Perspective Crop Based Egocentric Hand Pose Estimation via Fisheye Stereo Vision

Hyejin Hur[1,2], Seongmin Baek[1], Younhee Gil[1], Sangpil Kim[2]

[1]Electronics and Telecommunications Research Institute (ETRI), South Korea
[2]Korea University, South Korea

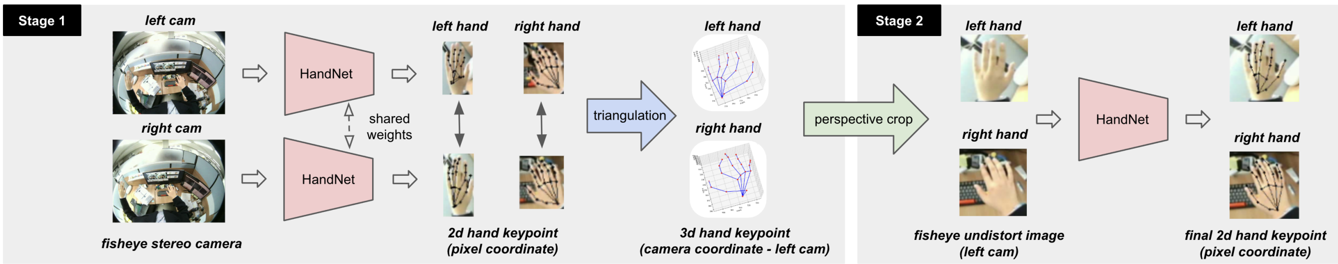

Figure 1: *Two-stage egocentric hand pose estimation pipeline*

Abstract

In this paper, we propose a method to improve the performance of hand pose estimation from egocentric view. To accurately capture hands moving within a wide range in daily activities, we mounted a fisheye stereo camera on a head mounted display to obtain wide-angle images from egocentric view. Our proposed two-stage method addresses the camera distortion introduced by this setup. The 2D hand keypoints estimated by stage-1 HandNet are converted into 3D hand keypoints through triangulation for perspective cropping. Stage-2 HandNet then predicts the final 2D hand keypoints from the undistorted hand crop image. To train stage-1 HandNet for perspective cropping, we built FisheyeEgoHAND dataset which consists of three categories of scenarios (separate hand, hand-hand, and hand-object) that reflect various hand interactions in an egocentric view. Through experiments, we demonstrated that two-stage 2D hand pose estimation outperforms one-stage approach without perspective cropping.

CCS Concepts

• ***Computing methodologies*** $\rightarrow$ *Computer vision;* ***Vision for robotics;***

1. Introduction

With the advancement of Virtual Reality (VR) and Augmented Reality (AR), the importance of hand pose estimation from egocentric view has also increased. To capture diverse hand movements in daily activities, fisheye cameras with a wide Field of View (FoV) are essential. While they offer a broader capture range than conventional cameras, they also introduce significant distortion. Thus, this leads to inaccurate hand pose estimation with fisheye cameras. Additionally, most existing hand pose estimation research has focused on third-person view, resulting in limited datasets for egocentric and fisheye-based hand pose estimation.

To address these challenges, we propose two-stage egocentric hand pose estimation method using a fisheye stereo camera. This method generates undistorted hand crop images through perspective cropping [SH22]. We employed U-Net [OR15] for HandNet and compared the performance of our model against SimpleBaseline [BX18] and HRNet [KS19].

2. Methods

As shown in Fig. 1, the proposed method works as follows.

- **2D hand pose estimation from two fisheye camera views**: Using images from fisheye stereo camera as input, stage-1 HandNet predicts the 2D hand keypoints of left/right hands in each view.
- **3D keypoints via triangulation**: The predicted 2D hand keypoints are used for triangulation with camera parameters. 3D

DOI: 10.2312/egp.20251016

hand keypoints are computed in the camera coordinate system with the left (or right) camera as the origin.

- **Perspective cropping**: Virtual cameras are generated from 3D keypoints, and undistorted hand crops are obtained via warping.
- **2D hand pose estimation from undistorted hand crop images**: Stage-2 HandNet predicts 2D hand keypoints from undistorted hand crop images obtained in the previous step.

Perspective cropping is a technique that generates a virtual camera for the Region of Interest (RoI) and crops the image accordingly. A virtual camera can be created when 3D points within the RoI are available. We set hand keypoints as the required points for perspective cropping and aimed to obtain 3D points from a stereo camera. To efficiently acquire 3D hand keypoints, we used triangulation for fast and accurate computation. For accurate depth values of these points, precise 2D hand pose estimation in stereo views is crucial for triangulation. Since we use a fisheye stereo camera, training 2D hand pose estimator on distorted images is essential.

Accordingly, we constructed a fisheye camera-based egocentric hand dataset, FisheyeEgoHAND (Fig. 2). To capture various hand interactions, we designed three categories of scenarios: separate hand (7 scenarios, 5,142 images), hand-hand (9 scenarios, 7,008 images), and hand-object (4 scenarios, 4,358 images), designing a total of 20 scenarios with 16,508 images. For each scenario, videos were recorded at 30 fps, with all frames extracted as images.

Figure 2: *Example of our FisheyeEgoHAND dataset*

3. Experiments

We adopted simplified version of U-Net as HandNet, which enables learning high-resolution representation features essential for hand pose estimation. We trained SimpleBaseline, HRNet, and HandNet for 200 epochs on the FisheyeEgoHAND dataset using a single RTX 4090. The batch size was 64 for SimpleBaseline and 128 for HRNet and HandNet, with IoU loss. We used SGD with a 0.1 learning rate and a multi-step scheduler at 40 and 100 epochs (gamma = 0.5). Fig. 3 (b) shows the perspective cropping result from stage-1 HandNet, demonstrating improved distortion correction over (a).

Since our proposed method employs perspective cropping, we generated a new ground truth derived from FisheyeEgoHAND. Specifically, the original 2D hand keypoints from FisheyeEgoHAND were projected onto virtual cameras obtained from perspective cropping. This dataset was then used to train stage-2 HandNet in the same way as stage-1. The results in Table 1 show that the two-stage hand pose estimation with perspective cropping achieved an EPE of 9.634, outperforming the one-stage estimation using the original fisheye camera image as input, which had an EPE of 12.626. Additionally, Fig. 3 (a) and (b) also show that the 2D hand pose estimation results improve with perspective cropping.

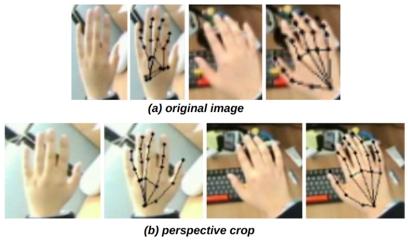

Figure 3: *Comparison of hand crops and 2D pose results*

Model	Avg EPE ↓	PCK ↑	AUC ↑
SimpleBaseline (r-50)	17.974	0.968	0.890
SimpleBaseline (r-50, w/ PC)	13.178	0.951	0.893
SimpleBaseline (r-101)	18.062	0.958	0.883
SimpleBaseline (r-101, w/ PC)	13.013	0.937	0.894
SimpleBaseline (r-152)	18.296	0.958	0.887
SimpleBaseline (r-152, w/ PC)	13.411	0.927	0.892
HRNet	17.918	0.971	0.890
HRNet (w/ PC)	12.724	0.937	0.896
HandNet (ours)	**12.626**	**0.984**	**0.917**
HandNet (ours, w/ PC)	**9.634**	**0.969**	**0.914**

Table 1: ***2D hand pose estimation on fisheye images and perspective cropped images.*** *"w/ PC" refers to "with perspective cropping".*

4. Conclusion and Future Work

We propose an effective method for egocentric hand pose estimation using a fisheye camera. We created FisheyeEgoHAND dataset and trained stage-1 HandNet on it for perspective cropping. The undistorted hand crops were then used to estimate the final 2D hand pose, outperforming one-stage approach. While 3D pose is more useful for VR and AR, our method estimates 2D pose. We aim to explore a distortion-robust 3D hand pose estimation approach.

5. Acknowledge

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZC1210, Research on hyper-realistic interaction technology for five senses and emotional experience, 100%]. This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City.

References

[BX18] BIN XIAO HAIPING WU Y. W.: Simple baselines for human pose estimation and tracking. In *ECCV* (2018). 1

[KS19] KE SUN BIN XIAO D. L. J. W.: Deep high-resolution representation learning for human pose estimation. In *CVPR* (2019). 1

[OR15] OLAF RONNEBERGER PHILIPP FISCHER T. B.: U-net: Convolutional networks for biomedical image segmentation. In *MICCAI* (2015). 1

[SH22] SHANGCHEN HAN PO-CHEN WU Y. Z. B. L. L. Z. Z. W. W. S. P. Z. Y. C. T. H. R. C. L. T. M. A. T.-H. Y. C. K. R. W.: Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia* (2022). 1