3D Pose Estimation Using Time Series Data in Event-based Video

Kakeru Koizumi Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering, Waseda University

Abstract

An event camera has a vision sensor that asynchronously detects pixel-by-pixel brightness changes. Compared to the conventional RGB camera, it provides better detection accuracy when objects move fast or are in dark areas. There are several studies in which a 3D pose is estimated from event-based video. However, joints with slight motion are not displayed, leading to a loss of pose estimation accuracy. In this paper, we propose a method to estimate joints that cannot be detected with a single frame. We use past and future frames as well as current frames for estimation. From the experiment, we confirm that the pose estimation accuracy is stabilized and improved.

Keywords: 3D Human Pose Estimation, Event-based Video, Temporal Convolutions

1. Introduction

3D pose estimation is a technology expected to be applied to various fields, such as robotics and automatic sports judgments. Generally, RGB cameras have been used for pose estimation, but there is a limit to the range of brightness and darkness that can be observed, making it difficult to capture images in dark areas. Event cameras provide events continuously and asynchronously by detecting luminance changes pixel by pixel. They also have a wider dynamic range than RGB cameras and can accurately detect almost all motion information of a subject, even at night. Event-based 3D pose estimation mainly uses multiple cameras. However, the difficulty of accurately synchronizing cameras with each other and the high cost of the cameras themselves can be obstacles to their practical use. For these reasons, monocular event-based 3D pose estimation is an attractive research topic that combines performance and practicality.

The conventional method for estimating event-based monocular 3D pose is frame-based estimation. In this method, the event signal stream is divided into a sequence of a fixed number of event packets, each event packet is converted into an image, and 3D pose estimation is performed from a single frame. However, this method does not take advantage of the attribute of event cameras, which retains almost all motion information. For example, when a person moves unevenly, such as only moving their arms, the accuracy of whole-body estimation becomes unstable due to insufficient collection of events representing joints with slight motion. Therefore, in this paper, we attempted to improve pose estimation accuracy by introducing time series information to estimate joints with little motion that cannot be acquired in a single frame. Specifically, 3D pose estimation is performed by adding features of the past and future frames to the one of the current frame.

2. Related Work

Event-based pose estimation is an area of research that has only recently begun. Event-based monocular 3D pose estimation has only been studied, to the best of the author's knowledge, by Scarpellini *et al.* [1]. The model for this study is the location circled in blue in Fig. 1. First, features are extracted using ResNet34 for a single frame. Next, each of the xy, yz, and zx heat maps, called marginal heat maps [3], are created from the features, and the operation to create a heat map is defined as a single stage. Then, to improve the accuracy, the created heat maps are reformatted into the form of the original features, and the operation of creating additional heat maps is repeated n times. n = 1 is defined as stage 1, and n = 3 is defined as stage 3.

There are also few event-based datasets for human pose estimation. DHP19 [2] is the first event camera-only dataset for human pose estimation. It consists of 17 subjects and 33 movements recorded using four cameras synchronized with markers on the subjects. In this way, 3D pose estimation using event cameras still needs to be studied, and the level of accuracy still needs to be improved.

3. Proposed Method

We propose a method to improve the accuracy of event-based monocular pose estimation using multiple frames. Joints that could not be detected from the current frame imply minimal motion. Therefore, they are almost identical to the joints detected in the frames immediately before and after the current frame.

As shown in Fig. 1, the features are extracted from each frame using ResNet34 and then concatenated to integrate the time series information. Specifically, as shown in Fig. 2, features of size X × Y × C are divided into 1 × Y × C, which are combined alternately in time-series order. The size of the combined features is 3X × Y × C. To make the size of the features large enough to

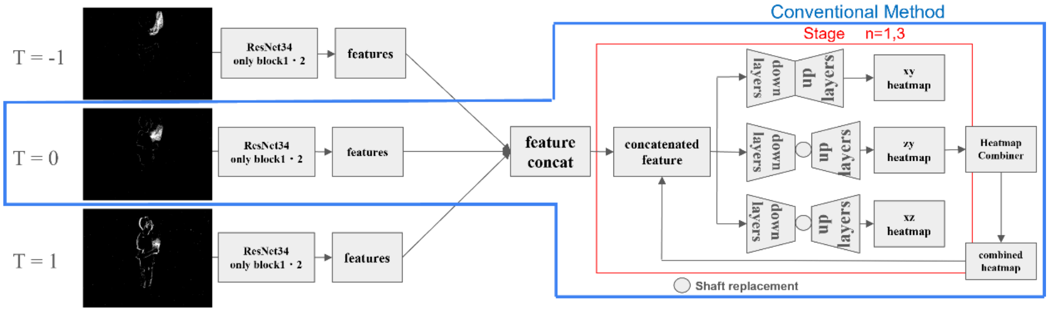


Figure 1: The network structure of the proposed pose estimation method.

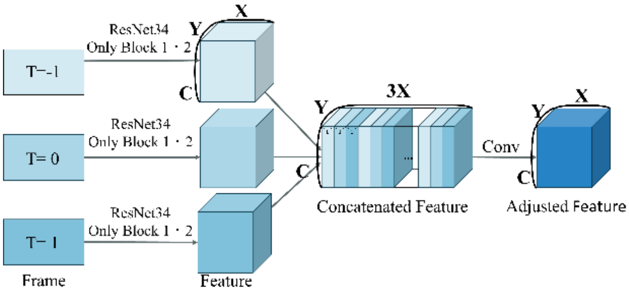


Figure 2: The way of feature concatenation.

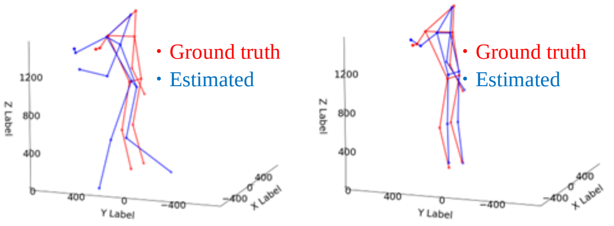
create marginal heat maps, the features are adjusted using a convolution layer of 3 × 3 kernels with 3 strides in the x direction and 1 stride in the y direction. In addition to using three frames, we propose a model that uses five frames for estimation. The specific network and feature combination methods are the same as those described in the previous section. However, the size of the combined features is adjusted by using a convolutional layer of 5 × 5 kernels with strides of 5 in the x direction and 1 in the y direction.

4. Experiments

In this section, we test the first and third stage models in the constant count representations and the correctly annotated DHP19 data set. We also test each model using 3 and 5 frames. The Mean Per Joint Position Error (MPJPE) and the number of parameters in the models are compared as quantitative values and are summarized in Table 1. The results show a 16.1 [mm] reduction in error and a significant 19.4% improvement in accuracy compared to the conventional method. Using stage 1 of our model improves the accuracy with fewer parameters than the most accurate model in conventional studies.

As a qualitative evaluation, actual 3D pose estimation results are shown in Fig. 3. Red is the ground truth, and blue is the estimation result. As shown in Figure 3(a), the conventional method estimates a largely collapsed posture for a frame with a part of the posture missing.

However, our method produces results closer to ground truth, as shown in Figure 3(b).



(a) Conventional method. (b) Proposed method.

Figure 3: Results of the pose estimation.

Table 1: Results refer to the DHP19 dataset.

Method	Stage	MPJPE(mm) ↓	Parameters
Scarpellini *et al.* [1]	1	84.84	6.99M
Proposed method (3 frames)	1	**68.38**	**7.58M**
Proposed method (5 frames)	1	**67.36**	**8.10M**
Scarpellini *et al.* [1]	3	83.06	18.68M
Proposed method (3 frames)	3	**68.66**	19.27M
Proposed method (5 frames)	3	**66.96**	19.80M

5. Conclusion

The contribution of this paper is the improvement and stabilization of the accuracy of event-based 3D pose estimation by introducing time series information. It also succeeded in lowering the computational complexity significantly compared to previous studies, improving both the practicality and performance of event-based 3D pose estimation.

References

[1] Gianluca Scarpellini *et al*., “Lifting monocular events to 3d human poses,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.1358-1368, Jun. 2021.

[2] E. Calabrese *et al*., “Dhp19: Dynamic vision sensor 3d human pose dataset,” IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1695-1704, Jun. 2019.

[3] A. Nibali, Z. He, S. Morgan, and L. Prendergast, “3D Human Pose Estimation with 2D Marginal Heatmaps,” IEEE Winter Conference on Applications of Computer Vision, pp.1477-1485, Jan. 2019.