17** , (3), pp. 2204-2219.

2. X. Chen, G. Wang, H.Guo, C. Zhang: 'Pose Guided Structured Region Ensemble Network for Cascaded Hand Pose Estimation', Neurocomputing, 2017, **395** , pp. 138-149.

3. S. Liu, G. Wang, P. Xie and C. Zhang: 'Light and Fast Hand Pose Estimation From Spatial-Decomposed Latent Heatmap', IEEE Access, 2020,**8** , pp. 53072-53081.

4. X. Zhang, S. Huang, Z. Ye: 'Accurate 3D hand pose estimation network utilizing joints information', Signal Processing:Image Communication, 2021, **90** .

5. Amir Rasouli, Iuliia Kotseruba: 'PedFormer: Pedestrian Behavior Prediction via Cross-Modal Attention Modulation and Gated Multitask Learning', IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 9844-9851.

6. T. Simon, H. Joo, I. Matthews, Y. Sheikh: 'Hand Keypoint Detection in Single Images Using Multiview Bootstrapping', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4645-4653.

7. A. Spurr, J. Song, S. Park and O. Hilliges: 'Cross-Modal Deep Variational Hand Pose Estimation', IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 89-98.

8. L. Ge et al.: '3D Hand Shape and Pose Estimation From a Single RGB Image', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10825-10834.

9. L. Yang and A. Yao: 'Disentangling Latent Hands for Image Synthesis and Pose Estimation', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9869-9878.

10.H.-X. Song, T.-J. Mu, R. R. Martin: 'Joint Hand and Object Pose Estimation from a Single RGB Image using High-level 2D Constraints', Computer Graphics Forum, 2022, **41** , 7, pp. 383-394.

Posted on 16 Oct 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.169744436.63648882/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

11. L. Chen, S. Lin, Y. Xie, et al: 'DGGAN: Depth-image Guided Generative Adversarial Networks for Disentangling RGB and Depth Images in 3D Hand Pose Estimation', IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 400-408.

12. D. Kong, H. Ma, Y. Chen, et al: 'Rotation-invariant Mixed Graphical Model Network for 2D Hand Pose Estimation', IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1535-1544.

13. L. Fan, H. Rao, W. Yang: '3D Hand Pose Estimation Based on Five-Layer Ensemble CNN', Sensors, 2021, **21** , pp. 649-664,2021.

14. X. Wang, J. Jiang, Y. Guo, et al: 'CFAM:Estimation 3D hand poses from a single RGB image with attention', Applied Sciences, 2020,**10** , pp.618-635.

15. Y. Wang, L. Chen, J. Li, et al.: 'HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation', IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5664-5673.

16. G. Moon, S. Yu, H. Wen, et al: 'InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image', European Conference on Computer Vision, 2020, pp. 548-564.