SCALE-ADAPTIVE REAL-TIME CROWD DETECTION AND COUNTING FOR DRONE IMAGES

Markus Küchhold, Maik Simon, Volker Eiselein and Thomas Sikora

Communication Systems Group, Technische Universität Berlin
Einsteinufer 17, 10587 Berlin
{kuechhold,simon,eiselein,sikora}@nue.tu-berlin.de

ABSTRACT

We propose a scale-adaptive crowd detection and counting approach for drone images. Based on local feature points and density estimation considering the image scale, we detect dense crowds over multiple distances and introduce an extremely fast counting strategy with high accuracy for our detected crowd regions. We compare our results with a recent CNN-based state-of-the-art approach and validate both methods for different scaling factors on a novel crowd dataset. The results show that our proposed method outperforms the pretrained CNN-based approach and receives very precise counting results for different zoom factors, resolutions and crowd sizes. Its low computational complexity makes it highly suitable for real-time analysis or embedded systems.

Index Terms— crowd counting, crowd detection, drone, real-time, surveillance

1. INTRODUCTION

The analysis of public events such as concerts, fan parks or sports events has recently emerged as a very important research field. For security agencies, police or crisis management teams, it is a challenging task to ensure security and avoid critical situations such as panics due to overcrowding. The Duisburg Loveparade 2010 or the 2014 Shanghai stampede are prominent examples for catastrophes caused by inadequate overview and coordination during overcrowding situations. Crowd detection as well as crowd counting techniques can help to prevent such accidents by providing crucial information about the number of people and crowd density in a scene. An important factor here is the need for real-time analysis and a good overview. As street cameras usually have a small coverage area and often have been mounted for other purposes, video drones can be an alternative.

Various approaches have been proposed for counting of smaller crowds based on street cameras [1, 2, 3, 4, 5, 6] but do not enable monitoring of dense crowds from different viewing angles or distances, which is necessary for drone videos. A large-crowd approach for high-altitude aerial images (optimized for 1000m altitude) is demonstrated in [7]. The method applies a FAST feature detector to compute a density map and uses an image segmentation method to filter out non-crowd features. Afterwards, neighbourhood filtering with a fixed disc-shaped size is used for clustering close features and obtaining the person count. Next to traditional methods, also CNN-based approaches get more attention and achieve drastically lower error rates for crowd counting [8]. A promising and novel CNN-based approach for crowd counting is given in [9]. The Cascaded-MTL approach learns globally relevant discriminative features and computes a density map to estimate the total count of people in the image. It allows different viewing angles and outperformed recent state-of-the-art methods for the highly challenging ShanghaiTech dataset.

We propose a scale-adaptive real-time crowd detection and counting method for drone images (SARCCODI) with a viewing perspective related to real crowd monitoring use cases. Current German Regulation prohibits to fly directly over crowds, and a distance of at least the flying altitude to people has to be kept. Likewise, the altitude is restricted to a maximum of 100m. Thus, the typical viewing angle for drones is a 45 degrees bird's eye view causing occlusions and perspective distortions which have to be taken into account.

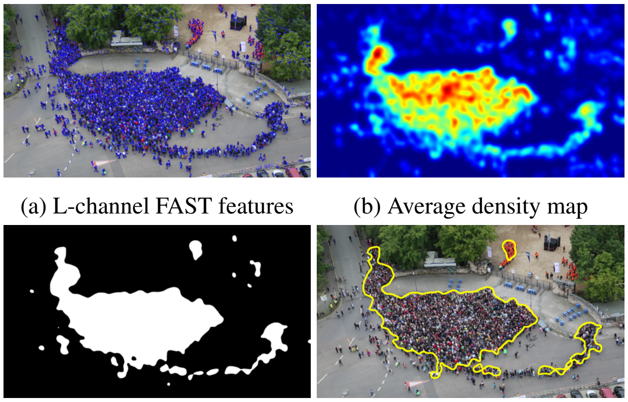
Similar as other approaches [7, 10, 11], SARCCODI is based on local feature points used as indication for the presence of crowds. In contrast to [7], we rely on features from the luminance channel which renders an additional image segmentation unnecessary and is thus faster. Kernel density estimation and thresholding allow for detection of dense crowds in the image. In order to deal with distortions by the viewing angle, a semi-automatic method using an affine transformation and a scale adaptation for multiple distances is used.

We compare our method with [7] and the CNN-based Cascaded-MTL approach [9] on pictures from a drone perspective which have been annotated manually for evaluation.

2. PROPOSED METHOD

2.1. Scale-Adaptive Crowd Detection

Following [7], we assume that FAST features [12] can serve as a basis for estimating initial crowd positions due



(a) L-channel FAST features (b) Average density map

(c) Segments by Otsu method (d) Filtered crowd segments

Fig. 1: Processing steps for crowd detection on a real-world image with queues at a concert venue.

to their ability to extract circular blob-like structures resembling human heads. Therefore, the input image is converted to CIELab color space and FAST features are computed on the L-channel (Fig. 1(a)). The resulting N points $\{x_i, y_i\}, i \in 1..N$ allow us to apply a probabilistic model to detect dense crowds: We compute a kernel density map over all feature points resulting in areas with high density values for dense crowd regions and areas with lower density indicating less people. The density value $p(x, y)$ for each pixel location is obtained by a discrete and bivariate Gaussian probability density function (pdf):

$$p(x,y) = \frac{1}{C}\sum_{i=1}^{N} exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma}\right) \quad (1)$$

where C is the normalization value to ensure $p(x, y) \in [0, 1]$.

Due to the camera angle, drone images are heavily affected by scale differences in the scene. Additionally, the scale between different frames can vary because the drone may start taking pictures at a certain height and then changes its altitude or position during the video. These factors are considered in our method by a) adapting σ to the camera view and b) applying a scale ratio to a pre-trained human height.

Using a user annotation step, the heights h_1, h_2, h_3 of three persons in the upper and lower picture border are determined and give an affine transformation M approximating a person's height $h_{est}(x, y)$ at any position in the image.

The bandwidth of the Gaussian kernel σ is then computed following the OpenCV[1] standard approach:

$$\sigma = 0.3 \cdot ((k_{size} - 1) \cdot 0.5 - 1) + 0.8 \quad (2)$$

using a frame-specific kernel window size

$$k_{size} = \frac{h_{est,cent\_low}}{h_{ref}} \cdot 101 = r_{zoom} \cdot 101 \quad (3)$$

[1] www.opencv.org

with $h_{est,cent\_low}$ as an estimate of a person's height at the lower picture center and h_{ref} as an according reference height obtained by training. r_{zoom} is the scale difference of the current frame to a pre-trained model. Our raw data is HD content, hence the rather large filter size. For robustness, the density maps (Fig. 1(b)) can be averaged over time.

In order to obtain the crowd position in the image, we apply Otsu's automatic thresholding [13] to the density map, resulting in a binary segmentation image (Fig. 1(c)). While mostly large crowds are of interest for an application, the result may also contain smaller segments. We filter out such false positives based on region size, expected number of features in a region and a minimally expected mean density. Fig. 1 (d) shows the final crowd detection result.

2.2. Crowd Counting Using the Image Scale

The previously segmented crowd regions in the image are further evaluated for counting. Generally, the FAST features in a crowd do not match the number of people in that region. Depending on the distance from the crowd and image resolution, several FAST features can be associated with one single person. To reduce unnecessary features, we use a grouping step with a circular shape and adaptively computed radius.

Firstly all feature points of crowd R_n will be permuted and for a randomly chosen feature at position (x_i, y_i) we compute, depending on the density $p(x_i, y_i)$ and the theoretical size of a person at (x_i, y_i), a specific radius $r(x_i, y_i)$. In order to avoid overfitting parameters to training images, this step is approximated by a linearization:

$$r(x_i, y_i) = (\alpha \cdot p(x_i, y_i) + \beta) \cdot r_{\text{sc}}(x_i, y_i) \cdot r_{zoom} \quad (4)$$

with the scaling factor r_{sc}

$$r_{sc}(x_i, y_i) = \frac{h_{est}(x_i, y_i)}{\max(h_1, h_2, h_3)} \quad (5)$$

accounting for the camera view and r_{zoom} as in (3). This allows us to use a single form (mostly) independent of the image scale. In our experiments, $\alpha = -16.41, \beta = 22.5$ have generated good results but may not extend to all use cases.

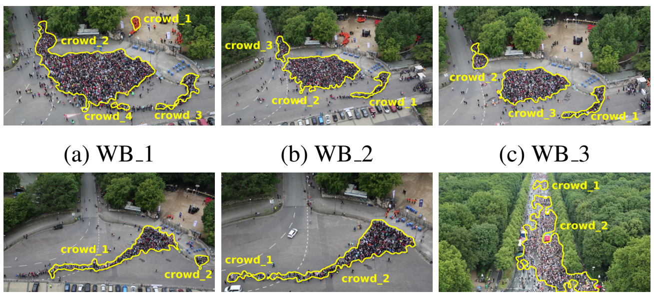
After assigning the radius, all other FAST features inside the disc shape are discarded and the steps are repeated until all features are processed. The number of circles then gives an estimate of the number of people in crowd R_n. In order to account for small variations between frames, we average the people count for each region R_n by buffering the counting results of previous frames. Therefore, regions are tracked over time using the intersection-over-union (IOU) principle.

3. EXPERIMENTS

Experimental validation is done on videos of two events from drone perspective which have been annotated manually. To

image	WB_1				WB_2				WB_3			
	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT
crowd segment		Part_A	Part_B			Part_A	Part_B			Part_A	Part_B	
1	27	37	16	33	125	93	45	108	119	77	50	99
2	1728	1582	893	1985	**1167**	1054	288	1223	131	**75**	22	72
3	79	77	63	96	**108**	77	22	112	**883**	710	439	880
4	**22**	14	15	21								
image	WB_4				WB_5				GS_1			
	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT
crowd segment		Part_A	Part_B			Part_A	Part_B			Part_A	Part_B	
1	580	369	196	526	33	14	19	20	83	52	15	72
2	50	37	24	40	**487**	342	258	485	**1319**	1096	686	1324

Table 1: Counting results on Full HD images - **bold** values indicate errors of less than 5% (GT: ground truth).



(a) WB_1 (b) WB_2 (c) WB_3

(d) WB_4 (e) WB_5 (f) GS_1

Fig. 2: Crowd detection results on Full HD resolution.

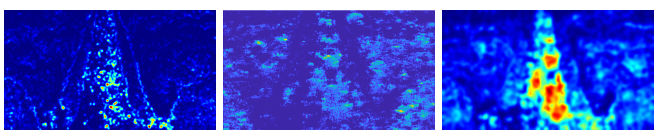


Fig. 3: Density maps for image GS_1 (Left: [7], Center: [9], Right: SARCCODI).

our knowledge, there are no accessible crowd datasets from drone perspective with ground truth (GT) for people counting. Our small set of test images is justified by the very time-consuming ground truth annotation for dense crowds. However, for future publications, we plan to release a dataset of our images with annotations for benchmarking purposes.

Fig. 2 shows results of our crowd detection for test images in HD resolution. SARCCODI is able to detect dense crowds for different zoom scales. Non-crowd areas like cars are left out, also the system is not trained to detect individuals. Nonetheless, we can detect crowds even in great distances as shown in Fig. 2(f).

Fig. 3 shows a comparison of SARCCODI's density maps with the feature-based approach from [7] and the Cascaded-MTL method [9]. For this method, we used the public models *Part_A* and *Part_B* [14] trained on the ShanghaiTech dataset for different viewing angles and scales. It can be seen that both methods estimate a significant crowd density in areas without people (e.g. on trees) while the main crowd is not segmented as a whole. It appears that especially the method from [7] optimized for higher camera altitudes rather founds salient color features than complete crowds. Therefore, for crowd counting, we will only consider the CNN-based approach from [9] which estimates the total number of people for a whole image without segmentation. To ensure a fair, segmentation-independent comparison, we thus multiply the density map from [9] with SARCCODI's segmentation of the currently considered crowd and leave out potential false detections in non-crowded areas reducing the accuracy of [9]. To obtain the ground truth for the considered crowd, we also use multiplication with the respective binary mask.

Results on HD images are shown in Tab. 1. For almost all crowds, SARCCODI outperforms both Cascaded-MTL models and achieves a much lower error (i.e. for the large crowds less than 1% on WB_3, WB_5 and GS_1). Smaller groups of less than 500 people are usually estimated with an error of less than ±20 people. Although trained for multiple viewing angles and scales, the Cascaded-MTL in general obtains much higher errors and only *Part_A* obtains acceptable results for very small crowds.

As an additional test, zooming has been simulated by downsampling the input images by a range of values from Full HD resolution to a scale of 0.25. Scale changes affect both the density estimate and also the segmentation when individual crowds in high resolution are merged to one single crowd in a scaled image (see Fig. 5). Our ground truth accounts for such segmentation changes.

Tab. 2 shows counting results of our scaling experiments using the biggest crowd in each image. Related error rates are shown in Fig. 4 (a-c). SARCCODI achieves mostly stable counting results for different scales in the range from HD720 to Full HD. Errors here are lower than 15%. In contrast, the error for Cascaded-MTL increases stronger for lower image resolutions and becomes higher than 50% for a scale of 0.5. This shows that our proposed scale-adaptive strategy works and is especially effective for scale changes up to 0.6. For lower resolution, the error increases due to a smaller number of FAST features.

The effect of the introduced scale-factor r_{zoom} can be seen in Fig. 4 (d). The red curve shows the counting error for different scales with a constant zoom factor $r_{zoom} = 1$

image	WB_1				WB_2				WB_3			
	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT
scaling factor		Part_A	Part_B			Part_A	Part_B			Part_A	Part_B	
1 (Full HD)	1728	1582	893	1985	**1167**	1054	288	1223	883	710	439	800
0.875	1736	1356	883	1972	1089	868	289	1224	859	606	403	797
0.75	1846	1326	565	2014	1096	819	143	1236	845	567	269	803
0.666 (HD720)	1887	1158	442	2022	1193	701	110	1360	**810**	505	160	803
0.625	1799	1186	407	2014	1092	704	97	1363	**769**	495	126	808
0.5	1743	1050	176	2135	980	528	26	1390	**783**	414	59	811
0.444 (FWVGA)	1740	812	131	2177	984	375	16	1407	724	317	32	827
0.375	1696	690	77	2229	841	292	13	1411	655	254	21	828
0.25	1221	328	7	2280	513	77	0	1410	387	77	4	835

image	WB_4				WB_5				GS_1			
	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT	SARCCODI	Cascaded MTL [9]		GT
scaling factor		Part_A	Part_B			Part_A	Part_B			Part_A	Part_B	
1 (Full HD)	580	369	196	526	**487**	342	258	485	**1319**	1096	686	1324
0.875	**527**	311	165	521	**478**	308	259	488	1121	882	509	1217
0.75	**535**	301	114	538	**497**	305	191	489	1101	825	346	1249
0.666 (HD720)	**568**	252	76	550	**510**	275	136	495	1122	718	212	1248
0.625	**564**	246	65	548	**519**	281	128	513	1108	682	207	1255
0.5	**557**	185	41	574	**484**	236	66	488	1132	527	109	1434
0.444 (FWVGA)	**583**	130	20	579	547	199	46	518	1159	420	67	1462
0.375	506	82	13	585	548	175	33	521	1074	320	51	1554
0.25	403	6	1	588	429	58	6	527	736	118	6	1579

Table 2: Counting results for different resolutions (errors less than 5% **bold**). GT varies with changing segmentation masks.

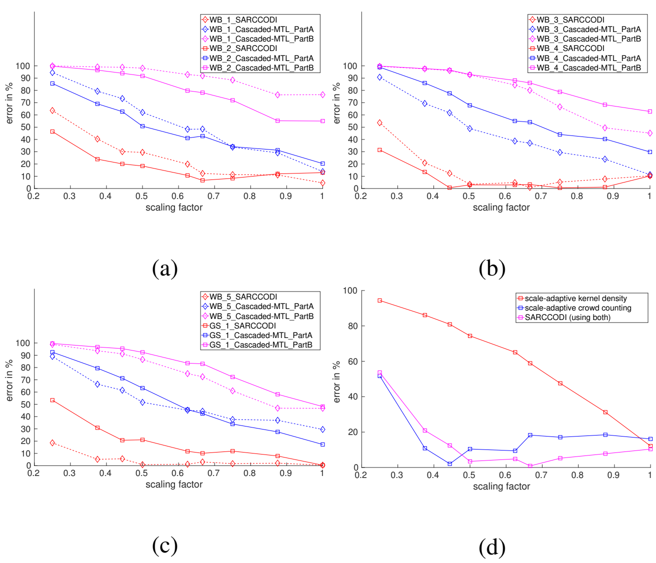
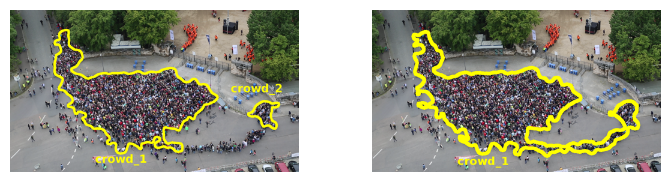


Fig. 4: Comparison of counting error for different scale factors (a-c). Error comparison for adaptive kernel bandwidth / adaptive counting strategy in SARCCODI on WB_3 (d).

in Eq. 4 and a varying r_{zoom} in Eq. 2, 3. The blue curve represents the opposite combination of a fixed bandwidth and scale-adaptive counting. The combination of both (pink) achieves a much lower counting error for most scales.

SARCCODI has a very low computational footprint and enables real-time crowd detection and counting for Full HD images. The currently required processing time of our complete single-threaded, non-optimized detection and counting method requires about 230ms per frame on a i7-7700 CPU @ 3.60GHz PC, including 6ms for the counting process.



(a) HD720 (b) FWVGA

Fig. 5: Scaling slightly changes the crowd detection results because individuals may be connected to crowds.

4. CONCLUSION

We proposed SARCCODI, a scale-adaptive crowd detection and counting method for drone images with real-time performance. Our approach outperforms the CNN-based Cascaded-MTL approach and is able to count extremely dense crowds with high precision. By introducing a scale-adaptive zoom-factor, we show that stable results can be achieved for a variety of different image scales which is important for drone applications. Thanks to its low computational complexity, SARCCODI could be run as an embedded method directly in drone systems. In our future work we plan to replace the semi-automatic scale adaptation through an automatic camera calibration system, which enables estimation of the zoom-factor without any additional human interaction.

5. ACKNOWLEDGEMENTS

The research leading to these results has received funding BMBF-VIP+ under grant agreement number 03VP01940 (SiGroViD).

6. REFERENCES

[1] Vincent Rabaud and Serge J. Belongie, "Counting crowded moving objects," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 705–711, 2006.

[2] A. B. Chan, Zhang-Sheng John Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–7.

[3] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 2913–2920.

[4] C. C. Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 2256–2263.

[5] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. Del Bimbo, "Real-time people counting from depth imagery of crowded environments," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2014, pp. 337–342.

[6] Tobias Senst, Volker Eiselein, Ivo Keller, and Thomas Sikora, "Crowd analysis in non-static cameras using feature tracking and multi-person density," in *21th IEEE International Conference on Image Processing*, Paris,France, Oct. 2014, pp. 6041–6045, ISBN: 978-1-4799-5750-7 DOI:10.1109/ICIP.2014.7026219.

[7] B. Sirmacek and P. Reinartz, "Automatic crowd density and motion analysis in airborne image sequences based on a probabilistic framework," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 898–905.

[8] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 589–597.

[9] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2017, pp. 1–6.

[10] B. Sirmacek and P. Reinartz, "Kalman filter based feature analysis for tracking people from airborne images," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVIII-4/W19, pp. 303–308, 2011.

[11] B. Sirmacek and P. Reinartz, "Automatic crowd analysis from airborne images," in *Proceedings of 5th International Conference on Recent Advances in Space Technologies - RAST2011*, June 2011, pp. 116–120.

[12] Edward Rosten and Tom Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[13] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[14] Vishwanath Sindagi, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting (single image crowd counting)," https://github.com/svishwa/crowdcount-cascaded-mtl, 2017.