Headquarters Papers
#UOET Proceedings 2023
https://www.milak.at/nike/uoet

doi: 10.5281/zenodo.10761610

Crowd Counting and Localization for Subsurface Operations

Roland PERKO*, Richard LADSTÄDTER, Michael HUBER, Sead MUSTAFIC,
Alexander ALMER, Manfred KLOPSCHITZ (JOANNEUM RESEARCH)

Abstract: Accurate crowd counting and localization are particularly important in subsurface environments to assist decision-makers in critical security situations. Thus, we introduce an approach based on artificial intelligence to count and localize individuals, both emergency personnel and uninvolved persons, to resolve crowded situations. Our method needs a single image to predict the overall crowd count, density map, and precise locations of each individual, which could then be mapped to a common operational picture. With the additional precise person location information more concise end user information can be made possible, while simultaneously improving absolute counting information.

Key Words: Crowd counting, crowd localization, artificial intelligence, subsurface operations.

1. Introduction

In subsurface and urban operations, one key element is to gather all possible information in a common operational picture (COP). This frontend supports the responsible operators who can cooperatively work with such a system. An example of a 3D COP was developed within the NIKE[1] SOMT project (Hofer et al., 2020; Gegenhuber et al., 2022) which can be used simultaneously by multiple operators within a virtual reality system and was applied in several NIKE activities such as SubMoveCon and DHQ-RADIV (Perko et al., 2021).

Since humans are of crucial interest in a subsurface or urban scenario, this work presents a crowd counting and localization approach based on optical and thermal cameras. The acquired multi-modal images, which are synchronized in time and co-registered in space, are the basis for a novel analysis method based on artificial intelligence (AI). The proposed method predicts the count of humans in the image and their precise locations. In case of fully calibrated cameras, the 2D results can directly be mapped in the 3D COP as described in (Perko, 2021). The underlying rational for using both optical and thermal cameras is the fact that such a system should also work under low or no light conditions. This is particularly important in subsurface environment but also for crowded events (like a music festival or a soccer game) that often take place during the night. Therefore, the presented approach can serve various applications, from small to large scale (border surveillance to demonstrations), with challenging illumination conditions.

2. Setup and Methods

This section reports on the proposed hardware setup, the algorithmic design, and the implementation for crowd counting and localization.

a. Hardware Setup

To set up a camera system for combined recording of optical (RGB) and thermal infrared (TIR) imagery we chose the cameras listed in Table 1, which represent well-known industry standard. The lenses are selected to provide approximately the same field of view, resulting in a ratio of RGB to TIR resolution of about 1:5 vertically and horizontally.

[1] NIKE is the abbreviation for „Nachhaltige Interdisziplinarität bei komplexen Einätzen unter Tage“ or „Sustainable interdisciplinarity for complex subsurface operations“.

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

Table 1: Technical specifications of the RGB and TIR camera.

Parameter	RGB camera	TIR camera
Model / Type	Allied Vision GT3400C	Optris PI 640
Interface	GigE	USB
Framerate	≤ 10 Hz	32 Hz
Sensor type	Interline CCD (Color)	FPA (uncooled Bolometer)
Pixel size	3.69 µm	17 µm
Resolution	3382 x 2702 pixel	640 x 480 pixel
Focal length	12 mm	10.5 mm
Field of view	55° x 45°	60° x 45°

For a first prototype of the camera system, both cameras were integrated in weatherproof housings and mounted side by side on a stable aluminium plate equipped with a tripod adapter (see Figure 1). The total weight of the camera system (without tripod) is about 8 kg. First test recordings were taken in December 2022 at the JR DIGITAL office looking down from the 4th floor to the street level. Observation of cars and walking people allows evaluation of image quality with moving objects at distances up to 250 m.

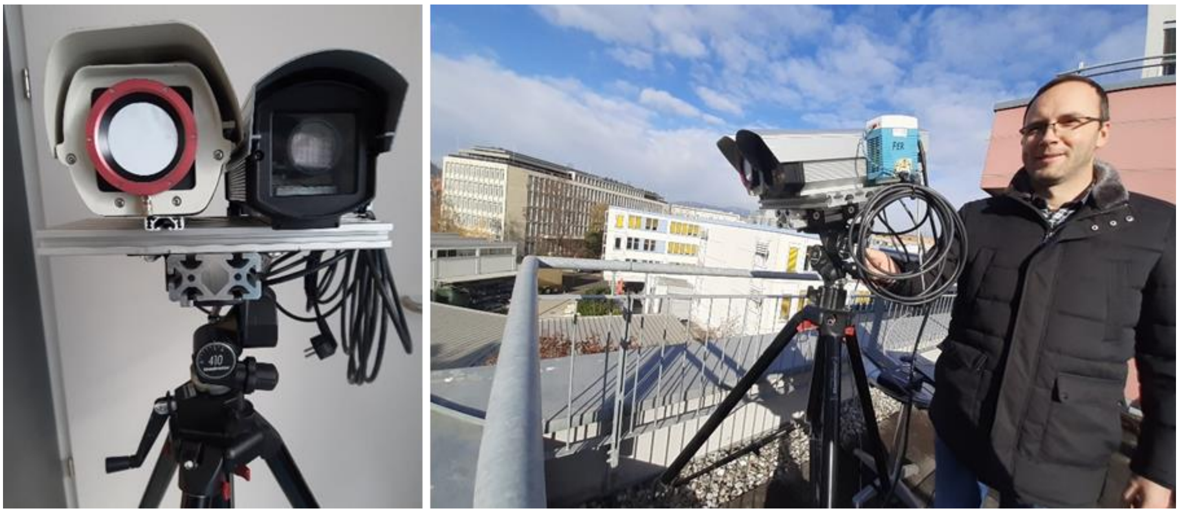

Figure 1: Combined RGB/TIR camera prototype mounted on a tripod, first test recordings at JR DIGITAL, Steyrergasse 17, Graz (4th floor).

Obviously, one very important aspect is to provide a synchronized recording of the RGB and TIR frames, respectively. The RGB camera can be triggered by an external source but the TIR camera is a free-running sensor, which only allows flagging the next frame taken after a trigger signal is set. Using the RGB camera as the "reference camera" to be configured at a certain frame rate does not lead to a stable synchronization and produces missing frames from time to time. Therefore, the other way round was implemented and the TIR camera is used as the trigger source for the RGB camera. As the 32 Hz frame rate is higher than the maximum RGB frame rate, the TIR sync signal has to be divided by a factor ≥ 3 before it can be used as the RGB trigger input signal. The according trigger electronic was developed in-house and can be controlled by a computer via a LAN interface. The trigger method described above is considered to be very stable and accurate, which is empathized by using the acronym S-RGBT (abbreviation for "Synchronized RGB and TIR" camera).

In order to get a well-registered RGB/TIR image it is also necessary calibrating both cameras for internal and relative orientation parameters (for details see our work on multi-modal multi-sensor calibration (Ladstädter et al., 2023)). This was done in the measurement lab of the Institute of Engineering Geodesy and Measurement Systems (IGMS) at Graz University of Technology (TUG), which is equipped with special (electrically heated) thermal markers (compare Figure 2, left side). After removing geometric distortions and determination of a projective transformation between both image modalities, the image sample in Figure 2 (right side) was generated. As can be seen from the walking person in the middle, the images are geometrically correct and synchronized in time.

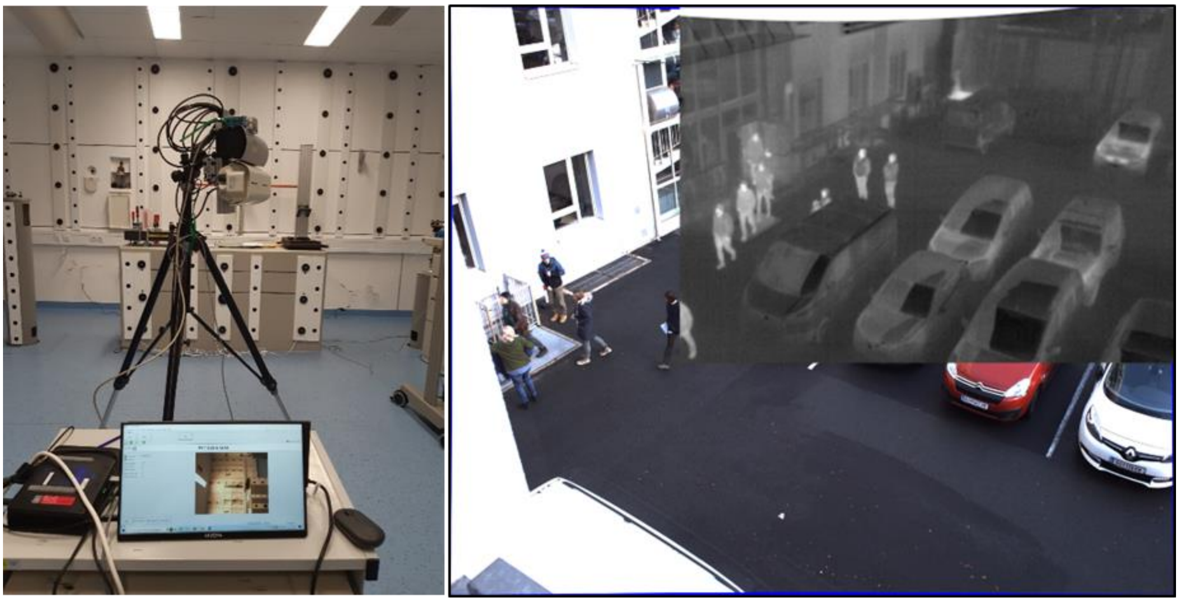

Figure 2: Geometric calibration of the S-RGBT camera system in the measurement lab (IGMS/TUG); Synchronized and co-registered image sample where parts of the TIR image are superimposed on the RGB image.

b. Algorithmic Design

A standard approach for crowd counting is to estimate the object density for each pixel in the given input image, where the sum over this density equals the object count (Lempitsky and Zisserman, 2010; Perko et al., 2013; Almer et al., 2016). Lately, various AI-based network architectures were proposed, which use an encoder-decoder scheme based on convolutional neural networks (CNN) (cf. the review in Perko et al., 2021). The exemplary high-level workflow for counting via density estimation is depicted in Figure 3. Such algorithms allow counting the objects of interest, but cannot localize individuals.

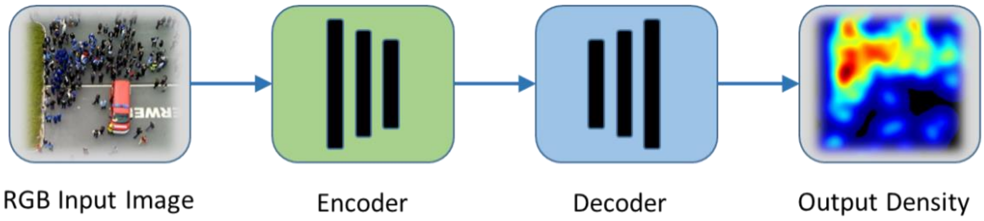

Figure 3: Deep learning architecture for crowd counting. Classical counting via density estimation where the network predicts an output density and, thus, indirectly the object count.

Our proposed network design uses an encoder mainly as a strong feature extractor, which is followed by two separate prediction heads. The first predicts a fixed number of 2D locations of objects in the image, whereas the second predicts the according confidences. Predicted points with a confidence below 0.5 get discarded. The remaining points define the localization of individuals where their count is directly given. Object densities could be simply generated by smoothing the point locations with a Gaussian kernel as, for example, as described in (Lempitsky and Zisserman, 2010; Perko et al., 2021). The resulting high-level workflow is depicted in Figure 4 and follows the principle of (Song et al., 2021).

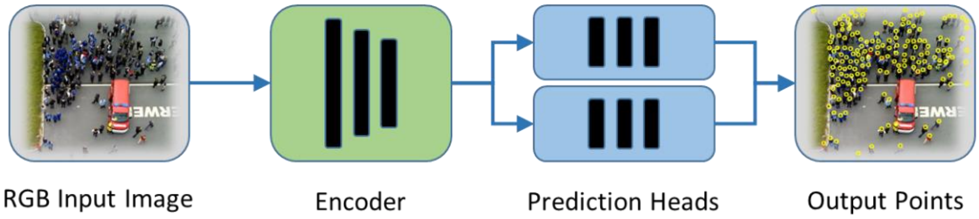

Figure 4: Deep learning architecture for crowd counting and localization. Our proposed design predicts the location of objects in form of 2D points, together with confidences and the overall human count.

c. Implementation

Our implementation follows the workflow of (Song et al., 2021) and represents a fully end-to-end trainable network. One crucial step is the comparison of point predictions from the prediction heads to the ground truth object locations. Here, a one-to-one matching is performed based on the Hungarian algorithm (Kuhn 1955), such that the point set can be divided into a positive and a negative set. This information is then the basis for training the deep neural network via backpropagation.

3. Results

This section reports the first results of the proposed hardware and software components. The ability to capture synchronized image data is demonstrated and also further improvements in regard to person counting and localization are shown.

a. Hardware

Despite the fact that first results with the S-RGBT camera prototype delivered satisfying results (see Figure 2), both the synchronization and the geometric co-registration need further investigation to reach the best possible results: (1) Synchronization: There is a constant, yet undefined time offset between the mid exposure of the TIR frame and the TIR sync pulse. This could be measured by a rotating target (visible to RGB and TIR cameras). If, in addition the variable RGB exposure time is considered, the mid exposure of the RGB and TIR camera could be synchronized even more precisely. (2) Co-registration: The baseline between the cameras is about 12 cm, which results in horizontal parallaxes for short object distances (~15 RGB pixel @ 25 m). This can be optimized by reducing the baseline even further, but this is limited by the camera housings. If a depth map is known for the camera view (or approximated by the viewing geometry) one could try to correct, for example, the TIR image for distance dependant parallaxes in a pre-processing step. This would improve the geometric co-registration significantly at the cost of additional computational effort.

b. Software

As a proof of concept, we trained and validated the proposed algorithm on two benchmarks and compared the results to state-of-the-art approaches. In particular, we chose the ShanghaiTech Part A (SHHA) (Zhang et al., 2016) and the NWPU (Wang et al., 2020) benchmarks and compared the results to the following other methods: (1) Multi Column Method (MCNN) (Zhang et al., 2016), (2) Congested Scene Recognition Network (CSRNet) (Li et al., 2018), and (3) Context-Aware Method (CAN) (Liu et al., 2019). For validation we used two metrics, namely the mean absolute error (MAE) and the root mean square error (MSE) of counting discrepancies over all images on the according test set (not used for training, cf. (Perko et al., 2021)). Results are reported in Table 2. Actually, the presented method outperforms the others in 3 out of 4 metrics. The large MSE value of our experiments at the NWPU benchmark stems from the fact, that we downsampled the training and testing images to a maximal width respectively height of 2048 pixel for speedup (corresponding to about 3 MP per image). Thus, very large images with up to 73 MP with human counts up to 20.000 lose their details, such that the persons are not recognizable in the downsampled imagery. We could change our procedure for such images however this was not the focus of this work.

Table 2: Performance indicators for the two benchmarks. Best results are marked in bold face.

architecture	SHHA		NWPU	
	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	232.5	714.6
CSRNet	68.2	115.0	121.3	387.8
CAN	62.3	100.0	106.3	**386.5**
ours	**53.1**	**82.5**	**102.8**	648.4

Since, the presented S-RGBT camera system will have its first appearance at the "Donauinselfest" in Vienna in June 2023, we show expected results for subsurface scenarios in Figure 5 and for urban scenarios in Figure 6.

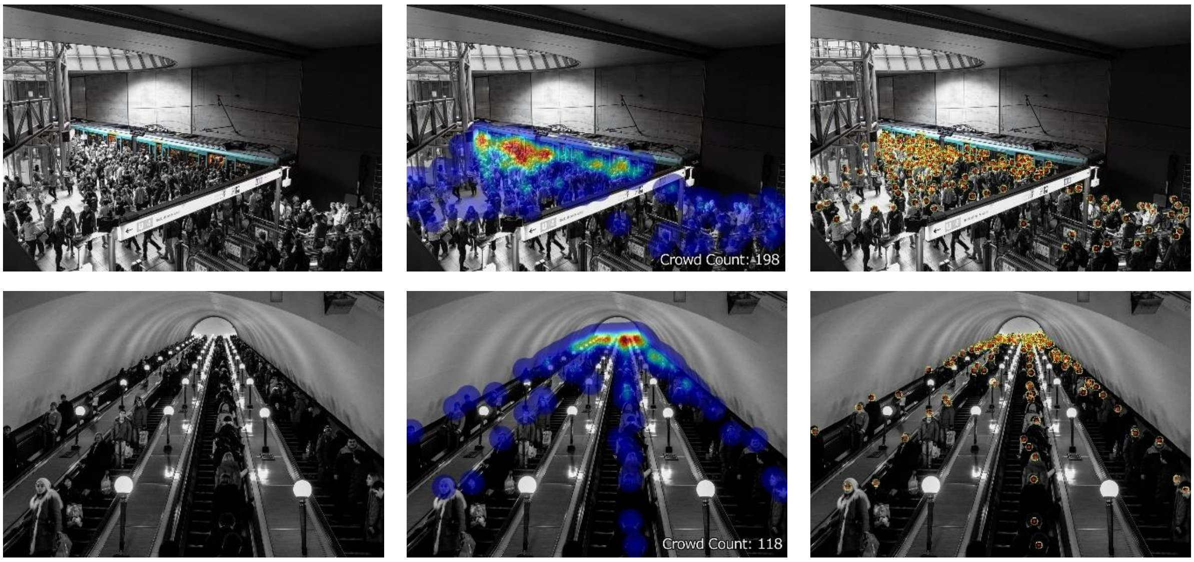

Figure 5: Examples of crowded subsurface scenes in Frankfurt (top) and in Budapest (bottom). Original image (left), human density overlay and estimated crowd count (middle), and localization of each individual (right). Image Source: www.pixabay.com.

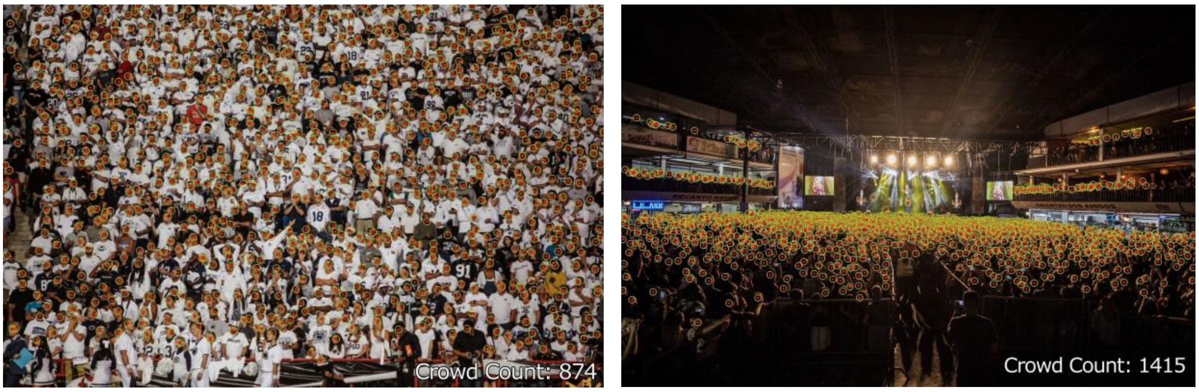

Figure 6: Examples of crowded urban scenes from a sport event (left) and from a concert (right). Shown are the localization of each detected individual. Image Source: NWPU benchmark.

4. Conclusion

This work presented our S-RGBT camera system, consisting of a synchronized RGB and TIR camera. The system is triggered with custom electronics and was also geometrically calibrated. In addition, we proposed a human counting method which, in contrast to various state-of-the-art methods, yields the precise location of each detected individual while also providing high counting accuracies. These first promising counting results were demonstrated on the RGB image modality.

The presented hard- and software concept will have its first appearance at the "Donauinselfest" in Vienna in June 2023. The focus will be on several aspects: First, to evaluate the ease of use of our recording hardware in a real scenario. Second, to validate the counting and location accuracy in real life. Third, to validate to which extend the TIR images are able to assist the counting procedure.

Future resources will be put into even better synchronization based on measuring the time offsets of mid exposures and into transformer-based architectures for crowd counting (cf. (Liang et al., 2022)). The latter should lead to faster convergence in training, since the patch and batch sizes could be increased, as those architectures need less GPU RAM.

5. Acknowledgment

The presented research activity is embedded into the project MUSIG #886355 (within the Austrian Security Research Programme KIRAS) funded by the Austrian Research Promotion Agency (FFG).

Publication bibliography

Almer, Alexander; Perko, Roland; Schrom-Feiertag, Helmut; Schnabel, Thomas; Paletta, Lucas (2016). Critical situation monitoring at large scale events from airborne video based crowd dynamics analysis. In AGILE International Conference on Geographic Information Science, 19, pp. 351-368.

Gegenhuber, Nina; Wenighofer, Robert; Nieß, Birgit (2022). IRON NIKE Forschungstätigkeiten 2022 am Zentrum am Berg. In Berg- und Hüttenmännische Monatshefte, 167(12), pp. 578–581.

Hofer, Peter; Strauß, Clemens; Wenighofer, Robert; Eder, Julian; Hager, Lukas. (2020). Die Rolle von Virtual Reality in der Bewältigung militärischer Einsätze unter Tage. In AGIT Journal für Angewandte Geoinformatik, 6, pp. 126-131.

Kuhn, Harold W. (1955). The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2), pp. 83-97.

Ladstädter, Richard; Gutjahr, Karlheinz; Perko, Roland; Gregorac, Ana (2023). Simultane Kalibrierung von RGB/TIR Multi-Kamerasystemen, In Internationale Geodätische Woche in Obergurgl, pp. 150-158.

Lempitsky, Victor; Zisserman, Andrew (2010). Learning to Count Objects in Images. In Advances in Neural Information Processing Systems, 23, pp. 1324–1332.

Li, Yuhong; Zhang, Xiaofan; Chen, Deming (2018). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1091–1100.

Liang, Dingkang; Xu, Wei; Bai, Xiang (2022). An end-to-end transformer model for crowd localization. In European Conference on Computer Vision, Part I, pp. 38-54.

Liu, Weizhe, Salzmann, Mathieu, Fua, Pascal (2019). Context-aware crowd counting. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 5099-5108.

Perko, Roland; Schnabel, Thomas; Fritz, Gerald; Almer, Alexander; Paletta, Lucas (2013). Airborne based high performance crowd monitoring for security applications. In Scandinavian Conference on Image Analysis, 7944, pp. 664-674.

Perko, Roland; Klopschitz, Manfred; Almer, Alexander; Roth, Peter M. (2021). Critical aspects of person counting and density estimation. Journal of Imaging, 7(2), p. 21.

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

Song, Qingyu; Wang, Changan; Jiang, Zhengkai; Wang, Yabiao; Tain, Ying, Wang, Chengjie; Li, Jilin; Huang, Feiyue; Wu, Yang (2021). Rethinking counting and localization in crowds: A purely point-based framework. In IEEE International Conference on Computer Vision, pp. 3365-3374.

Wang, Qi; Gao, Junyu; Lin, Wei; Li, Xuelong (2020). NWPU-crowd: A large-scale benchmark for crowd counting and localization. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), pp. 2141-2149.

Zhang, Yingying; Zhou, Desen; Chen, Siqin; Gao, Shenghua; Ma, Yi (2016). Single-image crowd counting via multi-column convolutional neural network. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 589-597.