Computer Vision Techniques for Crowd Counting

Andrew Aquilina
Department of Artificial Intelligence
University of Malta
Msida, Malta
andrew.aquilina.18@um.edu.mt

***Abstract*—The crowd counting task is defined as estimating the number of people in crowd images. It is characterised by a number of challenges, such as crowd occlusion and varied crowd density. Over the years, different approaches have been proposed, such as counting by detection, clustering or regression. However, modern methods employing the use of deep neural networks, have gained considerable attention due to their feasibility in handling the named challenges within densely crowded environments. Techniques developed for crowd counting can also be applied to a number of other tasks, such as vechile counting. In this report, two traditional crowd counting detection-based and two alternative deep-learning techniques are reviewed, evaluated and compared.**

***Index Terms*—crowd counting, convolution neural network, detection based**

I. Introduction

From individual person tracking [1] to crowd segmentation [2], the analysis of crowd dynamics opens up a wide spectrum of interests for the Computer Vision (CV) community. Crowd counting remains a non-trivial problem to this day, due to its real-time necessity in public safety and security applications. Unfortunately it is not a rare occurrence for tragedies to occur as a result of dense and large crowds [3]. Such incidents can be avoided by enforcing better crowd control through proper analysis. Additionally, the COVID-19 pandemic has also required businesses to find ways to comply with the restricted number of visitors. Sindagi and Patel [18] identified a number of challenges highly present within crowd counting: "occlusions, high clutter, non-uniform distribution of people, non-uniform illumination, intra-scene and inter-scene variations in appearance, scale and perspective". Fig. 1 outlines a few of these challenges. The structure of this report is as follows: Section II briefly reviews the considered traditional and modern techniques, Section III discusses the approach taken to compare and evaluate these techniques, and Section IV reports the produced results. Concluding remarks are made in Section V.

Fig. 1. Illustration of various crowd scenes [18].

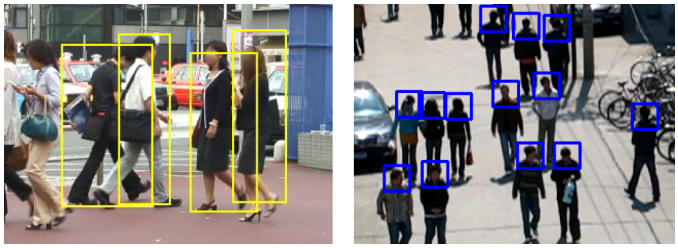

Fig. 2. Monolithic (left) and part-based (right) detection [4].

II. Literature Review

In this review, techniques are categorised either as traditional or modern, where modern methods employ the use of deep learning and Convolutional Neural Networks (CNN).

A. Traditional Methods

Loy et al. [4] survey a number of traditional crowd counting techniques and recognise three prominent categories: counting by detection, clustering and regression. In this report, the former will be discussed. Detection-based methods are straightforward, requiring a trained classifier to be applied in a sliding window fashion across some image. The types of features to gather depend on the choice of detection used.

Also known as the typical pedestrian approach, monolithic detection extracts human anatomy attributes through a selection of hand-crafted features. Gradient-based [5], edgelets [6], and Haar wavelets [7] are all viable options. Through these features, the full-body appearance of a pedestrian can be recognised, as shown in Fig. 2. Careful deliberation is instrumental when it comes to choosing the detector's features and classifier type, especially if to be used in a real-time scenario. Loy et al. [4] note that linear classifiers, such as Support Vector Machines (SVM) with a linear kernel, provide a decent trade-off between speed and quality. On the other hand, SVMs with a radial kernel are used for scenarios where quality is prioritised over speed. As one would presume, this would suffer in scenes where clutter is highly present [8]. This is called as the partial occlusion problem. Part-based detectors are used to overcome it.

By relaxing the assumption that the visibility of the whole body must be present, part-based detection methods yield better results in crowded scenes. A common implementation for a part-based detector involves the construction of a boosted classifier for specific body parts, such as the head or shoulder.

Gao et al. [9] note that the head region alone is not enough for reliable detection, concluding that including the upper body, as shown in Fig. 2, enables better performance. Thus it is no surprise that part-based detection has been adopted in many works [5, 10, 11].

Various other methods have been proposed to overcome the partial occlusion problem. Zhao et al. [12] and Ge and Collins [13] propose the use of a stochastic process to estimate the number of people in a scene through parametrised geometric shapes. Such work produces the count, location, but also the pose of the individuals within some scene. Yang, Guibas, et al. [14] and Ge and Collins [15] explore the potential of crowd counting methods through the use of multiple cameras. This helps in resolving scene uncertainties and benefit from improved detection speed. Unfortunately, such multi-sensor setup with overlapping views is rarely available in the real world. Another problem with generic pedestrian detectors is the requirement for scene training. Applying a detector to a new scene does not guarantee satisfactory results [8], and training scene by scene is impractical and time-consuming work. Over the years, transfer learning models have been introduced to address this issue [16, 17]. By exploiting scene structures and various view points, detector knowledge can be transferred from one scene to another without human supervision. It is to be said that despite advances over the years, detection-based crowd counting techniques suffer from dense crowds with non-trivial backgrounds.

B. Modern methods

Advances in CNN techniques continue to inspire various CV fields. The one at hand is no exception. Sindagi and Patel [18] categorise crowd counting models employing CNNs into four distinct types: basic, scale-aware, context-aware and multi-task. Basic CNNs were the first models to be proposed for crowd counting, such as the early work of Wang et al. [19]. Scale-aware models take in consideration the scenes' scale varations. This is usually achieved by multi-column CNNs [20]. Context-aware models incorporate the scenes' available contextual information, such as local count estimation [21]. Multi-task models focus on combining crowd counting along with other tasks, such as background subtraction [22]. The recent survey of Gao et al. [23] distinguishes CNN-based crowd counting techniques into three categories: Basic, Multi-column and Single-column. In this report, we will focus on two popular CNN models: MCNN and CSRNet. The former is a Scale-aware Multi-column CNN, proposed by Zhang et al. [20] while the latter is a CNN adopting dilated convolution layers, proposed by Li et al. [24].

1) MCNN: Zhang et al. [20] proposed a multi-column CNN architecture for images with a variety of crowd density and perspectives. As shown in Fig. 3, the network constitutes of three columns, each column corresponding to different sized filters (large, medium, small) so that the features learned by each column cater for the variation in people and head size. The input image of such model can be of any size,

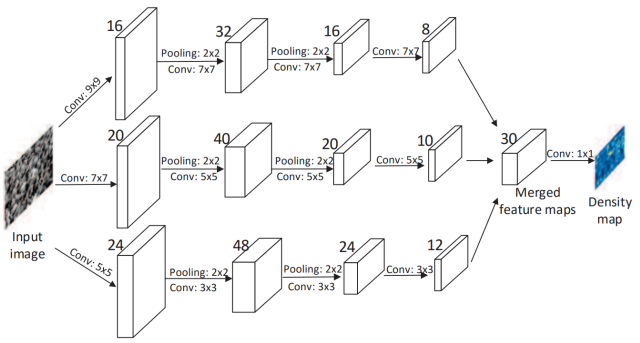

Fig. 3. Overview of MCNN architecture [20].

Fig. 4. Images from the *Shainghaitech* dataset and corresponding crowd density maps [20].

whereby the output is an estimate of crowd density from which the overall count can be extracted. This is achieved using geometry-adaptive kernels. The authors provide the following justifications for using density estimation maps for crowd count extraction: *a)* Density maps contain information which enable local area analysis, and, *b)* the CNN filter learning density maps can adapt to size variations. Zhang et al. also collect a large-scale crowd dataset named *Shainghaitech*, consisting of 1,200 images with around 330,000 annotations. The dataset is compromised of two parts. The first consists of images randomly crawled from the internet, and the second consists of images taken from the busy streets of China's biggest city, Shanghai. Fig. 4 shows sample images from this dataset and their respective crowd density maps. Using the mean absolute error (MAE) and the mean squared error (MSE) as evaluation metrics, experimentation on the *Shainghaitech* and other benchmark datasets using the MCNN architecture outperformed the state-of-the-art crowd counting methods.

2) CSRNet: Li et al. [24] proposed a CNN for Congested Scene Recognition, called CSRNet. It is composed of two major networks: one for feature extraction (front-end) and another using dilated kernels for crowd density map generation (back-end). It uses the VGG-16 architecture [25] for feature extraction to exploit its strong transfer learning ability and flexible architecture. The back-end utilises dilated convolution layers to provide an alternative to pooling layers. This expands the receptive fields while also maintaining its resolution. Previous research demonstrated the use of dilated convolutional layers to significantly improve accuracy [26], as shown in Fig. 5. The authors experimented their approach on five different public datasets, one of them being the gold-standard *Shainghaitech* dataset. They also use the MAE and the MSE evaluation metrics. Table I compares the estimation errors of the aforementioned CNN models on the *Shainghaitech* dataset.

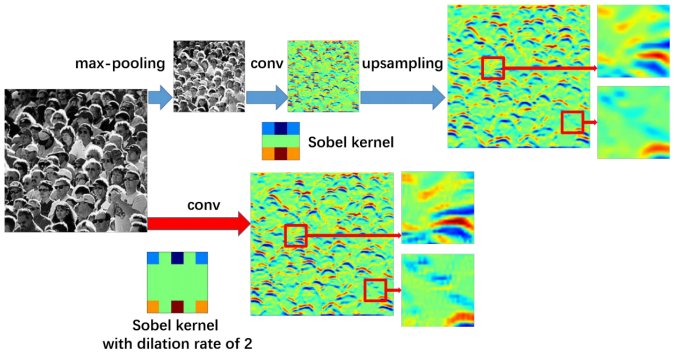

Fig. 5. The provided comparison by Li et al. [24] between pooling and using a dilated kernel.

	Part_A		Part_B	
Method	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3
CSRNet	68.2	115.0	10.6	16.0

TABLE I
MCCN AND CSRNET ESTIMATION ERRORS ON THE SHAINGHAITECH DATASET REPORTED BY LI ET AL. [24]

III. METHODOLOGY

The reviewed literature indicated that the MCNN and CSRNet models are excellent examples of the recent developments aimed to solve the crowd counting problem. Open source pre-trained implementations have been obtained for both MCNN[1] and CSRNet[2] to re-create the results reported in the work of Li et al. [24]. Furthermore, two traditional detection-based implementations, namely a linear SVM detector using Histogram of Oriented Gradients (SVM+HOG), and a Haar-cascade detector (HAAR), have been retrieved to note how these techniques have fared over the years. The former monolithically detects individuals, while the latter detects individuals by their upper-body. The OpenCV[3] library ships with a pre-trained HOG and Linear SVM pedestrian-detector model[4] and a trained upper-body Haar cascade[5]. The Python interface for the library was used to take advantage of these resources. An additional script, also making use of OpenCV, was developed to visually compare the performance of both pedestrian detectors.

IV. EVALUATION

A. Datasets

For this research, the *Shainghaitech* crowd dataset (discussed in Section II-B1) was obtained[6] to evaluate the aforementioned techniques. Fig. 6 highlights the non-uniform distribution of such dataset, as noted by Sindagi and Patel [18]. The considered traditional methods resulted in very poor performance using this dataset, due to the inability of properly

[1]https://github.com/svishwa/crowdcount-mcnn
[2]https://github.com/jaysondale/Size.AI-Deployment
[3]https://opencv.org/
[4]https://bit.ly/2WUhu4B
[5]https://bit.ly/3nTt63Y
[6]https://www.kaggle.com/tthien/shanghaitech

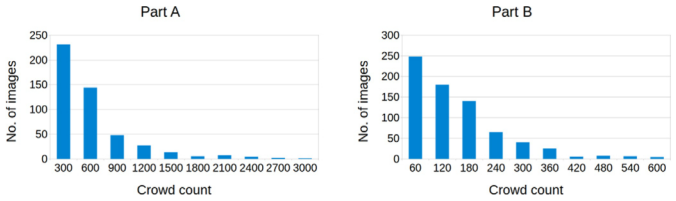

Fig. 6. Imbalanced distribution of crowd counts in the *Shanghaitech* dataset [18].

detecting individuals within dense crowds. As a result, a subset of another dataset *PETS2009*[7] was utilised to analyse the detection performance in sparse crowds. The subset taken was the S1 frame sequences, which contains three distinct sets (L1, L2, L3) used for Person Count and Density Estimation.

B. Results

As done in previous literature [4, 24, 26], the MAE and MSE metrics were chosen to evaluate the considered crowd counting methods. Fig. 7 depicts the considered methods' estimations errors, successfully reproducing known results [24].

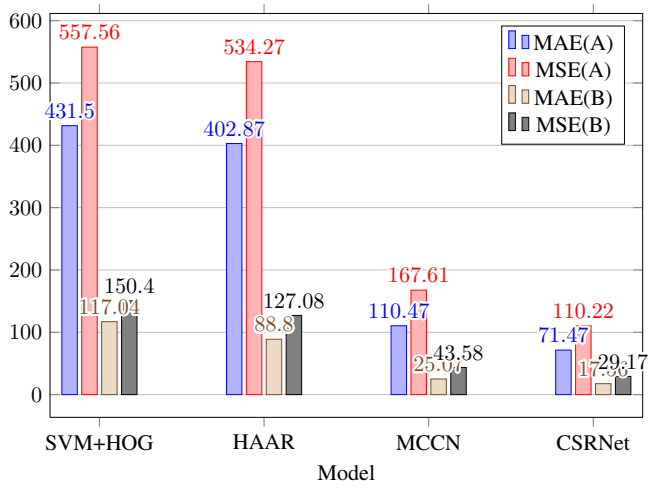

Fig. 7. Produced estimation errors on the *Shainghaitech* dataset.

Each frame from the *PETS2009*-S1 dataset was inputted to both SVM+HOG and HAAR pedestrian detectors, bounding each detected individual with a coloured rectangle. The outputted frames are concatenated and saved as a playable video, as to compare the performance of each detector side-by-side. Fig. 8 shows a cropped instance from one of the outputted videos generated by the developed script.

V. CONCLUSION

This study has focused on comparing two traditional and two modern crowd-counting techniques with each other. Notably, modern methods achieved better performance within the utilised benchmark dataset. CSRNet was found to be

[7]http://cs.binghamton.edu/~mrldata/pets2009

Fig. 8. Side by side comparison of SVM+HOG (green rectangles) and HAAR (red rectangles) from a particular frame of the *PETS2009* dataset.

the best model yielding satisfactory results for both parts. Findings show that the part-based method performed slightly better than the monolithic approach. It has been noted that recent developments in the field has superseded detector-based methods. The code and the datasets used in this study are available at https://github.com/wendru18/crowd-counting.

VI. References

[1] Chen Change Loy, Tao Xiang, and Shaogang Gong. "Time-delayed correlation analysis for multi-camera activity understanding". In: *International Journal of Computer Vision* 90.1 (2010), pp. 106–129.

[2] Mikel Rodriguez et al. "Density-aware person detection and tracking in crowds". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2423–2430.

[3] Dirk Helbing et al. "Simulation of pedestrian crowds in normal and evacuation situations". In: *Pedestrian and evacuation dynamics* 21.2 (2002), pp. 21–58.

[4] Chen Change Loy et al. "Crowd counting and profiling: Methodology and evaluation". In: *Modeling, simulation and visual analysis of crowds*. Springer, 2013, pp. 347–382.

[5] Bo Wu and Ramakant Nevatia. "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 1. IEEE. 2005, pp. 90–97.

[6] Payam Sabzmeydani and Greg Mori. "Detecting pedestrians by learning shapelet features". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.

[7] Paul Viola and Michael J Jones. "Robust real-time face detection". In: *International journal of computer vision* 57.2 (2004), pp. 137–154.

[8] Piotr Dollar et al. "Pedestrian detection: An evaluation of the state of the art". In: *IEEE transactions on pattern analysis and machine intelligence* 34.4 (2011), pp. 743–761.

[9] Chenqiang Gao et al. "People counting based on head detection combining Adaboost and CNN in crowded surveillance environment". In: *Neurocomputing* 208 (2016), pp. 108–116.

[10] Haroon Idrees, Khurram Soomro, and Mubarak Shah. "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning". In: *IEEE transactions on pattern analysis and machine intelligence* 37.10 (2015), pp. 1986–1998.

[11] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. "Estimation of number of people in crowded scenes using perspective transformation". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 31.6 (2001), pp. 645–654.

[12] Tao Zhao, Ram Nevatia, and Bo Wu. "Segmentation and tracking of multiple humans in crowded environments". In: *IEEE transactions on pattern analysis and machine intelligence* 30.7 (2008), pp. 1198–1211.

[13] Weina Ge and Robert T Collins. "Marked point processes for crowd counting". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 2913–2920.

[14] Danny B Yang, Leonidas J Guibas, et al. "Counting people in crowds with a real-time network of simple image sensors". In: *null*. IEEE. 2003, p. 122.

[15] Weina Ge and Robert T Collins. "Crowd detection with a multiview sampler". In: *European Conference on Computer Vision*. Springer. 2010, pp. 324–337.

[16] Meng Wang and Xiaogang Wang. "Automatic adaptation of a generic pedestrian detector to a specific traffic scene". In: *CVPR 2011*. IEEE. 2011, pp. 3401–3408.

[17] Meng Wang, Wei Li, and Xiaogang Wang. "Transferring a generic pedestrian detector towards specific scenes". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3274–3281.

[18] Vishwanath A Sindagi and Vishal M Patel. "A survey of recent advances in cnn-based single image crowd counting and density estimation". In: *Pattern Recognition Letters* 107 (2018), pp. 3–16.

[19] Chuan Wang et al. "Deep people counting in extremely dense crowds". In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 1299–1302.

[20] Yingying Zhang et al. "Single-image crowd counting via multi-column convolutional neural network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 589–597.

[21] Chong Shang, Haizhou Ai, and Bo Bai. "End-to-end crowd counting via joint learning local and global count". In: *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2016, pp. 1215–1219.

[22] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. "Counting in the wild". In: *European conference on computer vision*. Springer. 2016, pp. 483–498.

[23] Guangshuai Gao et al. "CNN-based Density Estimation and Crowd Counting: A Survey". In: *arXiv preprint arXiv:2003.12783* (2020).

[24] Yuhong Li, Xiaofan Zhang, and Deming Chen. "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1091–1100.

[25] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[26] Fisher Yu and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122* (2015).