Commentary: Multi-person 3D pose estimation from a single image captured by a fisheye camera

Qiyao Zhou
Z5379852

1. Introduction

Modern digital photography and computer vision techniques are advancing with technology and the times, wide angle and fisheye cameras are involved in many real life scenarios, including but not limited to video surveillance, virtual reality and automotive applications. A fisheye image is an image taken by fisheye lens, which is extreme wide-angle lens that resembles fish's eye in order to access most photographic angles of view, so called the name "fisheye lens". Compared to wide-angle cameras, fisheye cameras are able to capture a greater amount of information, but with correspondingly greater distortion parameters. Based on the promising research gap in the field, this 2022 paper will implement the estimation of multi-person 3D poses from a single fisheye image .

There are three main problems to be solved to achieve this, firstly, due to the inherent characteristics of fisheye images, the distance from the center point affects the estimation of one's posture, secondly, the distance between the target portrait and the camera is not fixed, and finally, predicting 3D human joint positions with absolute depth is more challenging than root-relative 3D pose estimates due to the inherent depth and scale ambiguity.

In order to estimate multi-person 3D poses from a single image captured by a fisheye camera, a top-down approach is proposed, using HPoseNet and HRootNet to estimate the relative 3D pose and absolute depth of the root joint, respectively, and then a reprojection module to connect the two branches to obtain a 3D pose consistent with the 2D truth.

If the approach adopted by the authors in this paper is indeed effective, it will have a direct impact on the use of fisheye images, for example in video surveillance, where the use of a fisheye lens to obtain images of the movements of people within the surveillance area can be managed with a guaranteed amount of information, and in autonomous driving, which will also benefit for the same reason.

2. Methods

The approach used in this paper is designed to address two main issues: difficulties in model construction caused by image distortion and the interference of global information. For the first problem, the authors first converted the initial fisheye image (2D) into a 3D image using the lifting module and obtain 3D human joint positions by minimizing the error between the projected 3D prediction and the 2D ground truth. Considering that the relative depth of the human joints is close to the distance from the body to the camera, it is reasonable to use perspective projection to calculate the 2D projection. In addition, the authors used a learning-based approach to predict camera parameters, including focal length, principal coordinates and distortion parameters during the training phase to achieve automatic calibration, avoiding dependence on actual camera parameters.

As for the global information problem, this method aggregates features from around humans to maintain global information used to estimate absolute depth and camera parameters, while applying an attention mechanism to enhance the contribution of features to model performance.

In the process of network training, three parts, including HPoseNet, HRootNet and the reprojection module, are used to construct a top-down pipeline. HPoseNet uses ResNet50 as the main body and contains three deconvolutional layers and two fully connected layers for estimating the root relative joint position of each individual. HRootNet also uses ResNet50 as backbone to estimate root joint locations by combining features from the original and cropped images. The reprojection module first sums the roots from HPoseNet and HRootNet to obtain the absolute 3D joints, then converts them to 2D poses using the previously predicted camera parameters, and finally, projects the results again into 3D while keeping them consistent with the 2D ground truth, thus, according to the authors, significantly reducing the negative effects of image distortion.

I think that the method proposed by the authors does improve the accuracy of 3D pose estimation for multiple people from fisheye images. Creating network branches to estimate the relative 3-dimensional pose and absolute depth of the root joint separately subtly mitigates the effect of image distortion on joint position and pose, and reprojection mitigates the effect of human scale variation due to unknown distortion. However, I think that repeated conversions between 2D and 3D may amplify errors in 2D pose estimation, and in addition, the method is biased towards dealing with multi-person pose estimation on flat ground, and the errors in dealing with ground with some undulations are not discussed and have some limitations.

3. Results

In terms of evaluation, this paper first tests the synthesis by adding different levels of image distortion to two public datasets, and then also collects real fisheye images of three common human activities: posing, talking and walking, in a top-down viewpoint. The assessment of the method in this paper uses two metrics: MPJPE and MRPE. The former represents the average position error per joint, while the latter represents the average of the root position errors. As a

comparison, this paper also compares two existing 3D human pose estimation algorithms.

First, for the two synthetic datasets, the authors' method outperforms the two existing methods. For the Modified CMU Panoptic dataset, the MPJPE obtained using this method is only 66.76 and the MRPE is 182.94, whereas the other two methods achieve an MPJPE of 102.83 and an MRPE of 783.42 and 367.47 respectively, a significant difference, which demonstrates the reliability of the present method.For the Modified Shelf dataset, the MPJPE obtained using this method is 132.45 and the MRPE is 589.19, while the other two methods achieve an MPJPE of 300.18 and an MRPE of 696.10 and 793.11 respectively, so it is clear that this method is also better here, but this method is better in The error value in this dataset is much larger compared to the previous one, which casts doubt on its reliability.

For the real dataset 3DhUman dataset used in this paper, the best performance was obtained using this method, with an MPJPE of 62.14 and an MRPE of 177.95, while the other two methods achieved an MPJPE of 73.29 and an MRPE of 1536.24 and 1661.02 respectively.

In addition to this, the generalizability of the methods is explored in this paper. The resulting HPoseNet and the two previously described methods are trained on the commonly used Human3.6m dataset for 3D human pose estimation from perspective images, and the resulting MPJPE is extremely close to that of the other two methods at 54.1 and 53.3, respectively.

More specifically, the contributions of HPoseNet, HRootNet and the reprojection module of the method are explored separately and the results show that each part of the method has a positive impact on the overall model construction.

Overall, this method has a good accuracy advantage over other methods for multi-person 3D pose estimation on fisheye images, but its erratic error performance on different datasets makes users hesitant to choose this method, and the error analysis provided in this paper is not intuitive enough, my suggestion is to use metrics like PCK(Percentage of Correct Key-points) and PCP(Percentage of Correct Parts) to evaluate and see if it produces satisfactory results for users.

4. Conclusions

To summaries the paper, an effective and promising method is proposed to solve the multiplayer 3D pose estimation of fisheye images. Through top-down theory, two networks are trained separately to estimate the relative 3D pose and absolute depth of the root joint, which are then connected by reprojection to reduce the effect of image distortion and obtain a relatively stable and reliable model.

Compared to previous multi-person 3D pose estimation algorithms, the present method has a much better performance for fisheye images, which have a large amount of information and distortion, with a significant improvement in the accuracy of the estimation, and all parts of the method have been shown to have a positive impact on the results.

Nevertheless, there are limitations to this method. Firstly, there are large unpredictable fluctuations in the error of the method for different datasets, and secondly, the dataset selected for this paper is not sufficiently complex, including and not limited to the number of people in the image, the complexity of the environment, etc., to be considered as a general method.

With regard to suggestions for further research, I believe that, depending on the prospect of modern applications such as video surveillance and autonomous driving, the algorithm could be extended to predict a person's next move by estimating their posture in order to warn them of a specific behavior or to record their behavior.

References

[1] Learning Monocular 3D Human Pose Estimation from Multi-view Images. CVPR 2018.

[2] Generalizing Monocular 3D Human Pose Estimation in the Wild. 2019.

[3] Ching-Hang Chen, Deva Ramanan. 3D Human Pose Estimation = 2D Pose Estimation + Matching. CVPR, 2017