Research on Somatosensory Interaction Based on Convolutional Neural Network

Hui Tang, Qing Wang, Hong Chen, Hao Guo
College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
wangqingait@cau.edu.cn

Abstract. Somatosensory interaction is an important part of human-computer interaction. The core of somatosensory interaction must accurately obtain the 3D spatial information of the human body. This article uses an RGBD camera to acquire both color and depth images. We perform 2D human pose estimation in color images using hourglass network with residual structure. At the same time, we use the sparse feature points to align the color map with the depth map in order to accurately map the detection results in color images to the corresponding depth images. The experimental results show that this method can accurately and quickly obtain the 3-dimensional posture information of the human body, which is an important guarantee for somatosensory interaction.

1. Introduction
Somatosensory interaction is an important application scene for human-computer interaction. Especially in the past two years, deep learning is widely used in the field of human-computer interaction. In this paper, color and depth images are read by an RGBD camera, and the color image is subjected to 2D human body pose estimation by a convolutional neural network, then the returned result of 2D human pose estimation in color images using hourglass network with residual structure. At the same time, we use the sparse feature points to align the color map with the depth map in order to accurately map the detection results in color images to the corresponding depth images. The experimental results show that this method can accurately and quickly obtain the 3-dimensional posture information of the human body, which is an important guarantee for somatosensory interaction.

2. Human Pose Estimation
This part is divided into 3 sub-sections. First, the human target detection is performed to obtain the human proposals. The second step is to send the human proposals into the joint detection network for
2D human pose estimation, finally, align the color images with the depth images and the detected joints information in color images is mapped to the depth map to complete the 3D human pose estimation task.

2.1. Human target detection

We use the most advanced target detection net YOLOV3 as the human target detector. Considering that the deeper the convolutional neural network is, the more favorable it is to extract the high-level semantic information in the picture, the feature extraction backbone network of yolov3 are traditional convolution layer, pooling layer, and relu layer stacking, which are difficult to extract the higher layer semantics information, We choose Resnet50 as the feature extraction backbone network, and its structure is shown in Figure 1.

![Resnet50 network structure](image)

The structure introduces the idea of the residual structure to solve serious problems such as over-fitting and degradation caused by the deep neural network. It can extract higher-level semantic information, reduce the information transmission loss through the residual module jump layer connection, and enhance the network output expression ability.

2.2. Pose estimation

The understanding human pose is especially important for advanced tasks such as somatosensory interactions. In fact, for the associated task of pose estimation, the different joint points of the whole body are not the best recognition accuracy on the feature map of the last layer of the convolutional neural network. In the previous network structure of the 2D pose estimation, most of the convolution features of the last layer are used, which will result in the loss of information. For example, the arm may be easily identifiable on the feature map of the third layer, and the head is easier to identify on the fifth layer, so it is necessary to design a feature map for the feature map generated by the different layers.

Convolutional Pose Machine is a sequential convolutional architecture that expresses spatial information and texture information by designing a multi-stage network. The input of the latter stage uses the detection result of the previous stage. and at the same time, the local and global key information of the human body are utilized, which the spatial information, the texture information, and the central constraint are better integrated. The Convolutional Pose Machine architecture is shown in Fig.2.

The architecture generates a belief map in all stage $t \in \{1, 2, ..., T\}$. in the step $t$, the output of belief predicts the position $z_{i} = (x_{i}, y_{i})$ for every joint point $i$ at each belief map $P$, so the output of the stage $t$ is $b_{t}^{*}(x_{i}, y_{i})$. we place a Gaussian peak in the ground truth of each body part in order to create an ideal
belief map $\delta_p^i(x_i, y_i)$ in an image. At the same time, we use L2 distance to minimize the loss function of the stage $t$ by using the predicting results and the ideal belief map for each body part. Therefore, the loss function can be defined as below:

$$f_t = \sum_{p=1}^{P} \sum_{i=1}^{N} \left\| b_p^i(x_i, y_i) - \delta_p^i(x_i, y_i) \right\|^2$$  \hspace{1cm} (1)

The total loss can be obtained by adding the loss at each step $f_t$. This is shown:

$$F = \sum_{t=1}^{T} f_t$$  \hspace{1cm} (2)

In our work, we define the number of $N$ is 14 and the number of $T$ is 6. We normalize the human proposals generated by the human target detection network as 368x368. and then input to the Convolutional Pose Machine, through multi-stage convolution, pooling, and downsampling, while supervising the losses generated in each stage. Eventually, it can produce a good prediction for 2D human pose estimation.

![Architecture of Convolutional Pose Machine](image)

2.3. Color and depth image alignment

Due to the limitation of manufacturing technology, the color image read by the RGBD camera is not inconsistent with the depth image, which causes errors in the task of 3D human pose estimation. For example, the Microsoft KinectV2 color and depth images are 1920*1080 and 512*424, respectively. Although Microsoft provides tools for the alignment, there are still a lot of errors. Figure 3(a) is the result of directly reading color and depth images using the Kinect SDK. Seeing that there are still some errors. For Somatosensory interaction task, this existing error can cause the task to fail.

In order to achieve better alignment. In this paper, a calibration method based on sparse feature points is proposed. We match the feature points on the color image and the feature points on the corresponding depth map to obtain accurate conversion parameters. Feature points are matched before. We use the SURF algorithm with a constant scale to extract the feature points of the image. SURF is a highly improved SIFT calculation. After the feature points are obtained, the feature points are described by the descriptors of the scale-invariant feature transformation in order to get the depth vector and the description vector of each feature point $i$. We use the k-means clustering algorithm to match the description vectors of a color image and depth image.
The numbers of 1, 2, 3 represent respectively the color image, the depth image, and the fusion of color and depth image. (a) represent no alienation, (b) represent alienation with our method.

Although the feature points in the depth image and the color image are matched and compared by the Euclidean distance comparison, there are still some mismatched point pairs in the feature point set. The RANSAC algorithm [32] is used to remove the mismatch, which further improves the matching accuracy. Fig.3(b) is the result of matching using the above, the fusion effect is much better than without alignment.

3. Experiment and result
This method implements Nvidia GeForce GTX970, Intel Core i5-6402P@2.80GHZ, memory 8G, Ubuntu16.04 operating system, Kinectv2 device. The project is used to implement on the deep learning Pytorch 1.0 framework.

3.1. Experimental details
We introduce a Resnet50 residual structure as a feature extractor in the target detection YOLOV3 network. We use the PCK (Percentage of Correct Keypoints) metric to evaluate the performance of our method in 2D human pose estimation tasks. PCK represents the percentage of detections that fall within a normalized distance of the ground truth, for MPII dataset, the distance(PCKh@0.5) is normalized by a fraction of the head size.

| Methods          | Head | Neck | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | PCK  |
|------------------|------|------|----------|-------|-------|-----|------|-------|------|
| Carreira et al.  | 95.7 | 95.4 | 91.7     | 81.7  | 72.4  | 82.8| 73.1 | 66.4  | 81.3 |
| Newell et al.    | 98.2 | 96.8 | 94.7     | 91.2  | 87.1  | 90.1| 87.4 | 83.6  | 90.9 |
| Chu et al.       | 98.5 | 97.6 | 96.3     | 91.9  | 88.1  | 90.6| 88.0 | 85.4  | 91.5 |
| Yang et al.      | 98.6 | 97.9 | 96.7     | 92.5  | 88.7  | 91.1| 88.6 | 86.0  | 92.0 |
| Nie et al.       | 98.6 | 98.2 | 96.9     | 93.0  | 89.1  | 91.0| 89.0 | 86.2  | 92.4 |
| Ours(without Resnet50) | 98.7 | 98.0 | 96.2     | 92.3  | 90.6  | 90.2| 89.3 | 85.7  | 91.8 |
| Ours(+Resnet50)  | 99.0 | 98.5 | 97.2     | 95.4  | 93.0  | 90.7| 92.8 | 87.6  | 97.7 |

We conduct a comparison test with Carreira et al.[5], Newell et al.[6], Chu et al.[7], Yang et al.[8] and Nie et al.[9] on the MPII dataset. The experimental results are shown in Table 1.
From the comparison results, our method is improved compared to the best method at each joint point, with an average increase of 5.73% (97.7/92.4), which explains that the introduction of resnet50 into the yolov3 network has a positive impact on human pose estimation.

### Table 2: Quantitative analysis of the effects of resnet50 and alignment strategies

| Methods                  | PCK  | Average MPJPE |
|--------------------------|------|---------------|
| Resnet50 (Without Alignment) | 94.7 | 95.4          |
| Alignment (Without Resnet50) | 84.3 | 75.9          |
| Resnet50 + Alignment     | 94.7 | 53.8          |

In order to verify the effect of this alignment method on 3D human pose estimation results, the experimental comparison results are shown in Table 2. We performed experiments on the famous 3D human pose data Human 3.6. This dataset has a total of 15 themes composed of 3.6 million color images of the human joint points and corresponding depth maps. We use the field-recognized evaluation index MPJPE (mean per joint position error) to calculate the error when the detected 2D joint points are mapped to the depth map after aligning the depths of the root joints.

When using our alignment strategy, MPJPE is reduced from 95.4 to 53.8, which can largely map 2D joint point information to the corresponding depth map. Therefore, the use of the Resnet50 network and alignment strategy in this paper plays an important role in 3D human pose estimation results.

**Fig 4 Visualize the result of 3D human pose estimation by our method**

### 3.2. Visualization of results

Fig 4 shows our test results. Color images and depth images are read from the RGBD camera. First, the YOLOV3 generates a human proposal, and then uses Pose estimation to estimate the single pose for each human proposal. The detection results are mapped to the corresponding depth map according to the alignment strategy. Finally, the color map and the depth map are combined to obtain accurate 3D human body posture information to complete the somatosensory interaction task.
4. Conclusion
In this paper, we propose a novel method for performing somatosensory interaction tasks by combining color images and depth images. The main innovations have 3 points:

1) With the advantage of the RGBD camera, 2D human pose estimation is performed on the color image, and then the depth image is used to complement the depth information.

2) The YOLOV3 is used as a human detector and use Resnet50 network for human feature extraction to improve the accuracy of human target detection.

3) Use the sparse feature point matching method to align the color image and the depth image so that the points on the color image can be accurately mapped back to the corresponding depth image.

In general, the method we propose can get accurate 3D human pose information, which can ensure the execution of somatosensory interaction successfully.

Acknowledgments
This work was financially supported by the National Science and Technology Support Program “Integration Research on Innovation Design and Performance Technology of Exhibition Space” (2015BAH28F01).

References
[1] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. 2018.
[2] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2016.
[3] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[J]. 2016.
[4] Han J, Shao L, Xu D, et al. Enhanced computer vision with Microsoft Kinect sensor: a review.[J]. IEEE Transactions on Cybernetics, 2013, 43(5):1318-1334.
[5] Carreira J, Agrawal P, Fragkiadaki K, et al. Human Pose Estimation with Iterative Error Feedback[J]. 2015.
[6] Ning G, Zhi Z, He Z. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation[J]. IEEE Transactions on Multimedia, 2017, PP(99):1-1.
[7] Chu X, Yang W, Ouyang W, et al. Multi-Context Attention for Human Pose Estimation[J]. 2017.
[8] Peng X, Tang Z, Yang F, et al. Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation[J]. 2018.
[9] Nie X. Human Pose Estimation with Parsing Induced Learner[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.