Theresian Military Academy
HEADQUARTERS PAPERS

Vol. 2, Special Issue: #UOET24 Proceedings

doi: 10.5281/zenodo.13840091

Transferable Crowd Counting and Localization for Urban Operations

Roland PERKO* (JOANNEUM RESEARCH)

Richard LADSTÄDTER (JOANNEUM RESEARCH)

Ana GREGORAC (JOANNEUM RESEARCH)

Christian REPASKI (eurofunk KAPPACHER)

Abstract: In urban operations, an intuitive common operational picture (COP) is essential for gathering necessary information for operators like fire fighters, police, and military personnel. This work emphasizes accurate crowd counting and localization in urban settings using video surveillance and AI, particularly deep learning with vision transformers. Results are visualized as a 2D human density layer in the COP. Addressing data set variability, the presented multi-resolution approach enables transferable AI models, overcoming the challenge of diverse training and real-world data.

Key Words: crowd counting, crowd localization, artificial intelligence, transferability

1. Introduction

In various urban operations, one key element is an intuitive and simple common operational picture (COP) where all necessary information is gathered. Such a visual frontend supports the responsible operators which, depending on the scenario, can be fire fighters, police officers, security units, disaster control task forces or military personnel. Exemplary geographical COPs based on a 2D representation are described in (Almer et al., 2016a; Frering et al., 2023) or based on 3D visualization in the NIKE[1] SOMT project (Hofer et al., 2020).

In this work, we focus on accurate counting and localization of large crowds in urban scenarios based on video surveillance systems. The underlying analysis methods build upon modern artificial intelligence (AI), in particular deep learning with vision transformers (Dosovitskiy et al., 2020). The results are then mapped onto a 3D surface showing the density of people to be visualized as an additional layer in the COP. One of the main challenges in counting and localization is a very generic one, namely that existing data sets used for training of such an AI architecture have different properties than the real data acquired for the specific scenario under investigation. The first option would be to annotate sufficient real data and admix it in the training process. Obviously, this strategy is very time consuming and does not provide a general solution. The second option is to understand the variations within the data sets and transfer the model trained on the reference data to the real data. The main contribution of this work is, thus, a multi-resolution approach allowing transferable crowd counting and localization over diverse data sets. The whole work is embedded into the KIRAS project MUSIG, where data is combined from various sources (in specific social media, mobile communication, and video cameras) for efficient event monitoring within a custom-tailored COP.

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

2. Data Set

Every year by end of June the *Donauinselfest* music festival takes place in Vienna. The festival has free admission and offers live acts on several stages distributed over the long-stretched island of the Danube River, attracting lots of visitors[2]. For that reason, the Donauinselfest 2023 (23rd to 25th of June) had been selected for data acquisition within the MUSIG research project. The camera position was chosen in front of the main stage, mounted at 12 m height looking towards the crowd (see Figure 1). The camera was already installed on 22nd of June, being operated during the whole festival up to five hours a day (from 7 pm to midnight), survived heavy rain and thunderstorms, and was deinstalled after the festival. As the camera was never moved, all images collected have the same field of view covering approximately an area of 1680 m².

Figure 1: Mounting of our camera system at the Donauinsel music festival held in June 2023 in Vienna.

The camera setup collected about 240.000 images at a mean frame rate of 6 Hz, resulting in 2.2 TB of synchronized RGB and thermal images (for all specifications of the multi-modal camera system we refer to (Perko et al., 2023)). This huge data set (referred as *DIF23* from now on) contains imagery at different light conditions, day and night, and records crowds of only a few persons up to quite high densities (5000 people visible). It is therefore considered to be a very valuable data set to develop and evaluate crowd counting methods. From the DIF23 data set, a total of 16 representative images were manually labeled and serve as reference for validation of the performance of the proposed AI systems. All 2D locations of human heads were annotated, resulting in a total of about 32.000 heads, and thus on average 2000 humans per image.

3. Methodology

a. Crowd Counting, Localization, and Motion Estimation

A traditional approach for crowd counting is based on object density estimation (Lempitsky and Zisserman, 2010; Perko et al., 2013; Almer et al., 2016b). The total count of objects is determined as the sum over all density pixels. Early AI-based approaches can be summarized as CNN encoder-decoder schemes (cf. the review in (Perko et al., 2021) and Figure 2). In this work MCNN (Zhang et al., 2016), CSRNet (Li et al., 2018), and CAN (Liu et al., 2019) were chosen – approaches which cannot localize individual objects.

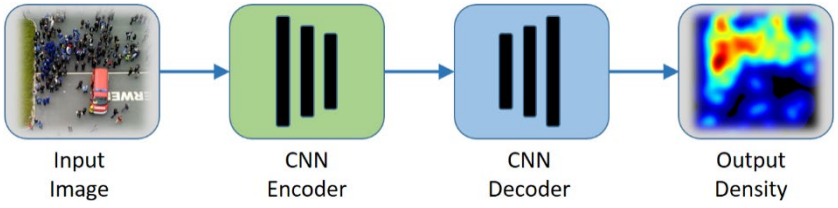

Figure 2: Deep learning architecture for crowd density estimation, based on CNN encoder and decoder (cf. (Li et al., 2018)). This design predicts the crowd density and the overall human count.

[2] https://donauinselfest.at/ (accessed on 6th of June 2024).

More recent AI methods allow counting and localization by combining CNN encoder with prediction heads (cf. Figure 3), which directly yield object locations with confidences. As reference method P2PNET (Song et al., 2021) was taken.

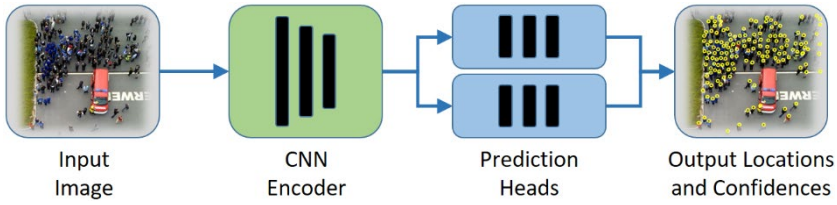

Figure 3: Deep learning architecture for crowd counting and localization, based on CNN encoder and prediction heads (cf. (Song et al., 2021)). This design predicts the location of objects in form of 2D points, together with confidences and the overall human count.

Finally, vision transformers outperform the previous methods. Here features are extracted by a CNN encoder, which are then fed to a transformer encoder-decoder module with prediction heads (cf. Figure 4). As exemplary method CLTR (Liang et al., 2022) was selected.

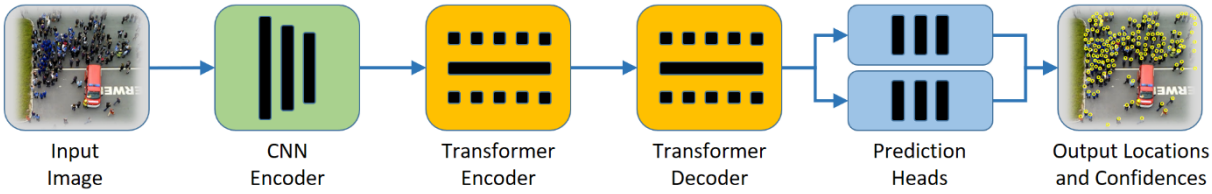

Figure 4: Deep learning architecture for crowd counting and localization, based on CNN encoder, Transformer encoder, Transformer Decoder, and prediction heads (cf. (Liang et al., 2022)). This design predicts the location of objects in form of 2D points, together with confidences and the overall human count.

On our real DIF23 data set, however, the approaches tend to overestimate the object count for images with low to medium sparse crowds and underestimates for dense crowds. This effect was also observed in (Zhang et al., 2016; Li et al., 2018) on other data sets (in specific on UCF CC 50 (Idrees et al., 2013) and ShanghaiTech Part A (Zhang et al., 2016)). In the present case, the count deviations stem from the fact that our images hold on average larger crowds than the NWPU-Crowd data set and that our images are of higher resolution (3382 x 2702 pixel). Therefore, the relative size of objects is not similar yielding degraded results. For this reason, we propose a multi-resolution approach in inference only. The idea is to perform object counting on several discrete scales, whereas the results are finally fused. Therefore, potential under- and overestimation cancel out to a certain degree. The underlying rationale is illustrated in Figure 5, which depicts the distribution of object counts per mega pixel within the NWPU-Crowd data set. The range of our real data set is indicated by the red rectangle, which is an outlier to the input distribution. By changing the scale of our input data, the effective window can be shifted (orange arrows), whereas a shift to the left steers our images more towards the reference data set, thus increasing the counting accuracy.

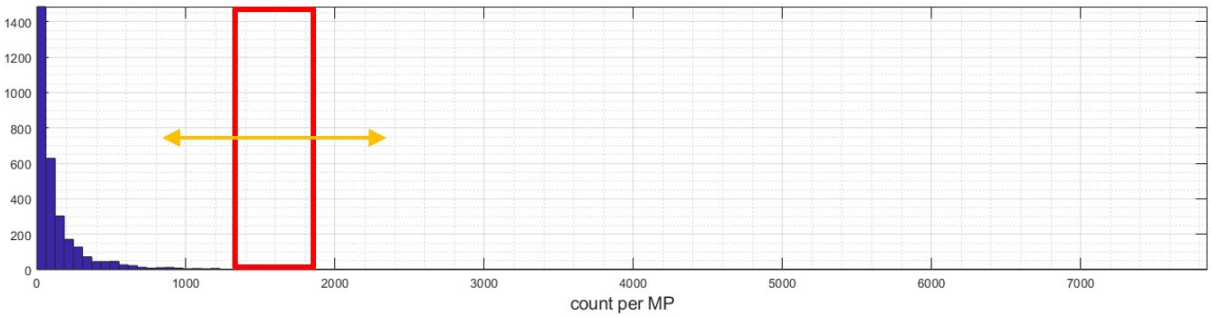

Figure 5: Distribution of the human count per megapixel over all images of the NWPU data set. The counts of our data set are much higher and are indicated by the red rectangle. The orange arrows represents that the distribution of our data can be shifted by appropriate rescaling.

In addition to the object count also the motion of humans is extracted via an optical flow estimation in image space (Chambolle and Pock, 2010; Almer et al., 2016b).

b. Geo-Referencing

As metric values are of high interest, human density and motion need to be transferred from image into object space. This can be easily done using a local coordinate system if the camera position (height over ground level) and camera tilt angle is known and approximately flat terrain can be assumed (as it is the case for the DIF23 data set). Projection of the camera field of view on the ground yields a trapezoid, which total area F can be calculated. This already allows to calculate a mean density value P [persons/m²] directly from the estimated human count N ($P = N/F$). However, to calculate local, high-resolution density values, it is necessary to project every single human head detection from the image into the ground plane (cf. Figure 6, middle, blue circles). In this step, also a mean human height needs to be considered to avoid displacements especially in the background. For every position on a regular 2D grid (e.g., 2x2 m²) within the camera footprint, the number of detected humans can be determined within a certain radius (e.g., 5 m) and local density values can be calculated. Using an intuitive colour scheme to depict low to high densities (blue to red colour), density distribution can be visualized, and density hotspots can be easily detected (cf. Figure 6, right).

To obtain metric velocity information (magnitude and direction), the same projection steps are performed for the starting and endpoints of the optical flow vector field estimated in image space. As for the density estimation this corrects for the perspective distortion of the image (displacements of a certain number of pixels in the foreground will result in a much lower velocity estimation in the background). Within the DIF23 data set, the crowd in the camera view was very homogeneous without significant motion, which results in velocity estimations around zero (not shown here).

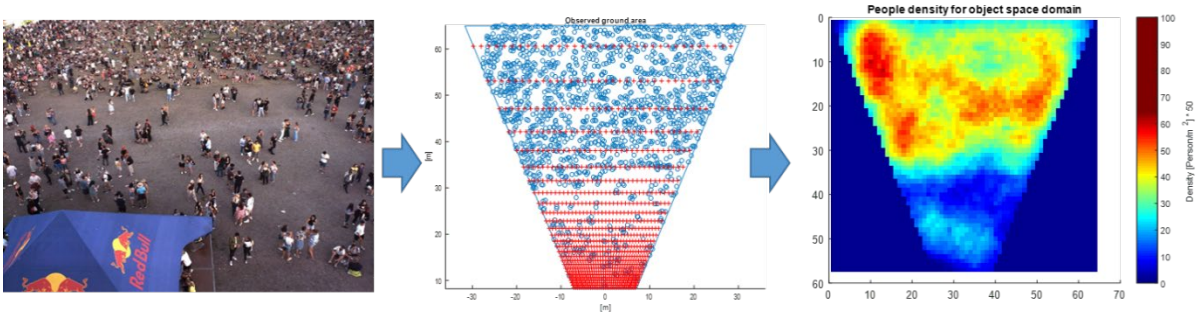

Figure 6: High-resolution density maps derived from estimated human head positions in the image projected into the ground plane (red lines: image lines @100 pixel spacing, emphasizing the non-equidistant distortion due to perspective geometry).

If the absolute position of the camera (e.g., by GNSS positioning) and also the viewing direction (azimuth) are known, the local map can be georeferenced in a global coordinate system (e.g., WGS84) and thus be viewed on top of a base map in a geospatial information system (see also Figure 9).

4. Results

a. Crowd Counting and Localization

In the first step, the accuracy of crowd counting was evaluated on two data sets, namely ShanghaiTech Part A (Zhang et al., 2016) and NWPU-Crowd (Wang et al., 2020), and for five AI methods, in particular MCNN (Zhang et al., 2016), CSRNet (Li et al., 2018), CAN (Liu et al., 2019), P2PNET (Song et al., 2021), and CLTR (Liang et al., 2022). Standard performance indicators are used, which are the mean absolute error (MAE) and the root mean square error (MSE), where smaller numbers indicate better performance. For the smaller data set ShanghaiTech Part A, P2PNET performs best, while on the large data set NWPU-Crowd the CLTR approach outperforms all other methods by a huge margin (cf. Table 1). For this reason, we stick to CLTR in this work.

Table 1: Performance indicators for two benchmarks and five AI-methods based on the validation set. Best results are marked in bold face.

architecture	ShanghaiTech Part A		NWPU-Crowd	
	MAE	MSE	MAE	MSE
MCNN (Zhang et al., 2016)	110.2	173.2	232.5	714.6
CSRNet (Li et al., 2018)	68.2	115.0	121.3	387.8
CAN (Liu et al., 2019)	62.3	100.0	106.3	386.5
P2PNET (Song et al., 2021)	**52.7**	**85.1**	77.4	362.0
CLTR (Liang et al., 2022)	56.9	95.2	**51.3**	**116.7**

In the second step, the CLTR AI model trained on NWPU-Crowd was utilized to predict the human count for our DIF23 data set. Results of single and multi-resolution approaches are depicted in Figure 7. Empirical tests show that three scales, corresponding to images dimensions where the image width is scaled to 1024, 2048, and 3072 pixels, yield best results.

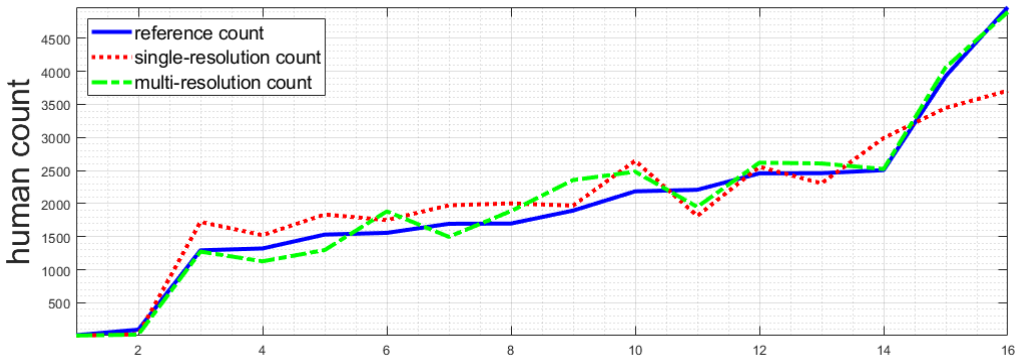

Figure 7: Counting results of 16 representative images gathered at the Donauinsel music festival. The blue line represents the manually annotated reference counts, the red dashed line the single-resolution results, and the green line our proposed multi-resolution results. Note, that our method yields relatively constant error rates also at very dense crowds, where the single-resolution approach degrades significantly.

Table 2 shows the accuracy metrics, where MAE and MSE are significantly better (actually better by a factor of 2). In addition, the mean normal absolute error (MNAE) is reduced by over 10%.

Table 2: Performance indicators on our DIF23 data for the standard single-resolution CLTR method and our proposed multi-resolution CLTR version. Best results are marked in bold face.

method	MAE	MSE	MNAE
Single-Resolution	342.3	446.8	25.9%
Multi-Resolution	**180.8**	**217.8**	**14.8%**

Exemplary results are given in Figure 8, where the predicted human locations are superimposed on the input images.

Reference count: 11 – estimated count: 6

Reference count: 1699 – estimated count: 1888

Reference count: 2460 – estimated count: 2607

Reference count: 4969 – estimated count: 4896

Figure 8: Visual examples of human detection on our DIF23 data set, with increasing crowd density (images have been downscaled and blurred for data protection reasons). The manually measured reference count and the AI-based estimated counts are given for each image, whereas the AI-based human locations are depicted as yellow circles.

b. Visualization within the Common Operational Picture

A COP was developed to allow a visualization of the human density on top of a base map. Figure 9 shows the georeferenced density map derived from the DIF23 data set with a temporal resolution of one minute and a spatial resolution of 1x1 m². Using the time slider on the bottom, changes in the crowd density can be monitored and further analyzed (four epochs shown). Density values are ranging from low (< 1 person/m², grey), middle (< 3 persons/m², green, magenta, orange) to high (> 6 persons/m², red).

Figure 9: Georeferenced human density maps for the 24th of June 2023 (density distributions between 7 pm and 9 pm on a 1x1 m² grid.

5. Conclusion

This study demonstrates the successful integration of deep learning techniques, specifically vision transformers, for accurate crowd counting and localization in urban scenarios. Utilizing the DIF23 dataset gathered at the Donauinselfest in 2023, the multi-resolution approach effectively addresses the challenge of dataset variability, significantly improving accuracy over single-resolution methods. The AI-generated human density maps, visualized in the COP, offer valuable insights for real-time decision-making in various urban operations, enhancing situational awareness for emergency responders and security personnel. The methods developed are promising for broader application across different urban environments and events, facilitating more effective crowd management and public safety measures.

6. 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; Schnabel, Thomas; Perko, Roland; Raggam, Hannes; Lukas, Sabine (2016a): Near real-time common operational picture (COP) for natural disaster management support. In Proceedings of *AGILE International Conference on Geographic Information Science*, number 19, pp. 1-3.

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

Chambolle, Antonin; Pock, Thomas (2010): A first-order primal-dual algorithm with applications to imaging. In *Journal of Mathematical Imaging and Vision*, 40, pp. 120-145.

Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; Houlsby, Neil (2020): An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, pp. 1-22.

Frering, Laurent; Köfler, Armin; Huber, Michael; Pfister, Sandra; Feischl, Richard; Almer, Alexander; Steinbauer-Wagner, Gerald (2023): Multi-Robot Support System for Fighting Wildfires in Challenging Environments: System Design and Field Test Report. In Proceedings of *IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 32-38.

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.

Idrees, Haroon; Saleemi, Imran; Seibert, Cody; Shah, Mubarak (2013): Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2547–2554.

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 Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091-1100.

Liu, Weizhe; Salzmann, Mathieu; Fua, Pascal (2019): Context-aware Crowd Counting. In Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5099–5108.

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

Perko, Roland; Schnabel, Thomas; Fritz, Gerald; Almer, Alexander; Paletta, Lucas (2013). Airborne based high performance crowd monitoring for security applications. In Proceedings of *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. In *Journal of Imaging*, 7(2), p. 21.

Perko, Roland; Ladstädter, Richard; Huber, Michael; Mustafic, Sead; Almer, Alexander; Klopschitz, Manfred (2023): Crowd Counting and Localization for Subsurface Operations. In Proceedings of *Urban Operations Expert Talks*, pp. 1-6.

Song, Qingyu; Wang, Changan; Jiang, Zhengkai; Wang, Yabiao; Tai, Ying; Wang, Chengjie; Li, Jilin; Huang, Feiyue; Wu, Yang (2021): Rethinking counting and localization in crowds: A purely point-based framework. In Proceedings of *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. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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 Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597.