 DOI: 10.3217/978-3-85125-869-1-15

Human Tracking and Pose Estimation for Subsurface Operations

Roland Perko[1], Hannes Fassold[1], Alexander Almer[1], Robert Wenighofer[2], and Peter Hofer[3]

***Abstract*— Human lives are particularly at risk in critical security situations in underground train stations compared to surface events. Due to the *closed situation* of such subsurface events, considerable obstacles to the safe and efficient evacuation of people after an attack must be taken into account. Thus, this work presents a computer vision system based on artificial intelligence that uses available surveillance cameras in the optical and the thermal spectrum to detect and track human beings, and to allow an activity classification based on a pose estimation. Those results are then transferred into a 3D common operational picture to assist subsurface operations.**

I. INTRODUCTION

Subsurface structures, like the whole subway infrastructure, are indispensable for modern societies. To ensure safety and efficient reaction to crisis, a deep understanding of the underground structure is necessary for specially trained and equipped personnel, aware of the associated risks and dangers – the so called *Subsurface Operators* [2]. In the special case of a terrorist attack available technical infrastructure can be employed to derive valuable information for those operators. Since most subsurface structures are equipped with surveillance cameras, the aim of this work is to analyse that data to assist the crisis team. From the computer vision perspective, three important queues can be derived: (1) Detection of objects of interest, in particular humans and vehicles, (2) tracking of those objects over time, and (3) activity recognition, in particular if humans are walking, standing, sitting, or lying.

With the knowledge of the coarse location and orientation of the cameras, the detections can be projected onto a map or a 3D model, which then serves as a common operational picture within a virtual reality system. To simulate the subsurface environment the test site *Zentrum am Berg* (ZaB) is chosen which allows underground research, development, training, and education at 1:1 scale [7]. An exemplary view of one of the tunnel tubes is depicted in Figure 1. This specific facility is equipped with multiple optical and thermal cameras (which are very important since critical events often occur in low or no light conditions) that serve as input for the developed computer vision and artificial intelligence system.

II. METHOD

This section reports an object detection and tracking, pose estimation for activity recognition, and the common operational picture.

[1]Roland Perko, Hannes Fassold, and Alexander Almer are with Joanneum Research, Austria {firstname.lastname}@joanneum.at

[2]Robert Wenighofer is with Montanuniversität Leoben, Austria robert.wenighofer@unileoben.ac.at

[3]Peter Hofer is with the Theresianische Militärakademie, Austria peter.hofer@bmlv.gv.at

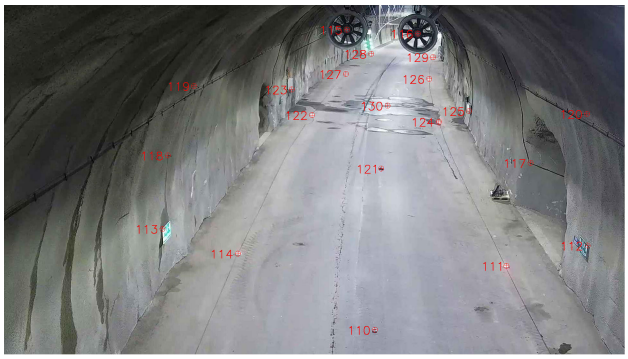

Fig. 1. Subsurface environment at ZaB with ground control points shown in red color for extrinsic camera calibration.

A. Object Detection and Tracking

For the detection and tracking of persons (and other objects), we base upon the *OmniTrack* algorithm [1]. It is real-time capable and combines a powerful deep learning based object detector (YoloV3 [8]) with high-quality optical flow methods (TV-L^1 [11]). Within this work, we updated the key components of the algorithm to more recent methods. Specifically, for the object detector component we switched from YoloV3 to the *Scaled-YoloV4* method [10]. It achieves higher accuracy by employing a cross-stage partial network and can be easily scaled to multiple resolutions. Additionally, instead of the classical TV-L^1 algorithm for optical flow we employ the recently proposed *RAFT* optical flow algorithm [9]. The RAFT optical flow method achieves high accuracy of the motion field and generalizes well to other domains (like thermal images which have a different characteristic than RGB images). Note that for both RGB and thermal input images, we use the standard Scaled-YoloV4 pretrained model, which has been trained on the MS COCO dataset [6] (consisting of RGB images). We do not fine-tune or retrain on a specific thermal image dataset. For the purpose of the project we only use the two classes humans and vehicles.

B. Human Pose Estimation

For human pose estimation, we employ the *EvoSkeleton* algorithm [5]. The method evolves a limited dataset to synthesize unseen 3D human skeletons based on a hierarchical human representation and heuristics inspired by prior knowledge. Via this special data augmentation procedure, EvoSkeleton achieves state-of-the-art accuracy on the largest public benchmark (Human3.6M [3]) and additionally generalizes well to unseen and rare poses. In order to calculate the poses (skeletons with 17 joints) for all detected persons in one frame, we proceed as follows. First, for all detected

persons the rectangular regions of interest are extracted to a list of sub-images. These sub-images are now processed in multiple batches, with the size of the batch set to 4 sub-images. The batching mechanism makes inference more efficient and ensures that the GPU memory is not exhausted. With a batch size of 4, roughly 5 GB GPU RAM are occupied. The thermal images are transferred to a different color range, which improves the performance of the pose estimation.

C. Common Operational Picture

Since all information is gathered in image geometry, it has to be transferred to the map projection of the 3D common operational picture. Therefore, the cameras are calibrated intrinsically using planar calibration random dot targets. The extrinsics are determined using ground control points (cf. the red points in Figure 1) within a least squares parameter adjustment. This calibration allows 2D information in image geometry to be intersected with an existing 3D tube model that was acquired via terrestrial laser scanning for the whole ZaB subsurface test site.

III. RESULTS

Figures 2 and 3 depict one video frame of an optical, respectively, thermal camera superimposed with the bounding boxes from the human detection and the human skeleton for each detection.

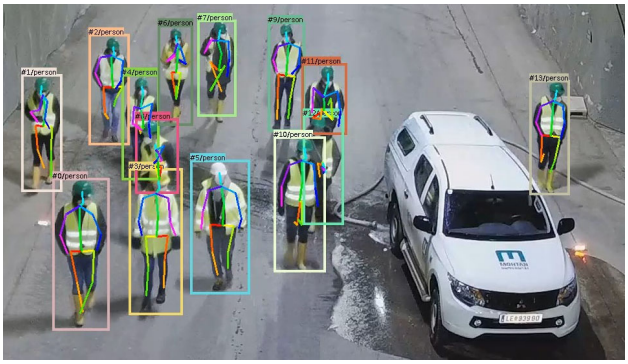

Fig. 2. Human detection and skeleton estimation for an optical image.

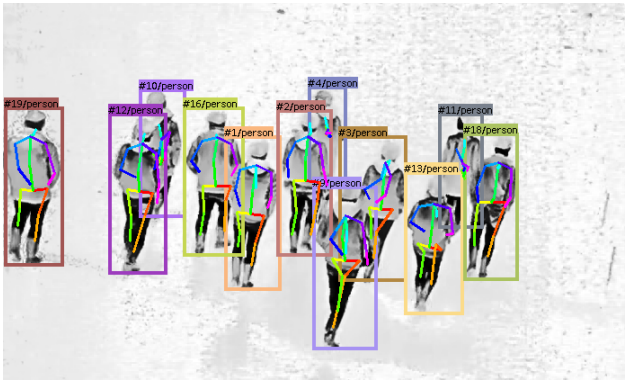

Fig. 3. Human detection and skeleton estimation for a thermal image. The color coding scheme of the thermal image was altered to improve the quality of the pose estimation.

Figure 4 depicts a screenshot of the 3D common operational picture within a virtual reality system where the subsurface operators get a simplified overview of the human detections and classifications.

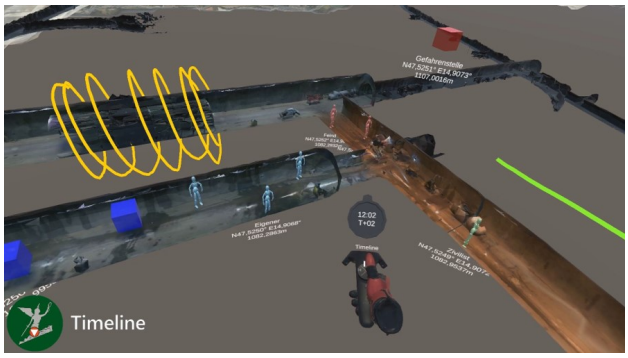

Fig. 4. The 3D common operational picture within a virtual reality system (illustration courtesy of [4]).

For RGB video, initial experiments show that both the object detection and tracking, and also the pose estimation work very well. For thermal video, the results are worse, especially for the pose estimation. This can be attributed to the *domain gap*, the fact that both methods have originally been trained on RGB image datasets and not on thermal images. Nonetheless, it seems that even on thermal video the result of the pose estimation is good enough for our task of activity classification of persons. Regarding runtime, the object detector and tracker works in real-time, whereas the pose estimation is not real-time capable. We will investigate techniques like 16-bit inference or frame subsampling in order to achieve real-time performance also for the pose estimation.

IV. CONCLUSION

A computer vision system for human tracking and pose estimation was presented, custom-tailored for subsurface operations, based on existing surveillance infrastructure. In the future, the results from the pose estimation, together with the motion information of the tracked persons, will be used for activity classification. Specifically, via the motion information a person could be classified either as stationary or moving (walking / running). Furthermore, the pose estimation information will be used to for activity recognition, in particular, whether a person is standing or lying, by analysing the person's spine orientation. Another future research focus is to preserve the privacy of people, where one option would be to use only thermal cameras.

ACKNOWLEDGMENT

The presented research activity is embedded into the project NIKE-SubMovCon #879720 within the Austrian Security Research Programme KIRAS, funded by the Austrian Research Promotion Agency (FFG).

REFERENCES

[1] H. Fassold and R. Ghermi, "OmniTrack: Real-time detection and tracking of objects, text and logos in video," in *Proc. ISM*, 2019.

[2] P. Hofer, "Coping with complexity. The development of comprehensive subsurface training standards from a military perspective," *BHM Berg-und Hüttenmännische Monatshefte*, vol. 164, no. 12, pp. 497–504, 2019.

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[4] Laabmayr, "Subsurface operation mission tool," www.laabmayr.at/tunnel-plus/rd/somt-subsurface-operation-mission-tool/, 2021, (accessed October 18, 2021).

[5] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3d human pose estimation with evolutionary training data," in *Proc. CVPR*, 2020.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[7] Montanuniversität Leoben, "Zentrum am Berg," https://www.zab.at/, 2021, (accessed October 14, 2021).

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[9] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," *ArXiv*, vol. abs/2003.12039, 2020.

[10] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," *ArXiv*, vol. abs/2011.08036, 2020.

[11] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*, 2008.