Binocular Multi-CNN System for Real-Time 3D Pose Estimation

Teo T. Niemirepo, Marko Viitanen, and Jarno Vanne
Ultra Video Group, Tampere University, Tampere, Finland
{teo.niemirepo, marko.viitanen, jarno.vanne}@tuni.fi

ABSTRACT

The current practical approaches for depth-aware pose estimation convert a human pose from a monocular 2D image into 3D space with a single computationally intensive convolutional neural network (CNN). This paper introduces the first open-source algorithm for binocular 3D pose estimation. It uses two separate lightweight CNNs to estimate disparity/depth information from a stereoscopic camera input. This multi-CNN fusion scheme makes it possible to perform full-depth sensing in real time on a consumer-grade laptop even if parts of the human body are invisible or occluded. Our real-time system is validated with a proof-of-concept demonstrator that is composed of two Logitech C930e webcams and a laptop equipped with Nvidia GTX1650 MaxQ GPU and Intel i7-9750H CPU. The demonstrator is able to process the input camera feeds at 30 fps and the output can be visually analyzed with a dedicated 3D pose visualizer.

CCS CONCEPTS

• Computing methodologies → Neural networks • Software and its engineering → Open source model

KEYWORDS

3D pose estimation, Convolutional neural network (CNN), Stereo vision, Multi-CNN fusion, Open-source software

ACM Reference format:

Teo T. Niemirepo, Marko Viitanen, and Jarno Vanne. 2020. Binocular Multi-CNN System for Real-Time 3D Pose Estimation. In *Proceedings of ACM Multimedia conference (ACMMM'20). ACM, Seattle, Washington, USA, 2 pages.* https://doi.org/10.1145/3394171.3414456

1 Introduction

Traditional 2D pose estimation algorithms [1], [2] use a *convolutional neural network* (*CNN*) to estimate human pose from a single monocular RGB image. Because they do not provide any depth information, manual or estimated depth mapping is required before their pose estimates are deployable in any 3D scenario. On the other hand, the latest monocular depth-aware pose estimation algorithms [3]-[11] are able to infer depth information from a single image but at a cost of high computational complexity. In addition, they often suffer from limited capability to estimate odd poses that need to be inferred by quantizing a 2D projection of the pose to the database of limited set of pre-captured 3D poses.

Hypothetically, more accurate depth information could be obtained with binocular schemes that use two cameras in stereo configuration and compute disparity values between every image pixel. However, these exhaustive schemes tend to be too computationally intensive and sensitive to invisible or occluded parts of the human body.

This paper introduces a hybrid binocular multi-CNN scheme that is robust to occlusions and provides accurate 3D pose estimates without letting complexity get out of hand. To the best of our knowledge, our proposal is the first open-source implementation for real-time binocular 3D pose estimation. It is distributed under the Apache 2.0 license and available online on GitHub at

https://github.com/ultravideo/Stereo-3D-Pose-Estimation

The rest of the paper is organized as follows. Section 2 provides an overview of the proposed 3D pose estimation algorithm. Section 3 presents the demonstration setup and visitor experience. Section 4 evaluates the performance of our proposal and compares it with the state-of-the-art approaches. Section 5 gives the conclusions and directions for future research.

2 Proposed 3D Pose Estimation Algorithm

Our solution uses two cameras in stereo setup. Individual camera calibration was avoided by applying similar intrinsic parameters to both cameras. In practice, the camera intrinsic matrix was obtained with OpenCV camera calibration [12].

First, the proposed algorithm downscales the images to 360×270 pixels regardless of the input resolution. This allows the underlying networks to generalize better and to be more robust, especially with low-light and noisy images.

The intermediate 2D pose data is obtained by executing a 2D CNN on both camera feeds. In this work, the underlying network for our scheme was adopted from [1] but also other CNNs could be used.

The stereo parallax effect allows us to calculate the horizontal pixel distance between the *x*-coordinates of every keypoint in the pose. This gives a sparse disparity map of the region of interest. Contrary to depth image sampling, our solution is immune to invalid depth values caused by occluded body parts.

However, pure disparity data is not enough to determine the position of a body keypoint in world space. Therefore, a disparity value is first converted into a depth value as

$$z_{world} = \frac{f \times B}{w \times (x_{left} - x_{right})}$$

where f is the focal length, B is the distance between the cameras, w is the width of the image in pixels, and x_{left} and x_{right} denote the x-coordinates of the pixels in the input feeds, respectively [13]. Finally, the parameters of the camera intrinsic matrix are used to transform the screen point $[u, v]$ along with the depth value z_{world} to a 3D point $[x_{world}, y_{world}, z_{world}]$ in world space as

$$x_{world} = (u - c_x) \times \frac{z_{world}}{f_x}\ ,\ y_{world} = (v - c_y) \times \frac{z_{world}}{f_y}$$

where c_x and c_y are the camera principal points and f_x and f_y are the camera focal lengths [14].

The accuracy of this solution is only limited by the reliability of the underlying CNN. Some jitter and noise can always be expected so the pose is filtered by taking a weighted average of the previous pose positions. The filtering trades the pose immediate responsiveness for smoothness. In the normal case, the noise can be measured in centimeters.

3 Demonstrator Setup and Visitor Experience

Fig. 1 illustrates the demonstrator setup. Two Logitech C930e webcams are used to capture live 1080p (1920×1080) video feeds at 30 *frames per second* (*fps*) for stereoscopic video. The proposed 3D pose estimation algorithm processes the camera feeds at 30 fps on a laptop equipped with Nvidia GTX1650 MaxQ GPU and Intel i7-9750H CPU. In addition, our test set includes pre-captured videos to better highlight different features of the pose estimation.

The visitors captured by the cameras are able to see their 3D pose estimates in a visualizer as exemplified in Fig. 2. This tailor-made 3D pose visualizer is built with the Unity 3D engine. Although the underlying CNN is able to recognize multiple people, the current version of the visualizer is limited to display a single pose at a time.

4 Performance Evaluation

With pre-captured videos, our algorithm can be executed at a stable frame rate of 50 fps on a high-end Nvidia GTX 1080 GPU and AMD Ryzen 9 3900X CPU. Switching to a faster underlying 2D CNN, such as [15], has the potential to double the performance on the same hardware.

According to our subjective evaluations, the estimated pose is accurate for realistic pose reconstruction with distances less than 10 m and with an inter-camera distance of 16 cm.

Because the 2D pose extraction is performed by a different thread than the CNN inference, the resulting pose is always one frame behind the input videos. Converting the dual 2D pose into 3D is not computationally intensive and the resulting 3D pose has a latency of $2 \times (frame\ rate)^{-1}$ seconds.

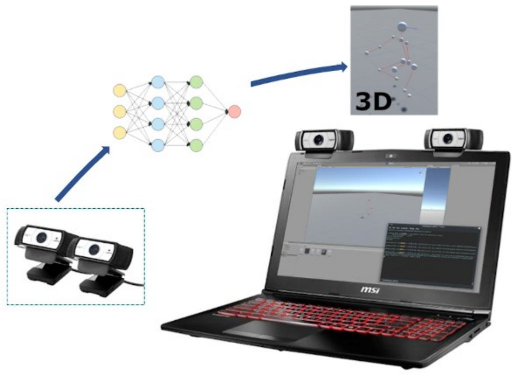

Figure 1: The demonstration setup and workflow.

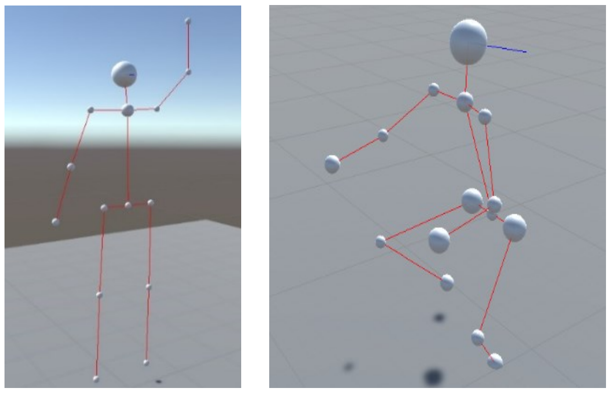
Figure 2: 3D pose visualizer in action.

Our multi-CNN fusion scheme outperforms the state-of-the-art implementations in execution speed. Many of the prior works in the field are largely theoretical and do not address actual implementations. The most competitive monocular implementation presented in [10] achieved a performance of ≈ 27 fps on a more powerful system than our demonstrator that can process pre-captured videos at ≈ 34 fps.

5 Conclusions

This paper proposed an open-source, fast, accurate, and robust implementation for binocular 3D pose estimation. It is able to provide realistic pose estimates for cost-efficient motion capture and other computer vision applications.

The future work includes offloading the applied CNN to a dedicated hardware, such as Google Coral Edge TPU [16] or Nvidia Jetson Nano [17]. In addition, the feasibility of CNN cloudification will be investigated with low-power mobile devices.

ACKNOWLEDGMENTS

This work was supported in part by the European ECSEL project NEWCONTROL (under the grant agreement 826653) and the Academy of Finland (decision no. 301820).

REFERENCES

[1] D. Osokin, “Real-time 2D multi-person pose estimation on CPU: lightweight OpenPose,” *arXiv:1811.12004 [cs.CV]*, Nov. 2018. [Online]. Available: https://arxiv.org/abs/1811.12004

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Early Access, July 2019.

[3] M. Ageeva, D. Osokin, and A. Kruglov. “Real-time 3D multi-person pose estimation demo,” [Online]. Available: https://github.com/Daniil-Osokin/lightweight-human-pose-estimation-3d-demo.pytorch

[4] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain, “Multi-person 3D human pose estimation from monocular images,” *in Proc. Int. Conf. 3D Vision*, Quebec City, Quebec, Canada, Sept. 2019.

[5] L. Wang, Y. Chen, Z. Guo, K. Qian, M. Lin, H. Li, and J. S. Ren, “Generalizing monocular 3D human pose estimation in the wild,” *in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Seoul, Korea, Oct. 2019.

[6] A. Zanfir, E. Marinoiu, and C. Sminchisescu, “Monocular 3D pose and shape estimation of multiple people in natural scenes: the importance of multiple scene constraints,” *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, June 2018.

[7] J. Liu, H. Ding, A. Shahroudy, L. Y. Duan, X. Jiang, G. Wang, and A. C. Kot, “Feature boosting network for 3D pose estimation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, Feb. 2020, pp. 494-501.

[8] S. Vosoughi and M. A. Amer, “Deep 3D human pose estimation under partial body presence,*” in Proc. IEEE Int. Conf. Image Processing*, Athens, Greece, Oct. 2018.

[9] S. Li, L. Ke, K. Pratama, Y. Tai, C. Tang, and K. Cheng, “Cascaded deep monocular 3D human pose estimation with evolutionary training data,” *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Virtual, June 2020.

[10] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H. P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, “XNect: real-time multi-person 3D motion capture with a single RGB camera,” *ACM Trans. Graph.*, vol. 39, no. 4, July 2020.

[11] S. Xia, Z. Zhang, and L. Su, “Cascaded 3D full-body pose regression from single depth image at 100 fps,” *in Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, Reutlingen, Germany, Mar. 2018.

[12] OpenCV (Open Source Computer Vision Library). [Online]. Available: https://opencv.org/

[13] A. Aslam and M. Ansari, “Depth-map generation using pixel matching in stereoscopic pair of images,” *arXiv:1902.03471 [cs.CV]*, May 2019. [Online]. Available: https://arxiv.org/abs/1902.03471

[14] H. Aghajan and A. Cavallaro, *Multi-Camera Networks: Principles and Applications*. Academic Press, Inc., 6277 Sea Harbor Drive, Orlando, Florida, USA, Apr. 2009.

[15] Lightweight mobile Tensorflow pose estimation. [Online]. Available: https://www.tensorflow.org/lite/models/pose_estimation/overview

[16] Google Coral Edge TPU. [Online]. Available: https://coral.ai/

[17] Nvidia Jetson Nano. [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit