Crowd Counting

Subjects: Computer Science, Information Systems
Contributor: Mostafa Mohamed

Crowd counting refers to estimating the number of individuals who share a certain region. This work present a survey for the main methods that calculate how many individuals are in a physical region. We start by reviewing traditional computer vision methods and then review deep learning-based methods.

Keywords: deep learning ; computer vision ; Crowd counting

1. Introduction

Automatic crowd scene analysis refers to investigating the behavior of a large group of people sharing the same physical area [1]. Typically, it counts the number of individuals per region, tracks the common individuals' trajectories, and recognizes individuals' behaviors. Therefore, automatic crowd scene analysis has many essential applications. It monitors the spread of the COVID-19 virus [2] via ensuring physical distance between individuals in stores, parks, etc. Securing public events, such as sports championships [3], carnivals [4], new year celebrations [5], and Muslim pilgrimage [6], is another application of automatic crowd scene analysis. Crowd scene analysis supplies surveillance camera systems with the ability to extract anomalous behaviors from a huge group of people [7][8][9]. Furthermore, analysis of crowd scenes of public places such as train stations, super stores, and shopping malls can show the effect of crowd path or the shortcomings of the design. Consequently, these studies can better safety considerations [10][11].

2. Crowd Counting

The following subsections review different methods that calculate how many individuals are in a physical region. For completeness, we start by reviewing traditional computer vision methods and then review deep learning-based methods.

2.1. Traditional Computer Vision Methods

2.1.1. Detection-Based Approaches

Early approaches used detectors to detect peoples' heads or shoulders in the crowd scene to count them, such as in [12] [13]. Counting by detection is usually performed either in monolithic detection or parts-based detection. In monolithic detection, the detection is usually performed based on pedestrian detection methods such as optical flow [[14], histogram of oriented gradient (HOG) [15], Haar wavelets [16], edgelet [17], Particle flow [18], and shapelets [19]. Subsequently, the extracted features from the former detectors are fed into nonlinear classifiers such as Support Vector Machine (SVM) [20]; however, the speed is slow. A linear classifier such as linear SVM, hough forests [21], or boosting [22] usually provides a trade-off between speed and accuracy. Then, the classifier is slid over the whole image to detect candidates and to discard the less confident candidates. The results of sliding give the number of people in the scene.

The former methods cannot deal with the partial occlusion problem [23] when it is raised; therefore, part-based detection is adopted. Part-based detection focuses on body parts rather than the whole body such as the head and shoulders as in [13]. Part-based detection is more robust than monolithic, as reported in [13]. Based on 3D shapes [24], humans were modelled with ellipsoids, which was employed as a stochastic process [25] to calculate the number and shape configuration that best explains a segmented foreground object. Later on, Ge et. al [26] extended the same idea with the Bayesian marked point process (MPP) [27] with a Bernoulli shape prototype [28]. Zhao et al. [29] used Markov chain Monte Carlo [30] to exploit temporal coherence for 3D human models across consecutive frames.

2.1.2. Regression-Based Approaches

Although counting by detection or part-based approaches achieves reasonable results, it fails in very crowded scenes and under heavy occlusion. Counting by regression tries to mitigate the former problems. Typically, this method consists of two main components. The first component is extracting low-level features, such as Foreground features [31], texture [32], edge

features [33], and gradient features [34]. The second component is mapping in a regression function, e.g., linear regression [35], piecewise linear regression [36], ridge regression [37], or Gaussian process regression, to map the extracted features into counts, as in [35]. The complete pipeline of this method is shown in Figure 1.

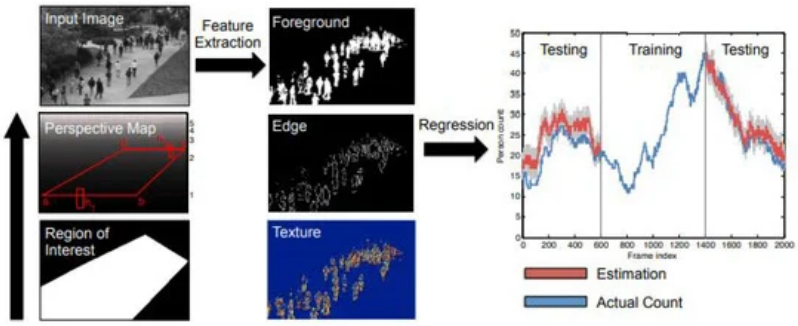

Figure 1. Crowd counting pipeline using the regression model. Image from [38].

2.1.3. Density Estimation-Based Approaches

These approaches build a density map to represent the number of individuals per region in an input image, as shown in Figure 2. In [39], the author built density maps via linearly mapping local patch features to its corresponding object. Formulating the problem in this way reduces the complexity of separating each object to count it and reduces the potential of counting errors in case of highly crowded scenes. Estimating the number of objects in this method equates to integration over local batches in the entire image.

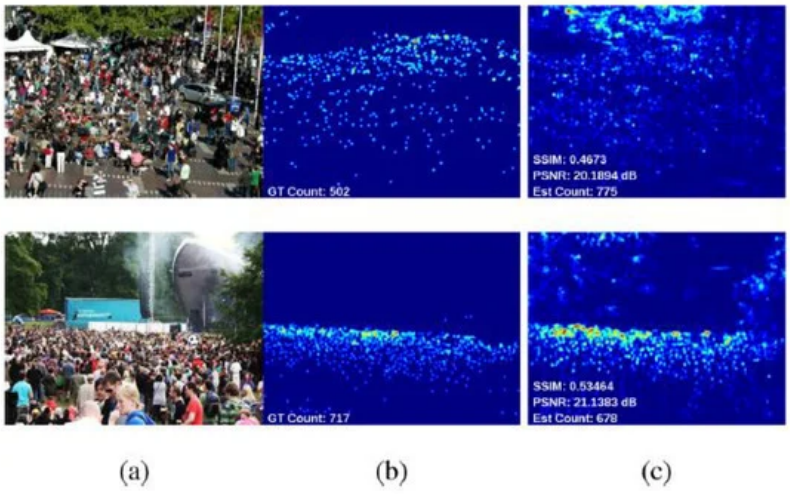

Figure 2. (**a**) Input image, (**b**) Ground truth, and (**c**) Estimated density maps. Image from [38].

2.2. Deep Learning Approaches

Convolutional Neural Networks (CNNs) are similar to plain Neural Networks (NNs) from the perspective that they consist of neurons/receptive fields that have learnable weights and biases. Each receptive field receives a batch input and performs a convolution operation, and then, the result is fed into a nonlinearity function [40] (e.g., ReLU or Sigmoid). The input image to CNN is assumed to be an RGB image; therefore, the hidden layers learn rich features that contribute to the performance of the whole network (hidden layers and classifier). This structure has benefits in terms of speed and accuracy since the crowd scene images have lots of objects that need computationally expensive operations to detect. End-to-end networks mean the network takes the input image and directly produces the desired output.

The pioneering work with deep networks was proposed in [41]. An end-to-end deep convolutional neural network (CNN) regression model for counting people of images in extremely dense crowds was proposed. A collected dataset from Google and Flickr was annotated using a dotting tool. The dataset consists of 51 images, each of which has 731 people on average. The least number of counts in this dataset is 95, and the highest count is 3714. The network was trained on positive and negative classes. The positive images were labelled with the number of the objects, while the negative images were labelled with zero.

Network architecture: This network consists of five convolutional layers and two fully connected layers. The network was trained on object classification with regression loss, as shown in Figure 3.

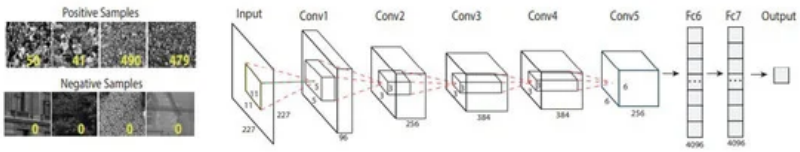

Figure 3. Convolutional Neural Network (CNN) architecture with positive and negative inputs. Image from [41].

One of the newest state-of-the-art methods for accurate crowd counting came out in [42]. The authors proposed an attention-injective deformable convolutional network called ADCrowdNet that they claim can work accurately in congested noisy scenes. The network consists of two sections: Attention Map Generator (AMG) and Density Map Estimator (DME). AMG is a classification network that classifies the input image into a crowd image or background image. The product of AMG is then used as input to DME to generate a density map of the crowd in the frame. This process is described in Figure 4. ADCrowdNet achieved the best accuracy for crowd counting on the ShanghaiTech dataset [43], UCF_CC_50 dataset [44], the WorldExpo'10 dataset [45], and the UCSD dataset [35]. In [46], Oh et al. proposed an uncertainty quantification method for estimating the count of the crowd. This method is based on a scalable neural network framework that uses a bootstrap ensemble. Method PDANet (Pyramid Density-Aware Attention-based network) [47] generates a density map representing the count of the crowd in each region of input images. This density map is generated by utilizing the attention paradigm, pyramid scale features, decoder modules for crowd counting, and a classifier to assess the density of the crowd in each input image. In DSSINet (Deep Structured Scale Integration Network) [48], structured feature representation learning and hierarchically structured loss function optimization are used to count the crowd. In [49], Reddy et al. tackled the problem of crowd counting by adaptive few-shot learning. In [50], an end-to-end trainable deep architecture was proposed. This approach uses contextual information, generated by multiple receptive field sizes and learning the importance of each such feature at each image location, to estimate the crowd count in input images.

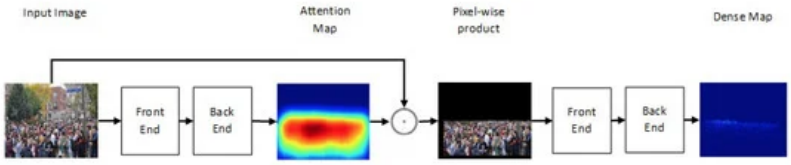

Figure 4. Structure of ADCrowdNet.

References

1. Musse, S.R.; Thalmann, D. A model of human crowd behavior: Group inter-relationship and collision detection analysis. In Computer Animation and Simulation'97; Springer: Berlin/Heidelberg, Germany, 1997; pp. 39–51.
2. Watkins, J. Preventing a Covid-19 Pandemic. 2020. Available online: https://www.bmj.com/content/368/bmj.m810.full (accessed on 8 May 2012).
3. Jarvis, N.; Blank, C. The importance of tourism motivations among sport event volunteers at the 2007 world artistic gymnastics championships, stuttgart, germany. J. Sport Tour. 2011, 16, 129–147.
4. Da Matta, R. Carnivals, Rogues, and Heroes: An Interpretation of the Brazilian Dilemma; University of Notre Dame Press Notre Dame: Notre Dame, IN, USA, 1991.
5. Winter, T. Landscape, memory and heritage: New year celebrations at angkor, cambodia. Curr. Issues Tour. 2004, 7, 330–345.
6. Peters, F.E. The Hajj: The Muslim Pilgrimage to Mecca and the Holy Places; Princeton University Press: Princeton, NJ, USA, 1996.
7. Cui, X.; Liu, Q.; Gao, M.; Metaxas, D.N. Abnormal detection using interaction energy potentials. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20 June 2011; pp. 3161–3167.
8. Mehran, R.; Moore, B.E.; Shah, M. A streakline representation of flow in crowded scenes. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2010; pp. 439–452.
9. Benabbas, Y.; Ihaddadene, N.; Djeraba, C. Motion pattern extraction and event detection for automatic visual surveillance. J. Image Video Process. 2011, 7, 163682.

10. Chow, W.K.; Ng, C.M. Waiting time in emergency evacuation of crowded public transport terminals. Saf. Sci. 2008, 46, 844–857.
11. Sime, J.D. Crowd psychology and engineering. Saf. Sci. 1995, 21, 1–14.
12. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. IEEE Trans. Patt ern Anal. Mach. Intell. 2012, 34, 743–761.
13. Li, M.; Zhang, Z.; Huang, K.; Tan, T. Estimating the number of people in crowded scenes by mid based foreground seg mentation and head-shoulder detection. In Proceedings of the 19th International Conference on Pattern Recognition (I CPR 2008), Tampa, FL, USA, 8 December 2008; pp. 1–4.
14. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2004; pp. 25–36.
15. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA, June 20 2005; Volume 1, pp. 886–893.
16. Viola, P.; Jones, M.J. Robust real-time face detection. Int. J. Comput. Vis. 2004, 57, 137–154.
17. Wu, B.; Nevatia, R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edge let part detectors. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05), Beijing, Ch ina, 17 October 2005; Volume 1, pp. 90–97.
18. Ali, S.; Shah, M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In Procee dings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 22 June 200 7; pp. 1–6.
19. Sabzmeydani, P.; Mori, G. Detecting pedestrians by learning shapelet features. In Proceedings of the Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, USA, 17 June 2007; pp. 1–8.
20. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2011, 2, 1–27.
21. Gall, J.; Yao, A.; Razavi, N.; Van Gool, L.; Lempitsky, V. Hough forests for object detection, tracking, and action recognit ion. IEEE Trans. Pattern Anal. Mach. Intell. 2011, 33, 2188–2202.
22. Viola, P.; Jones, M.J.; Snow, D. Detecting pedestrians using patterns of motion and appearance. Int. J. Comput. Vis. 20 05, 63, 153–161.
23. Zhang, T.; Jia, K.; Xu, C.; Ma, Y.; Ahuja, N. Partial occlusion handling for visual tracking via robust part matching. In Pro ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24 June 2014; p p. 1258–1265.
24. Kilambi, P.; Ribnick, E.; Joshi, A.J.; Masoud, O.; Papanikolopoulos, N. Estimating pedestrian counts in groups. Comput. Vis. Image Underst. 2008, 110, 43–59.
25. Whitt, W. Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues; Spri nger Science & Business Media: Berlin/Heidelberg, Germany, 2002.
26. Ge, W.; Collins, R.T. Marked point processes for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20 June 2009; pp. 2913–2920.
27. Chatelain, F.; Costard, A.; Michel, O.J. A bayesian marked point process for object detection: Application to muse hyper spectral data. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22 May 2011; pp. 3628–3631.
28. Juan, A.; Vidal, E. Bernoulli mixture models for binary images. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, 26–26 August 2004; Volume 3, pp. 367–370.
29. Zhao, T.; Nevatia, R.; Wu, B. Segmentation and tracking of multiple humans in crowded environments. IEEE Trans. Pat tern Anal. Mach. Intell. 2008, 30, 1198–1211.
30. Geyer, C.J. Markov Chain Monte Carlo Maximum Likelihood; Interface Foundation of North America: Fairfax Station, V A, USA, 1991.
31. Bouwmans, T.; Silva, C.; Marghes, C.; Zitouni, M.S.; Bhaskar, H.; Frelicot, C. On the role and the importance of feature s for background modeling and foreground detection. Comput. Sci. Rev. 2018, 28, 26–91.
32. Tuceryan, M.; Jain, A.K. Texture analysis. In Handbook of Pattern Recognition and Computer Vision; World Scientific: S ingapore, 1993; pp. 235–276.
33. Mikolajczyk, K.; Zisserman, A.; Schmid, C. Shape rEcognition With Edge-Based Features. 2003. Available online: http s://hal.inria.fr/inria-00548226/ (accessed on 11 September 2020).

34. Hwang, J.W.; Lee, H.S. Adaptive image interpolation based on local gradient features. IEEE Signal Process. Lett. 200 4, 11, 359–362.
35. Chan, A.B.; Liang, Z.S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people model s or tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anch orage, AK, USA, 24 June 2008; pp. 1–7.
36. Paragios, N.; Ramesh, V. A mrf-based approach for real-time subway monitoring. In Proceedings of the 2001 IEEE Co mputer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8 December 2 001.
37. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature mining for localised crowd counting. In Proceedings of the British Mac hine Vision Conference; BMVA Press: Surrey, UK, 2012; Volume 1, p. 3.
38. Loy, C.C.; Chen, K.; Gong, S.; Xiang, T. Crowd counting and profiling: Methodology and evaluation. In Modeling, Simul ation and Visual Analysis of Crowds; Springer: Berlin/Heidelberg, Germany, 2013; pp. 347–382.
39. Lempitsky, V.; Zisserman, A. Learning to count objects in images. In Advances in Neural Information Processing Syste ms, Proceedings of the Neural Information Processing Systems 2010, Vancouver, BC, Canada, 6 December 2010; Neu ral Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2010; pp. 1324–1332.
40. Karlik, B.; Olgac, A.V. Performance analysis of various activation functions in generalized mlp architectures of neural ne tworks. Int. J. Artif. Intell. Expert Syst. 2011, 1, 111–122.
41. Wang, C.; Zhang, H.; Yang, L.; Liu, S.; Cao, X. Deep people counting in extremely dense crowds. In Proceedings of the 23rd ACM International Conference on Multimedia; ACM: New York, NY, USA, 2015; pp. 1299–1302.
42. Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; Wu, H. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long B each, CA, USA, 16–20 June 2019; pp. 3225–3234.
43. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural netw ork. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Caesars Palace, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 589–597.
44. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source multi-scale counting in extremely dense crowd images. In Pro ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23 June 2013; pp. 2547–2554.
45. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 833 –841.
46. Oh, M.H.; Olsen, P.A.; Ramamurthy, K.N. Crowd counting with decomposed uncertainty. In Proceedings of the Associat ion for the Advancement of Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 11799–11806.
47. Amirgholipour, S.; He, X.; Jia, W.; Wang, D.; Liu, L. PDANet: Pyramid Density-aware Attention Net for Accurate Crowd Counting. arXiv Preprint 2020, arXiv:2001.05643.
48. Liu, L.; Qiu, Z.; Li, G.; Liu, S.; Ouyang, W.; Lin, L. Crowd counting with deep structured scale integration network. In Pro ceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 1774–1783.
49. Reddy, M.K.K.; Hossain, M.; Rochan, M.; Wang, Y. Few-shot scene adaptive crowd counting using meta-learning. In Pr oceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 Marc h 2020; pp. 2814–2823.
50. Liu, W.; Salzmann, M.; Fua, P. Context-aware crowd counting. In Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition, Long Beach, CA, USA, June 16 2019; pp. 5099–5108.