Object Detection and Person Tracking in CathLab with Automatically Calibrated Cameras

Jiang, Y.; Dai, R.; Zeng, J.; Butler, R.M.; Vijfvinkel, T.S.; Wang, Y.; van den Dobbelsteen, J.J.; van der Elst, M.; Dauwels, J.H.G.

Publication date
2022

Document Version
Final published version

Published in
42nd WIC Symposium on Information Theory and Signal Processing in the Benelux (SITB 2022)

Citation (APA)
Jiang, Y., Dai, R., Zeng, J., Butler, R. M., Vijfvinkel, T. S., Wang, Y., van den Dobbelsteen, J. J., van der Elst, M., & Dauwels, J. H. G. (2022). Object Detection and Person Tracking in CathLab with Automatically Calibrated Cameras. In J. Louveaux, & F. Quitin (Eds.), *42nd WIC Symposium on Information Theory and Signal Processing in the Benelux (SITB 2022)* (pp. 57-57)

Object Detection and Person Tracking in CathLab with Automatically Calibrated Cameras

Yingfeng Jiang* **Renjie Dai*** **Jinchen Zeng*** **Rick Butler** **Teddy Vijfvinkel**
Yanbo Wang **John van den Dobbelsteen** **Maarten van der Elst** **Justin Dauwels**
Delft University of Technology, 2611 DX, Delft, Netherlands

Workflow analysis is a young research field that has been gaining traction in recent years. Work in this field aims to improve the efficiency and safety in operating rooms by analysing surgical processes and providing feedback or support, where observations are made and evaluated by algorithms rather than human experts. For our study, we mount five cameras from different angles in a Catheterization Laboratory (CathLab) to observe and analyse Cardiac Angiogram procedures. To automate the classification of workflow and personnel activities, we propose a pipeline that first automates the camera calibration of the 5-camera network then detect locations of medical equipment and track personnel activities.

One of the most common ways to calibrate camera networks is to use coded targets or markers [1]. However, cameras might accidentally move after the calibration, causing all 3D data calculated wrongly, therefore, the cameras need to be re-calibrated. The re-calibration can be time-consuming. An alternative is to use techniques based on Structure from Motion (SfM) which exploit the image correspondences between different views. But these are often difficult to establish when there is limited overlap between different views. To address the above limitations, we propose an automatic camera calibration framework in the CathLab, which relies on Scaled-YOLOv4 [2] to detect fixed objects and uses automatic model based on artificial neural networks to extract selected key-point features from each image frame. Then point-correspondences between the image frame and the 3D coordinates are used to compute one calibration set. A RANSAC-based filtering and aggregation algorithm is used to generate a robust estimate of the extrinsic parameters of each camera. The calibration framework is shown in Fig.1.(b).

For object detection, we apply the state-of-the-art method Scaled-YOLOv4 [2], as it has extremely fast processing speed and decent precision. However, we find Scaled-YOLOv4 still has difficulties detecting transparent objects such as lead shields In order to detect such objects, we propose an object detection algorithm based on Scaled-YOLOv4 with an auxiliary Dice Loss. The Dice Loss [3] establishes the right balance between objects and backgrounds automatically, making the boundary of objects attract more attention and contribute to more discriminative features. The proposed algorithm is consequently more powerful in detecting transparent objects. Moreover, Scaled-YOLOv4 is tailored for datasets of single view without taking advantage of the information contained in multiple views. In this work, we also design a filter following the object detection algorithm to refine the bounding boxes of objects by considering detection results from different cameras. The full pipeline of our object detection algorithm is shown in Fig.1.(a).

Multi-person tracking hinges on the accurate estimation of 3D human poses from multiple views. Most previous 3D pose estimation methods train their models directly on a 3D pose dataset, however, such data is unavailable for the task at hand. To solve this problem, we decompose 3D human pose estimation task into two stages (see Fig.1.(c)), avoiding the need to large amounts of 3D pose data: we fine-tune Scaled-YOLOv4 and HRNet for 2D pose estimation in the first stage, and use a matching algorithm [4] to match corresponding 2D poses from multiple views and then reconstruct 3D poses in the second stage. The proposed 3D human pose estimation algorithm is orthogonal to the traditional multi-view tracking algorithm, and hence can be integrated with them flexibly. Once trained, our method can be easily generalized to different Cathlabs.

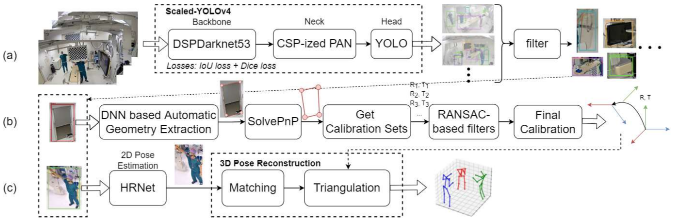

Fig. 1: Diagram of the proposed pipeline: (a) Object detection algorithm; (b) Automatic Extrinsic Calibration Framework; (c) Multi-person tracking algorithm.

References

[1] J. L. Schönberger *et al.*, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[2] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," 2020. [Online]. Available: https://arxiv.org/abs/2011.08036

[3] F. Milletari *et al.*, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016.

[4] J. Dong *et al.*, "Fast and robust multi-person 3d pose estimation from multiple views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

*Equal contribution