Subsurface Movement Control

Roland PERKO*, Alexander ALMER, Hannes FASSOLD, Anna MALY (JOANNEUM RESEARCH)

Robert WENIGHOFER (Montanuniversität Leoben)

Julian EDER (IL – Laabmayr und Partner ZT GesmbH)

Peter HOFER (Theresianische Militärakademie)

Abstract: Human lives are particularly at risk in critical security situations in subsurface infrastructures. Thus, this work presents concepts and initial results of the KIRAS project NIKE-SubMoveCon, which focuses on subsurface movement control. All critical information is extracted from cameras (optical and thermal) and from microphone arrays. The results are transferred into a 3D virtual reality common operational picture, which assists the subsurface operators and allows efficient movement control.

Key Words: Subsurface operations, movement control, visual and acoustic tracking, 3D common operational picture.

1. 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 (Hofer 2019; Hofer 2020). The presented work, embedded in the projects NIKE-SubMoveCon (KIRAS) and NIKE-DHQ Radiv (FORTE), focuses on information gathering on human beings who are currently in such a subsurface structure, being friend or foe, to allow an efficient subsurface movement control by the subsurface operators. In the first phase of the project, all information is collected remotely in real-time by optical and thermal cameras and by microphone arrays. The results are forwarded to an immersive 3D common operational picture (COP), which can be used by multiple operators within a virtual reality system. In the future, in-situ measurements will complement the COP with human-worn devices, the so-called human sensor, equipped with cameras, microphones and a gas sensor.

To simulate the subsurface environment the test site *Zentrum am Berg*[1] (ZaB) is chosen, which allows underground research, development, training, and education at 1:1 scale. An exemplary view of one of the tunnel tubes is depicted in Figure 5. 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. In addition, the installed *Acoustic Incident Detection in Tunnels*[2] (AKUT) system allows efficient real-time processing of audio signals. All sensor systems are geometrically calibrated, allowing the mapping from sensor geometry to 3D world coordinates used in the COP.

Overall, the system in its current state focuses on the extraction of the following human activities: (1) Detection of objects of interest, in particular humans and vehicles, (2) tracking of those objects over time, (3) activity recognition, in particular, if humans are walking, standing, sitting, or lying, (4) detection of acoustic events, in particular screaming, talking, or gun shooting. With the knowledge of the coarse location and orientation of the cameras and microphones, 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.

[1] https://www.zab.at [30.05.2022].
[2] https://www.akut-tunnel.com [30.05.2022].

*(corresponding author) Steyrergasse 17, 8010 Graz, Austria; roland.perko@joanneum.at

Finally, the project is accompanied by psychological studies that consider and evaluate human stress in this specific subsurface scenario.

2. Methods

This section reports on the proposed methodology, in particular, on sensor analysis, calibration, the common operational picture, and the overall system architecture.

a. Sensor Analysis

For object detection, tracking, and pose estimation a computer vision system is developed which combines the *Scaled-YoloV4* detector (Wang et al., 2020) with the *RAFT* optical flow (Teed and Deng, 2020) for tracking humans and cars. In principle, there is no limit to the number of humans, that can be tracked simultaneously. Although, in practice a higher number of humans in the scene means that there are more occlusions, which lead to more lost tracks (if a person is occluded, obviously it cannot be tracked anymore). The human poses are estimated via the *EvoSkeleton* algorithm (Li et al., 2020). For a detailed description, refer to (Perko et al., 2021) and (Fassold and Ghermi, 2019). The generated metadata is subsequently used for detecting the action of all persons within the camera images. The aim of the action recognition algorithm is to distinguish following human actions: standing, walking, running, lying, and sitting. Thus, the persons' trajectory is analyzed in a time interval covering, for example, the last two seconds, and the average speed is determined in kilometers per hour (km/h). In order to calculate the average speed, the 3D information delivered by the calibration (cf. next section) is taken into account. For walking persons, the average speed will be approximately 3 – 5 km/h, whereas for running people it will approximately 7 – 10 km/h. In order to distinguish standing from lying or sitting persons, the result of the pose estimation (skeleton) is used. Specifically, the orientation of the persons's spine is inspected. A (more or less) horizontal spine orientation indicates that the person is lying. From an architectural point of view, it is important to note that real-time processing of video streams is computationally demanding, such that each stream is fed into one processing unit equipped with a suitable graphical processing unit (GPU).
For acoustic event detection, a central audio server processes the streamed audio signals of all sensors. After acoustic feature extraction, detectors decide whether an event is present or not. Possible events are voices and bangs especially in strong reverberant underground environments.
The results of the sensor analysis are then transferred to a central server including the geo-location of the detected events. Currently, objects are not tracked from one segment to another segment, such that double detections are possible if a human or car is visible in two cameras.

b. Calibration

Since cameras only perceive a 2D projection of the 3D world, the location of a detected person cannot be directly projected into the 3D COP. Therefore, the cameras are calibrated intrinsically using a specially designed planar calibration random dot target (cf. Figure 1). The materials are optimized such that the dark dots are visible for optical and thermals cameras (when heating the target by sunlight or an artificial light source). This calibration step determines the focal length, principal point, and lens distortion for each camera. The extrinsics are then determined using precisely measured ground control points (cf. the red points in Figure 5) within a least squares parameter adjustment, where the coarse camera location and orientation serve as staring point. The presented 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. Within this step for each pixel on the ground plane of the tunnel (roadway and sidewalk) the absolute 3D location is extracted in the subsurface reference projection of ZaB (EPSG:31253). The latter information allows determining the velocity of tracked objects used in activity recognition. Overall, this procedure combines photogrammetry and computer vision as envisioned in (Perko, 2021).

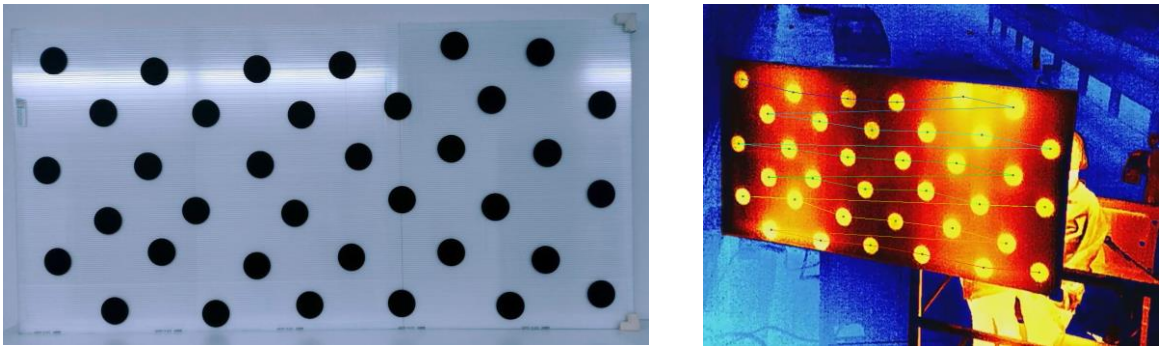

Figure 1: Specially designed camera calibration target (left) imaged by a thermal camera installed at ZaB (right). In the thermal image the detected and identified points are superimposed (points and lines), that are used for intrinsic camera calibration. Many of such images are used to calibrate each camera in the testing facility.

c. Common Operational Picture

For visualization of the results, the Subsurface Operation Mission Tool (SOMT) is used. SOMT is a Virtual Reality (VR) based Command and Control Information System (C2IS), developed to give operators an immersive look into the 3D geometry of a mission environment and, therefore, a comprehensive COP. It allows displaying information within its 3D environment in form of so-called Points of Interest (PoI). To display the information gathered by the NIKE-SubMoveCon system two types of PoIs are employed:

1. Camera Streams are used to directly display the results of the object detection (as seen in Figure 4) or unedited livestreams from cameras inside the COP. The video feed is displayed on a flat plane above the position of the camera in the model. The position of the camera is known based on the extrinsic camera calibration described above.
2. Tactical Graphics are used to display the results of the object detection through cameras and microphones symbolically. Meaningful symbols for the tracking categories were chosen from the MIL-STD 2525C (Department of Defense, 2008). Figure 2 gives some examples of the chosen symbols. Note that the shape and color of a symbol encode its affiliation (blue = friendly, red = hostile, green = neutral, yellow = unknown). As the affiliation cannot be detected automatically, all events are displayed as unknown initially and a user can assign other affiliations manually within SOMT.

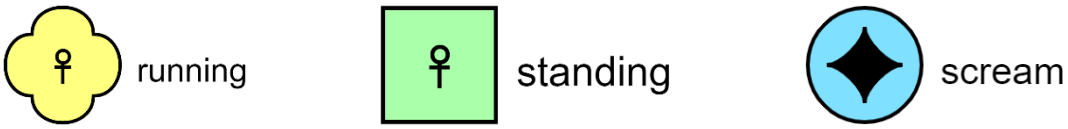

Figure 2: Symbols for a detected running person with affiliation "Unknown" (left), a standing person with affiliation "Neutral" (center) and a sensor detecting a Scream (right).

d. System Architecture

The main design rationale is based on simple but effective interfaces, where the sensor analyses are separated from the mission tool. Thus, two servers are deployed, which are the *Subsurface Operations Communication and Analysis* (SOCA) server and the *Subsurface Operations Mission Tool* (SOMT). SOCA collects all processing results and directs them, after an initial pre-processing, to SOMT via Rest-API and Streaming URL (cf. Figure 3). SOMT on the other side pipes the information into the 3D common operational picture. Technically, both systems can run on one single computer. The computing servers are hardwired within the same IP network, where also all cameras can be accessed via the real time streaming protocol (RTSP).

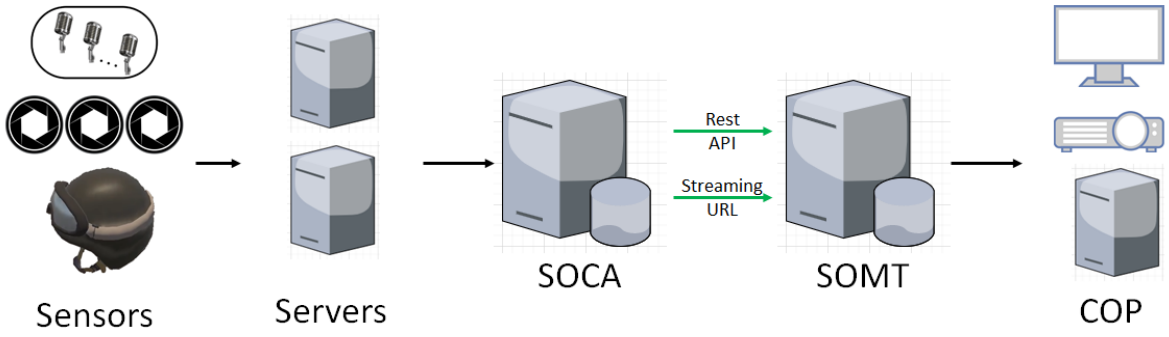

Figure 3: System overview depicting the main components used in the presented subsurface movement control.

3. Results

Within this section, exemplary results are visualized for the main components of the subsurface movement control system. Figure 4 shows human detection, tracking, and skeleton estimation for an optical camera and for a thermal camera in the test facility. All individuals are detected and also the skeletons, in particular the orientation of spines for activity recognition, are estimated correctly. This is surprising, since the *EvoSkeleton* algorithm was not trained on thermal data. This domain gap was reduced by applying color transformations such that the thermal data looks more similar to optical imagery. Figure 5 depicts one image of an optical camera with GCPs superimposed in red color used for extrinsic camera calibration. The colored region on the right side represents the determined 3D world coordinate of each pixel on the tunnels ground plane, where the East, North, and height values are interpreted as pseudo RGB colors.

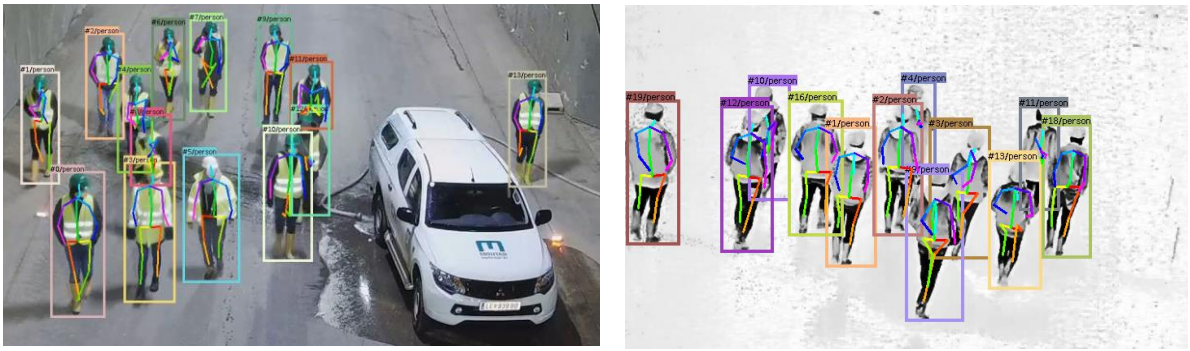

Figure 4: Human detection and skeleton estimation for an optical image (left) and a thermal image (right). The color coding scheme of the thermal image was altered to improve the quality of pose estimation.

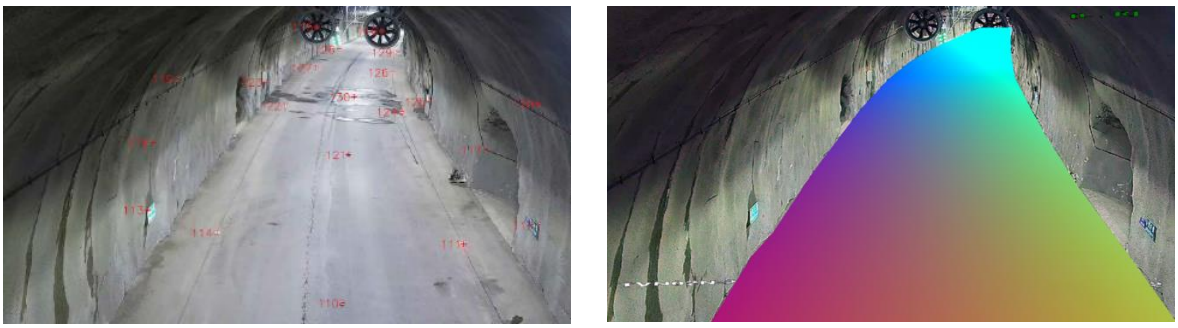

Figure 5: Subsurface environment at ZaB with ground control points shown in red color for extrinsic camera calibration (left) and extraction of the absolute 3D world coordinates for each pixel on the tunnels ground plane, represented by the colored region.

The application of the acoustic tunnel monitoring system AKUT for critical security situations already convinced various operating companies of tunnels where the detection of critical events via images is not sufficient. Especially, when multiple video streams are observed by a human operator, critical events might be missed. By employing the installed audio sensors and additional usage of sensors on helmets and robots, acoustic event detectors immediately determine possible critical situations like voices where no people are

expected. In poor visibility situations, acoustic event detection also serves as a security backup. Furthermore, the combination of these acoustic detectors with visual tools assists the decision-making unit in a very efficient way.

Figure 6 shows how the visualization will look like in the following imaginary scenario. Note that the system has not been end-to-end tested yet and, therefore, the underlying test data is manually created.

Scenario: "A violent attack was committed at the end of the street-tunnel at ZaB. Through the live camera feed, the operator can see that no individuals are located in the northern tube of the tunnel. In the southern tube, the system shows the detection of multiple persons. One person has already been classified as "neutral" by the operator. The detection of a scream, as well as a running and a lying individual in the eastern part of the tunnel system, allows the operator to draw the conclusion that the attack is located in that area and plan the response accordingly."

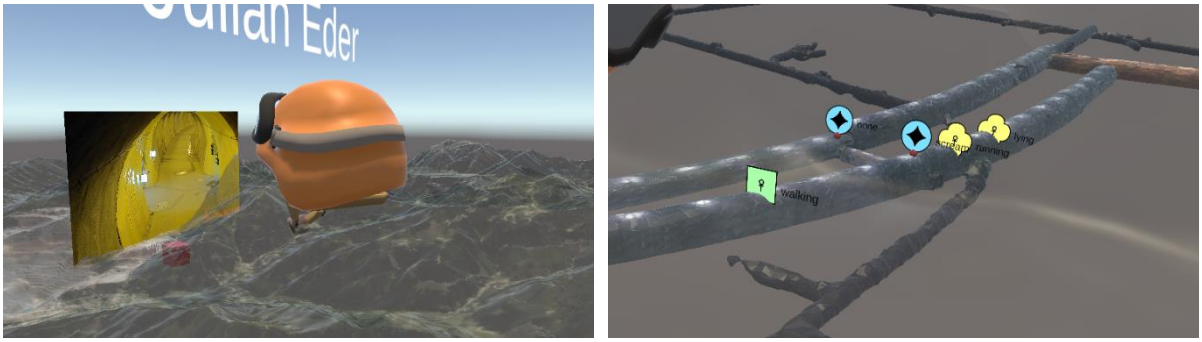

Figure 6: VR-user looking at a livestream from a georeferenced tunnel surveillance camera (left) and tactical symbols in the model of ZaB, viewed from a south-west perspective (right).

4. Conclusion

For optimal assistance of subsurface operators, this work presented concepts and results developed within the projects NIKE-SubMoveCon and NIKE-DHQ Radiv, which (also) focus on subsurface movement control. The system within its current state use cameras (optical and thermal) and microphones installed at the test facility at Zentrum am Berg to extract critical information on human presence and activity. The extracted data was transferred to a virtual reality system, which acted as a 3D common operational picture, which can be used by many operators simultaneously. One bottleneck in terms of scalability is the video-based detection and tracking system, where each camera stream is processed by one GPU. Thus, for a major subterranean system a decently sized GPU cluster is needed.

The future work includes four areas: (1) The current system will be demonstrated at the IRON NIKE demonstration in July 2022 at Zentrum am Berg in Eisenerz. (2) In the second project phase, in-situ measurements will complement the COP with human-worn devices, the so-called human sensor, equipped with cameras, microphones, and other sensors (e.g., a gas sensor). This also includes mobile multi-sensor platforms mounted on tripods that can rapidly be placed on strategic locations, which cannot be observed by the stationary installed sensors. (3) Employing multi-camera tracking to avoid duplicate detection. (4) Finally, it would be of high interest to compare the proposed system to the smart 3D surveillance system from Leica[3], which also incorporates a Lidar sensor.

5. Acknowledgment

The presented research activity is embedded into the projects NIKE-SubMovCon #879720 (within the Austrian Security Research Programme KIRAS) and NIKE-DHQ Radiv #886302 (within the Austrian Defense Research Program FORTE), funded by the Austrian Research Promotion Agency (FFG).

[3] https://shop.leica-geosystems.com/de/leica-blk/blk247 [30.05.2022].

Publication bibliography

Department of Defense (2008): Joint Military Symbology. MIL-STD-2525C. US Department of Defense. https://worldwind.arc.nasa.gov/milstd2525c/Mil-STD-2525C.pdf [30.05.2022].

Hannes Fassold and Ridouane Ghermi (2019): OmniTrack: Real-time detection and tracking of objects, text and logos in video. In Proc. ISM, pp. 245–246.

Peter Hofer (2019): Coping with complexity. The development of comprehensive subsurface training standards from a military perspective. In Berg Huettenmaenn Monatsh 164, no. 12, pp. 497–504.

Peter Hofer (2020): The SubSurface Operations Cell: High-value Asset for Decision-Making in Complex SubTerranean/SubSurface Operations. In Berg Huettenmaenn Monatsh 165, no. 12, pp. 666–672.

Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, and Chi-Keung Tang, and Kwang-Ting Cheng (2020): Cascaded deep monocular 3D human pose estimation with evolutionary training data. In Proc. CVPR, pp. 6173–6183.

Roland Perko, Hannes Fassold, Alexander Almer, Robert Wenighofer, and Peter Hofer (2021): Human tracking and pose estimation for subsurface operations. In Austrian Association for Pattern Recognition Workshop, pp. 77–79.

Roland Perko (2021): Photogrammetric Computer Vision in Remote Sensing. Habilitation, Graz University of Technology, p. 369.

Zachary Teed and Jia Deng (2020): RAFT: Recurrent all-pairs field transforms for optical flow. In ArXiv, vol. abs/2003.12039, pp. 1–21.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao (2020): Scaled-YOLOv4: Scaling cross stage partial network. In ArXiv, vol. abs/2011.08036, pp. 1–10.