Fraunhofer IOSB

FRAUNHOFER INSTITUTE FOR OPTRONICS, SYSTEM TECHNOLOGIES AND IMAGE EXPLOITATION IOSB

SPONSORED BY THE

Federal Ministry of Education and Research

Link to academic profile

TEMPORAL EXTENSION FOR ENCODER-DECODER-BASED CROWD COUNTING

Thomas Golda, Florian Krüger, Jürgen Beyerer

MOTIVATION

- In the past, mass events allured an increasing number of visitors. Hence, keeping an eye on the overall person count is crucial for safety
- Crowd manager have to keep an eye on the whole area all the time
- Pedestrian flow simulation beforehand is an important tool for event planning
- Image-based CC serves two purposes:
  - Simulations have to be varified
  - Crowd manager has to be supported

Density-based Crowd Counting in the Wild

- Videos come with temporal coherence which is rarely tackled by crowd counting methods
- Users get unsettled by jumping or non-steady counts for a static environment
- We examined how to use temporal information aiming for two improvements
  - Lower overall error
  - Smoother count estimates

MARE as Smoothness Measure

- Roughness ρ_i: standard deviation of succesive frame counts

$$\rho_i^2 := \frac{1}{m-1} \sum_{j=0}^{m-2} \left(\Delta_{i,j} - \overline{\Delta}_i\right)^2$$

$$\mathrm{MARE} := \frac{1}{l} \sum_{i=0}^{l-1} |\hat{\rho}_i - \rho_i|$$

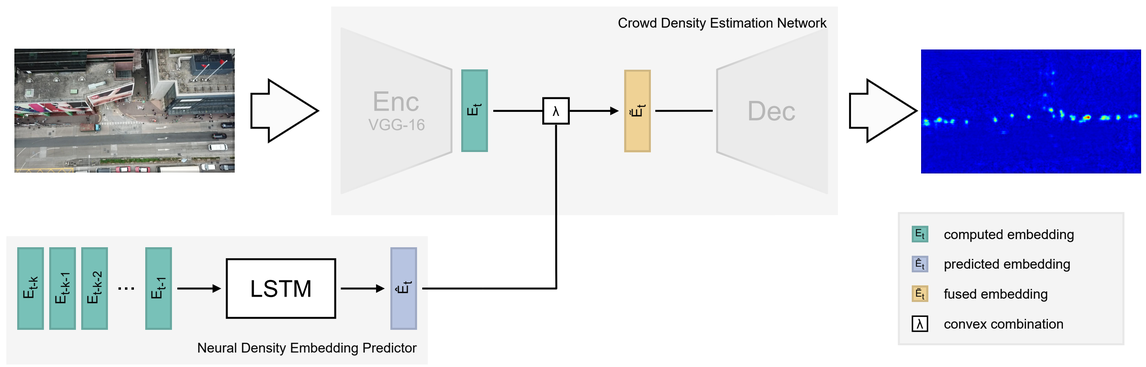

Figure 1: Visualization of the M2O architecture and complete pipeline. Both, the current embedding generated by the single frame processing and the temporal predicted embedding get fused to generate a new embedding. This newly constructed embedding is input to the existing decoder part of the CNN.

- MARE: average deviation from the overall trend of progression

Information on Experiments

- We evaluated our extension on three state of the art Encoder-Decoder architectures using the DroneCrowd [4] dataset
- All models were pretrained on DLR-ACD [2]
- Moving average brings no significant improvement for estimation performance but for smoothness
- Combination of recurrent unit and fixed merging generates best results for all adapted architectures

Conclusion

- Modelling temporal context inside model increases smoothness and accuracy of count estimates
- Gap between actual count and estimate can still differ by a quite large gap

RESULTS

Moving Avg.	Arch.	MAE	RMSE	MARE
k = 3	CSRNet	35.4 (-0.1)	43.2 (-0.2)	0.46 (-2.70)
	MRCNet	46.6 (-0.1)	58.2 (-0.1)	0.54 (-2.65)
	SFANet	39.4 (-0.2)	48.0 (-0.3)	**0.44 (-2.30)**
k = 5	CSRNet	35.3 (-0.2)	43.1 (-0.3)	0.63 (-2.53)
	MRCNet	46.5 (-0.2)	58.2 (-0.1)	0.67 (-2.52)
	SFANet	39.3 (-0.3)	47.8 (-0.5)	0.73 (-2.01)

Table 1: Three baseline models were trained from scratch on the DLR-ACD [2] dataset. Afterwards, all models were finetuned and evaluated on the DroneCrowd [4] dataset. The outputs were then processed by a moving average operator with window sizes of k = 3 and k = 5. This led to an improved performance for all models and all metrics. However, for the MAE and RMSE the improvement was not significant.

Method	Arch.	MAE	RMSE	MARE
TE-M2O	CSRNet [1]	**31.5 (-4.0)**	39.8 (-3.6)	2.61 (-0.55)
	MRCNet [2]	46.7 (±0.0)	59.8 (+1.5)	2.05 (-1.14)
	SFANet [3]	46.0 (+6.4)	55.5 (+7.2)	2.35 (-0.39)
MTE (fixed)	CSRNet	31.9 (-3.6)	**38.7 (-4.7)**	2.17 (-0.99)
	MRCNet	44.3 (-2.4)	56.9 (-1.4)	1.58 (-1.61)
	SFANet	33.2 (-6.5)	41.8 (-6.5)	1.82 (-0.92)

Table 2: The temporal extension achieved just for the CSRNet improvement for all three metrics, whereas for the MRCNet and SFANet the performance decreased. By adding the convex combination for merging the temporal and single frame embedding, these results could be improved. However, the moving average achieved better results regarding the smoothness of the results.

Thomas Golda, M. Sc.
Video Exploitation Systems (VID)
thomas.golda@iosb.fraunhofer.de
Phone +49 721 6091-631

References

[1] Li et al. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. International Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[2] Bahmanyar et al. MRCNet: Crowd Counting and Density Map Estimation in Aerial and Ground Imagery. BMVC Workshop on Object Detection and Recognition for Security Screening (BMVC-ODRSS), 2019.

[3] Zhu et al. Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting. arXiv:1902.01115, 2019.

[4] Du et al. VisDrone-CC2020: The Vision Meets Drone Crowd Counting Challenge Results. Computer Vision - ECCV 2020 Workshops (ECCV), 2020.

Karlsruhe Institute of Technology