Crowd Scene Analysis for Zeyarat Al-Arabaeen: A Comprehensive Literature Survey

Faisel G. Mohammed

University of Baghdad – Collage of Science College – Dep., of Remote Sensing & GIS

faisel.mohammed@sc.uobaghdad.edu.iq

Noor N. Thamer

Ministry of Education, General Directorate of Education Iraq

noor.thamir1201@sc.uobaghdad.edu.iq

Abstract:

Understanding how people behave in crowded places is an important endeavor with several uses, like controlling the spread of COVID-19 and boosting security. In-depth study of crowd scene analysis methods, including both crowd counting and crowd activity detection, is included in this survey article. This article fills the gap by exhaustively examining the spectrum up to contemporary deep learning techniques, whereas current studies frequently focus primarily on certain aspects or traditional approaches. Paper proposes the innovative idea of Crowd Divergence (CD) evaluation as a matrix for evaluating crowd scene analysis approaches, which was motivated by information theory. Contrary to conventional measurements, CD quantifies the agreement between expected and observed crowd count distributions. This paper makes three key contributions: an examination of readily available crowd scene datasets, the use of CD for thorough technique evaluation, and a thorough examination of crowd scene methodologies. The investigation starts with conventional computer vision methods, closely examining density estimates, detection, and regression strategies. Convolutional neural networks (CNNs) become effective tools as deep learning progresses, as seen by new models like ADCrowdNet and PDANet, which make use of attention mechanisms and structured feature representation. To evaluate algorithmic effectiveness, a variety of benchmark datasets including ShanghaiTech, UCF CC 50, and UCSD are carefully examined. Computer vision's exciting and challenging topic of "crowd scene analysis" has numerous applications, from crowd control to security surveillance. This survey article offers a comprehensive viewpoint on crowd scene analysis, bringing several approaches under a single heading and presenting the CD measure to guarantee reliable assessment. This article provides a complete resource for researchers and practitioners alike through an elaborate investigation of methods, datasets, and cutting-edge evaluation approaches, paving the way for improved crowd scene analysis techniques across a variety of fields.

Keywords: Crowd behavior analysis, Crowd scene methodologies, Crowd Divergence (CD) evaluation, Deep learning techniques, Benchmark datasets, Crowd control and security.

1. Introduction

The study of crowd scene analysis involves examining the behavior of groups of people in the same physical area [1]. It typically includes counting the number of individuals, in regions tracking their movements and identifying their behaviors. This type of analysis has applications. One such application is controlling the spread of COVID 19 by ensuring distancing in places like stores and parks [2]. It also plays a role in ensuring security during events such as sports championships, carnivals, New Year celebrations and Muslim pilgrimages [3 6]. Automatic crowd scene analysis enables surveillance camera systems to detect behaviors within groups of people [7–9]. Additionally analyzing crowd scenes in places like train stations, supermarkets and shopping malls can provide insights, into crowd movement patterns. Identify design flaws. These studies contribute to improving safety considerations [10,11].

As was previously shown, it is crucial to analyze crowd scenes, hence various survey papers have been suggested. However, the survey articles now in publication either compel the use of conventional computer vision techniques for the analysis of crowd situations or focus on just one component of crowd analysis, such as crowd counting [12]. This survey paper aims to provide an in-depth analysis of the development of crowd scene analysis techniques up to the most modern deep learning [13] techniques. The two key components of crowd analysis are (1) crowd counting and (2) crowd activity recognition, which are both included in this survey.

Additionally, this study suggests the crowd divergence (CD) evaluation matrix for crowd scene analysis techniques, which is motivated by information theory. When compared to popular evaluation matrices like mean squared error (MSE) [14] and mean absolute error (MAE) [15], CD provides an accurate assessment of how well the predicted crowd count distribution corresponds to the real distribution. The suggested metric calculates the difference between the estimated and actual counts.

This study offers three contributions:

1. Examine the datasets for crowd scenes that are accessible.
2. recommending crowd divergence (CD) for a thorough assessment of crowd scene analysis techniques.
3. Analyzing crowd scene analyzing techniques using deep learning.

The remainder of the survey is structured as follows. The crowd counting approach is discussed in Section 2. The crowd scene datasets are reviewed in Section 2. The paper is discussed in Section 4 of the document. Section 5 conclusion and future directions are noted it.

2. Crowd Counting

Crowd counting refers to determining how many people reside in a specific area. The subsections go through various approaches to estimating the population density of a given geographic area. To be thorough, we first discuss conventional computer vision techniques before reviewing deep learning-based techniques.

2.1. Traditional Computer Vision Methods

2.1.1. Density Estimation-Based Approaches

As seen in Figure 2, these techniques create a density map to depict the number of people per region in an input image. Through the linear mapping of local patch features to their respective objects, the author of [16] constructed density maps. The complexity of isolating each thing to count it and the possibility of counting errors in situations with a lot of objects are both reduced by formulating the problem in this way. This approach integrates over local batches in the entire image to estimate the number of objects.

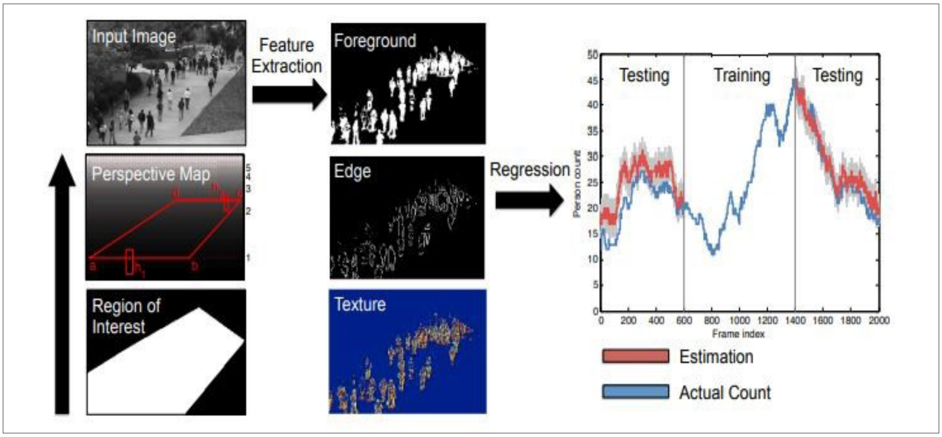


Figure:1. Crowd counting pipeline using regression model. Image from [17]

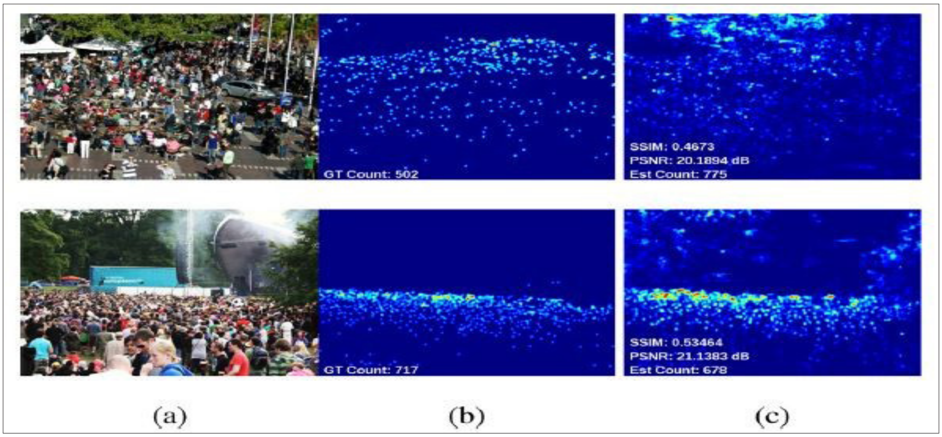


Figure:2. (a) Input image, (b) Ground truth, and (c) Estimated density maps. Image from [17].

A loss function that optimizes the regularized risk quadratic cost function was used to build the density map in .[18] Cutting-plane optimization was used to complete the solution .[19] The work in[18] was improved by Pham et al .in [50] by learning nonlinear mapping. They voted on density of several target items using random forest regression .[21 ,20] They also achieved real-time performance ,and in place of mapping dense features and creating a density map ,they computed the embedding of subspaces created by picture patches.

An approach for density estimation that is scale -and resolution-invariant was proposed by Sirmacek et al .To determine the probability density functions) pdfs [22] (of various places in successive frames, this technique uses Gaussian symmetric kernel functions .[23] The value of the generated pdfs is then used to estimate the number of persons per spot .The three primary categories of the conventional crowd counting approach are listed in Table.1

2.1.2.Detection-Based Approaches

Early methods ,like those in ,[24,25] relied on detectors to find people's heads or shoulders in crowd scenes to count them .Typically, monolithic detection or parts-based detection are used for counting by detection .For monolithic detection ,pedestrian detection techniques including optical flow ,[26] histogram of oriented gradient) HOG(,[27]Haar wavelets ,[28] edgelet ,[29] particle flow ,[30] and shapelets [31]are typically used as the foundation for the detection .The former detectors 'collected characteristics are then input into nonlinear classifiers like the Support Vector Machine) SVM ,[32] (however the pace is poor .A linear classifier that offers a trade-off between speed and accuracy is typically linear SVM ,hough forests ,[33] or boosting .[34] The classifier is then moved across the entire image to identify candidates and exclude the less confident ones .The results of sliding reveal how many persons are present.

When the partial occlusion problem [35] arises ,the earlier approaches are unable to handle it ;as a result ,part-based detection is used .Instead of focusing on the entire body ,like the head and shoulders as in ,[25] part-based detection concentrates on specific body components .According to ,[25] part-based detection is more reliable than monolithic .Humans were modeled using ellipsoids based on3 D shapes ,[36] and a stochastic approach was used to determine the number and shape configuration that best explains a segmented foreground item .[37] Later ,Ge et al [38] expanded the same concept using a Bernoulli form prototype [39] and a Bayesian marked point process) MPP .[40] (The Markov chain Monte Carlo was utilized by Zhao et al [41] .to take advantage of temporal coherence for3 D human models across consecutive frames.

2.1.3.Regression-Based Approaches

Even if counting by detection or part-based methods produces acceptable results ,they fall short in densely populated areas and where there is significant occlusion .Regression counting tries to address the prior issues .This approach typically consists of two key parts .Extraction of low-level features ,such as foreground features

,[42,43]texture ,[45] edge features ,[46] and gradient features,[47] is the first part of the process .The second step entails converting the collected features into counts using a regression function ,such as linear regression ,[48] piecewise linear regression ,[49] ridge regression ,[50]or Gaussian process regression ,as in .[48] Figure 1 depicts this method's whole workflow.

In ,[51] York et al .suggested a multi-feature technique for precise crowd counting .They combined many features ,such as head locations ,[52]SIFT interest points ,[53] Fourier interest points ,uneven and nonhomogeneous texture ,into one overall feature descriptor .Then, to estimate counts ,a multi-scale Markov Random Field) MRF[54] (was utilized using this global descriptor .The authors also offered a fresh dataset) UCF-CC .(50-Regression-based methods typically produce decent results ,but because they rely on a global count ,they lack spatial information.

Table 1. Comparison of Traditional Counting Methods.

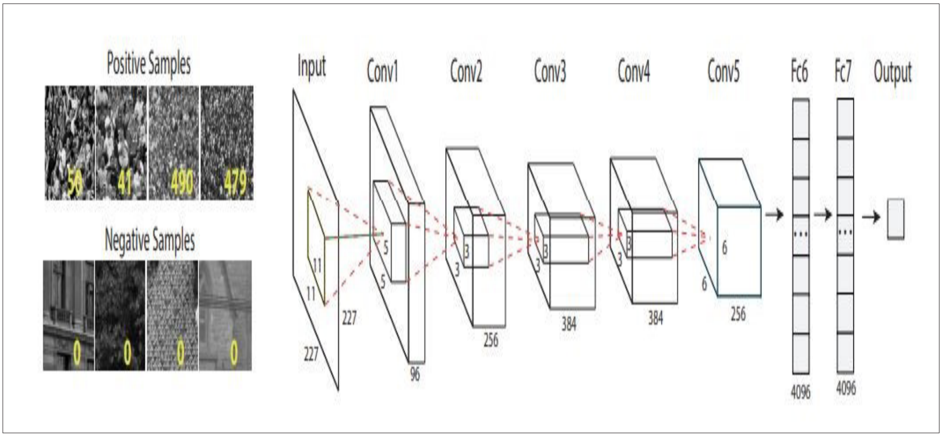
Traditional Counting Approaches	How they act	Benefits and Defects
Density Estimation-based Approaches	Convert crowd image to density map.	Utilize spatial information to minimize counting errors.
Regression-based Approaches	Feature extraction and regression modeling at low levels.	Results lack spatial information due to global count.
Detection-based Approaches	Detect heads and shoulders in crowd scenes using detectors.	Results fail in crowded and heavy occlusion scenes.

2.2. Deep Learning Approaches

Convolutional neural networks (CNNs) are like neural networks (NNs) in that they are made up of neurons and receptive fields with biases and weights that may be learned. The output of each receptive field's convolution operation is fed into a nonlinearity function when it receives a batch of inputs [55]. (e.g., ReLU or Sigmoid). CNN is presuming that the input image is an RGB image, so the hidden layers acquire rich information that improve the performance of the entire network (hidden layers and classifier). Since there are several items to be detected in the crowd scene photos, this structure has advantages in terms of speed and accuracy. End-to-end networks are those in which the network directly generates the required output after receiving the input image.

Deep network pioneering work was put forth in [56]. For counting individuals in photographs of incredibly dense crowds, an end-to-end deep convolutional neural network (CNN) regression model was put out. Using a dotting tool, a dataset compiled from Google and Flickr was annotated. There are an average of 731 persons in each of the dataset's 51 photos. In this dataset, 95 counts are the lowest and 3714 counts are the greatest. On both positive and negative classes, the network was trained. The number of the objects was labeled on the positive photographs, whereas zero was labeled on the negative images.

Five convolutional layers and two fully linked layers make up the network's structure. Figure 3 illustrates how the network was trained on object categorization with regression loss.



Figur: 3. CNN positive and negative inputs in architecture.

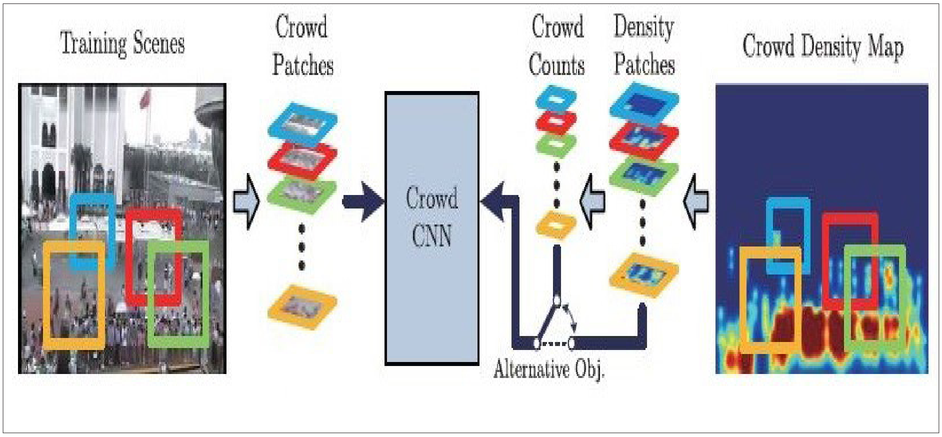
Following the first CNN-based technique ,[57] a real-time crowd density estimation method based on the multi-stage ConvNet was proposed .[58] The main presumption behind this strategy is that some CNN connections are superfluous .As a result ,similar feature maps from the second stage and their connections can be eliminated.

Network architecture :The network is made up of two multi-stage cascaded classifiers .[59] One convolutional layer and a subsampling layer make up the first stage .The second stage utilizes the same architecture .A fully connected layer with five outputs makes up the last layer ,which categorizes the crowd scenario as either very low, low ,medium ,high ,or very high .The authors optimized this stage because just 1/7 of the features come from the feature maps from the first stage .The optimization process was based on comparing how similar the maps were .To reduce processing time ,this map will be eliminated if the similarity is below a predetermined threshold.

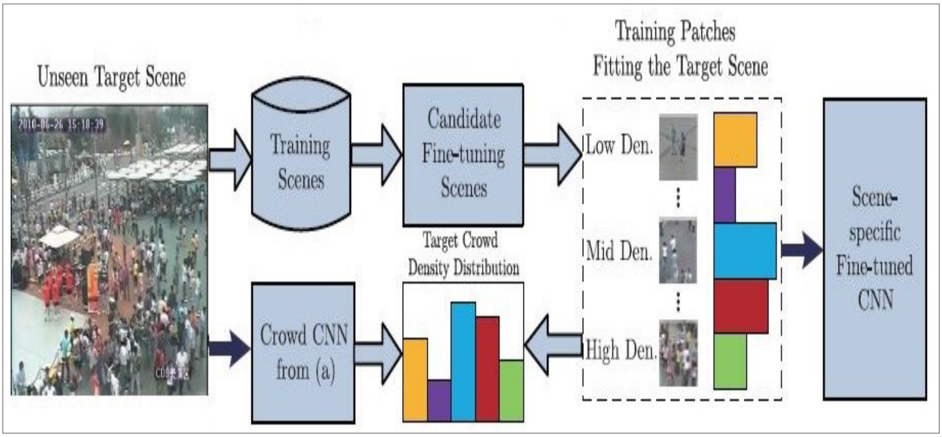
In ,[60] the author noted that performance dramatically decreased when the trained network was used with unknown data .As a result ,a new CNN mechanism was trained with switchable objectives on both crowd counts and density maps ,as illustrated in Figure .4 Another contribution to this work is the nonparametric fine-tuning module .The primary goal was to close the domain gap between the distribution of training data and the distribution of unobserved data .Candidate scene retrieval ,patch retrieval ,and local patch retrieval are all included in the nonparametric module .The primary goal of the candidate scene retrieval was to locate training scenes across all training scenes that share perspective maps with the target scene .The local patch retrieval scene seeks to identify comparable patches with densities that are like those in the test scene ,as shown in Figure

Generative adversarial network (GAN) is another framework that is used to create the crowd scene [61]. The parent patch and the child patch were provided as two inputs to the network by the author in [62]. The child patch is made up of two sub-patches, whereas the parent patch is the entire image. The goal of this architecture is to reduce the number of times the parent and child patches cross scale boundaries.

Network structure Parent Glarge and Child Gsmall are the two generators present in the framework. The input crowd image patch is mapped end-to-end by the generator network G to a density map with the same scale. To address scale fluctuation, each generator has an encoder and a decoder [63], placed back-to-back.

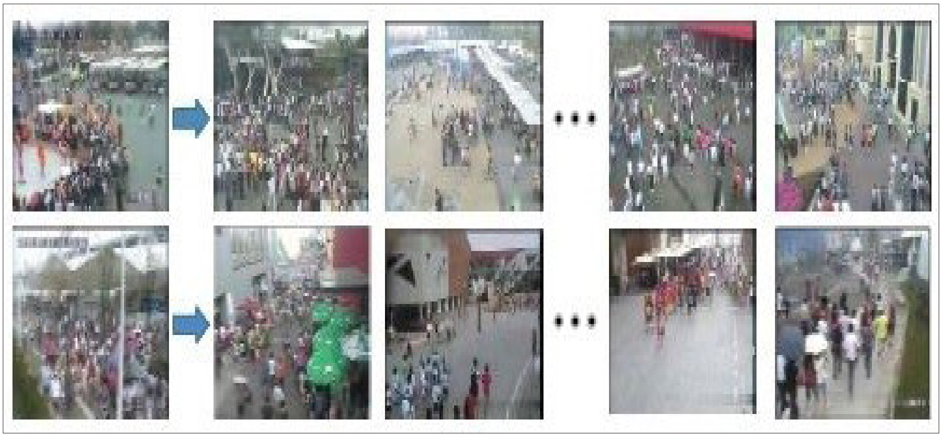


(a) training scenes for the crowd CNN

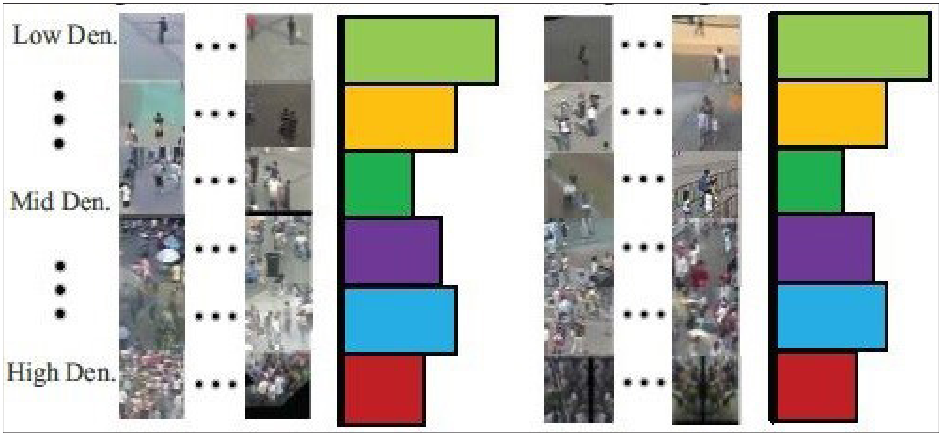

Unseen Target Scene
2010-06-26 15:10:39
Training Scenes
Candidate Fine-tuning Scenes
Training Patches Fitting the Target Scene
Low Den.
Mid Den.
High Den.
Scene-specific Fine-tuned CNN
Target Crowd Density Distribution
Crowd CNN from (a)

(b)A target scene is adjusted(fine-tuned)by the pre-trained CNN in[a]

Figure: 4. The cross-scene network's internal organization includes a fine-tuning scene module to enable generalization for unobserved data.



(a) Testing crowd scenes



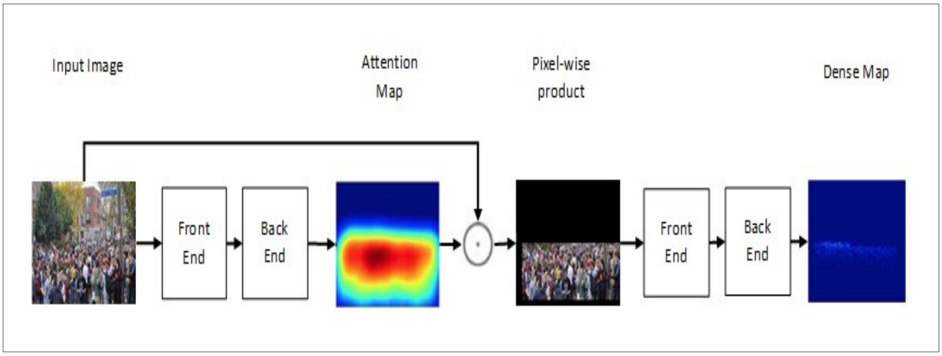
(b) Right side represents similar training patches that fit the target scene ,while the left side displays patches and distribution in the target scene.

Figure 5. The nonparametric module [60].

The authors in [64] offered two models for crowd and item counts. Counting CNN) CCNN ,(the first model ,is trained to map an image to a matching density map .The second model put forth ,called Hydra CNN ,is capable of estimating object densities in extremely cluttered environments without being aware of the scene's geometry.

On One of the most up-to-date ,cutting-edge techniques for precisely counting crowds was published in .[65] An attention-injective deformable convolutional network named AD Crowd Net was proposed by the authors ,and they claim that it can operate accurately in crowded noisy environments .The Attention Map Generator) AMG (and Density Map Estimator are the two components of the network) DME .(The input image is classified by the AMG classification network as either a crowd image or a background image .The output of AMG is then fed into DME as an input to produce a map of the crowd's density in the frame .Figure 6 details this procedure .On the Shanghai Tech dataset ,[66] UCF CC 50 dataset ,[42] WorldExpo 10'dataset ,[60] and UCSD dataset ,[39] AD Crowd Net had the highest accuracy for crowd counting .Oh et al .provided a method for quantifying uncertainty for crowd estimation in.[67]

This approach is built on a bootstrap ensemble-based scalable neural network framework .The PDA Net) Pyramid Density-Aware Attention-based network (method [68] produces a density map that shows how many people are present in each area of the input photos. The attention paradigm ,pyramid scale features ,decoder modules for crowd counting ,and a classifier to determine the crowd density in each input image are all used to create this density map .Structured feature representation learning ,and hierarchically structured loss function optimization are utilized to count the population in DSSI Net) Deep Structured Scale Integration Network .[69] (Reddy et al. addressed the issue of crowd counting in [70] using an adaptive few-shot learning method .An end-to-end trainable deep architecture was suggested in .[71] This method leverages contextual information to estimate the number of people in the input photographs by generating several receptive field sizes and learning the significance of each such characteristic at each image location.



3. Crowd Scene Datasets

Numerous datasets can be utilized to test and/or train crowd scene algorithms, as indicated in Table 3. The Shanghai Tec dataset is the most used, particularly in deep learning algorithms [66]. There are 1198 photos with annotations, including street view images and internet images. The 108 security cameras that were watching Shanghai World Expo 2010 produced the WorldExpo'10 dataset [60]. There are 1132 annotated video clips in this collection.

There are 50 annotated crowd frames in the UCF dataset _CC 50 [42]. Due to the wide variation in crowd sizes and scenario types, this dataset is one of the most difficult to analyze. The crowd size typically ranges from 94 to 4543 people. 2000 annotated photos with a dimension of 158 by 238 pixels each make up the UCSD dataset [39]. The maximum number of persons is 46, and the ground truth is labeled in the middle of each object. There are varied densities in the mall [41]. Various static and dynamic activity patterns are also present.

There are older datasets like Who do What at some Where (WWW) [85], UCLA [72], and Dyntex++ [73] that are still utilized in crowd scene counts.

Table 1. Comparison of Traditional Counting Methods.

Dataset name	Total images no.	Res.	Min	Avg.	Max	Total Count
ShanghaiTech Part A [66]	482	Varied	33	501	3139	241,677
ShanghaiTech Part B [66]	716	768 × 1024	9	123	578	88,488
UCF_CC_50 [51]	50	Varied	94	1279	4543	63,974
Mall [50]	2000	320 × 240	13	-	53	62,325
UCSD [48]	2000	158 × 238	11	25	46	49,885
WorldExpo'10 [60]	3980	576 × 720	1	50	253	199,923

4. Discussion

The examination covers the theoretical underpinnings of crowd counting, a crucial component of crowd scene analysis. A detailed analysis of density estimation-based, discovery-based, and regression-based approaches to traditional computer vision techniques is provided. Density estimation methods provide density maps that graphically represent the crowd distribution because they are adept at geographically reducing count mistakes. Regression-based techniques attempt to overcome the restrictions seen in highly crowded or dead-end settings by converting low-level traits into numbers using regression models, as opposed to discovery-based tactics, which employ detectors to label subjects' heads or shoulders for counting.

Convolutional neural networks (CNNs) are discussed in the narrative in an elegant way as a paradigm change in crowd analysis throughout the deep learning period. The authors set the stage for understanding the revolutionary potential of deep learning by describing the structure and learning mechanism of CNNs. Networks like AD Crowd Net, PDA Net, and DSSI Net are included in the debate as well as other significant advances. These models show hierarchical structures, attention processes, and generative adversarial networks (GANs), ushering in a new age of precise crowd counting even under difficult circumstances.

The study extensively examines a variety of data sets, including the Shanghai Tech data set, UCF_CC_50 data set, UCSD data set, and WorldExpo'10 data set, acknowledging the crucial role that data sets play in the advancement of research. The complexity and diversity of real-world crowd situations are highlighted by the offered dataset definition, highlighting the necessity for advanced analytic methods. The statistics highlight the diversity of crowd sizes, scene kinds, and obstacles, highlighting the necessity for adaptable research approaches.

5. Conclusions and Future directions

The significance of crowd scene analysis and its usefulness in enhancing public safety and urban planning serve as an excellent summary of the paper's investigation. An overview of the study's contributions, including a dataset scan, a proposed CD measure, and a summary of deep learning methodologies, is provided. The paper concludes by outlining potential future directions and suggesting potential directions for more research and innovation. The crowd scene analysis landscape shows potential for novel solutions that might transform event planning, public safety, and urban development as technology advances.

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. [CrossRef]

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. [CrossRef]

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. [CrossRef]

10. Chow, W.K.; Ng, C.M. Waiting time in emergency evacuation of crowded public transport terminals. *Saf. Sci.* **2008**, *46*, 844–857. [CrossRef]

11. Sime, J.D. Crowd psychology and engineering. *Saf. Sci.* **1995**, *21*, 1–14. [CrossRef]

12. Sindagi, V.A.; Patel, V.M. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognit. Lett.* **2018**, *107*, 3–16. [CrossRef]

13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

14. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]

15. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]

16. Loy, C.C.; Chen, K.; Gong, S.; Xiang, T. Crowd counting and profiling:

Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 347–382.

17. Teo, C.H.; Vishwanthan, S.; Smola, A.J.; Le, Q.V. Bundle methods for regularized risk minimization. *J. Mach.*

18. *Learn. Res.* **2010**, *11*, 311–365.

19. Goffin, J.L.; Vial, J.P. Convex nondifferentiable optimization: A survey focused on the analytic center cutting plane method. *Optim. Methods Softw.* **2002**, *17*, 805–867. [CrossRef]

20. Pham, V.Q.; Kozakaya, T.; Yamaguchi, O.; Okada, R. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 3–17 December 2015; pp. 3253–3261.

21. Liaw, A.; Wiener, M. Classification and regression by randomforest. *News* **2002**, *2*, 18–22.

22. Sirmacek, B.; Reinartz, P. Automatic crowd density and motion analysis in airborne image sequences based on a probabilistic framework. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–11 November 2011; pp. 898–905.

23. Scaillet, O. Density estimation using inverse and reciprocal inverse gaussian kernels. *Nonparametric Stat.* **2004**, *16*, 217–226. [CrossRef]

24. Cha, S.H. Comprehensive survey on distance/similarity measures between probability density functions.

25. *City* **2007**, *1*, 1.

26. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art.

27. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [CrossRef]

28. Li, M.; Zhang, Z.; Huang, K.; Tan, T. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008), Tampa, FL, USA, 8 December 2008; pp. 1–4.

29. 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.

30. 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.

31. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]

32. Wu, B.; Nevatia, R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17 October 2005; Volume 1, pp. 90–97.

33. Ali, S.; Shah, M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 22 June 2007; pp. 1–6.

34. 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.

35. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 1–27. [CrossRef]

36. Gall, J.; Yao, A.; Razavi, N.; Van Gool, L.; Lempitsky, V. Hough forests for object detection, tracking, and action recognition. *IEEE Trans. Pattern Anal.*

Mach. Intell. **2011**, *33*, 2188–2202. [CrossRef] [PubMed]

37. Viola, P.; Jones, M.J.; Snow, D. Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vis.* **2005**, *63*, 153–161 [CrossRef]

38. Zhang, T.; Jia, K.; Xu, C.; Ma, Y.; Ahuja, N. Partial occlusion handling for visual tracking via robust part matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24 June 2014; pp. 1258–1265.

39. Kilambi, P.; Ribnick, E.; Joshi, A.J.; Masoud, O.; Papanikolopoulos, N. Estimating pedestrian counts in groups. *Comput. Vis. Image Underst.* **2008**, *110*, 43–59. [CrossRef]

40. Whitt, W. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2002.

41. 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.

42. Chatelain, F.; Costard, A.; Michel, O.J. A bayesian marked point process for object detection: Application to muse hyperspectral 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.

43. 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. 33. Zhao, T.; Nevatia, R.; Wu, B. Segmentation and tracking of multiple humans in crowded environments. *Ieee Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1198–1211. [CrossRef]

44. Geyer, C.J. *Markov Chain Monte Carlo Maximum Likelihood;* Interface Foundation of North America: Fairfax Station, VA, USA, 1991.

45. Bouwmans, T.; Silva, C.; Marghes, C.; Zitouni, M.S.; Bhaskar, H.; Frelicot, C. On the role and the importance of features for background modeling and foreground detection. *Comput. Sci. Rev.* **2018**, *28*, 26–91. [CrossRef]

46. Tuceryan, M.; Jain, A.K. Texture analysis. In *Handbook of Pattern Recognition and Computer Vision*; World Scientific: Singapore 1993; pp. 235–276.

47. Mikolajczyk, K.; Zisserman, A.; Schmid, C. Shape rEcognition With Edge-Based Features. 2003. Available online: https://hal.inria.fr/inria-00548226/ (accessed on 11 September 2020)

48. Hwang, J.W.; Lee, H.S. Adaptive image interpolation based on local gradient features. *IEEE Signal Process. Lett.* **2004**, *11*, 359–362. [CrossRef]

49. Chan, A.B.; Liang, Z.S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, USA, 24 June 2008; pp. 1–7.

50. Paragios, N.; Ramesh, V. A mrf-based approach for real-time subway monitoring. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8 December 2001.

51. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature mining for localised crowd counting. In *Proceedings of the British Machine Vision Conference*; BMVA Press: Surrey, UK, 2012, Volume 1, p. 3. [CrossRef]

52. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR,

USA, 23 Jun 2013; pp. 2547–2554.

53. Vu, T.H.; Osokin, A.; Laptev, I. Context-aware cnns for person head detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7 December 2015; pp. 2893–2901.

54. Lindeberg, T. Scale Invariant Feature Transform. 2012. Available online: https://www.diva-portal.org/ smash/get/diva2:480321/FULLTEXT02 (accessed on 11 September 2020).

55. Li, S.Z. *Markov Random Field Modeling in Computer Vision*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.

56. Lempitsky, V.; Zisserman, A. Learning to count objects in images. In *Advances in Neural Information Processing Systems, Proceedings of the Neural Information Processing Systems 2010, Vancouver, BC, Canada, 6 December 2010*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2010; pp. 1324–1332.

57. Karlik, B.; Olgac, A.V. Performance analysis of various activation functions in generalized mlp architectures of neural networks. *Int. J. Artif. Intell. Expert Syst.* **2011**, *1*, 111–122.

58. 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.

59. Fu, M.; Xu, P.; Li, X.; Liu, Q.; Ye, M.; Zhu, C. Fast crowd density estimation with convolutional neural networks. *Eng. Appl. Artif. Intell.* **2015**, *43*, 81–88. [CrossRef]

60. Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; LeCun, Y. Pedestrian detection with unsupervised multi-stage feature learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, OR, USA,

23–28 June 2013; pp. 3626–3633.

61. Sun, Z.; Wang, Y.; Tan, T.; Cui, J. Improving iris recognition accuracy via cascaded classifiers. *IEEE Trans. Syst. Man Cybern. Part Appl. Rev.* **2005**, *35*, 435–441. [CrossRef]

62. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 833–841.

63. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.;

64. Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21 July 2017; pp. 4681–4690.

65. Shen, Z.; Xu, Y.; Ni, B.; Wang, M.; Hu, J.; Yang, X. Crowd counting via adversarial cross-scale consistency pursuit. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake Cite, UT, USA, 18–22 June 2018; pp. 5245–5254.

66. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]

67. Onoro-Rubio, D.; López-Sastre, R.J. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 615–629.

68. 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 Beach, CA, USA, 16–20 June 2019; pp. 3225–3234.

69. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. 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.

70. Oh, M.H.; Olsen, P.A.; Ramamurthy, K.N. Crowd counting with decomposed uncertainty. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 11799–11806.

71. 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.

72. Liu, L.; Qiu, Z.; Li, G.; Liu, S.; Ouyang, W.; Lin, L. Crowd counting with deep structured scale integration network. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 1774–1783.

73. Reddy, M.K.K.; Hossain, M.; Rochan, M.; Wang, Y. Few-shot scene adaptive crowd counting using meta-learning. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2814–2823.

74. Liu, W.; Salzmann, M.; Fua, P. Context-aware crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 16 2019; pp. 5099–5108.

75. Andersson, M.; Rydell, J.; Ahlberg, J. Estimation of crowd behavior using sensor networks and sensor fusion. In Proceedings of the 12th International Conference on Information Fusion, Seattle, WA, USA, 6–9 July 2009; pp. 396–403.

76. Beal, M.J.; Ghahramani, Z.; Rasmussen, C.E. The infinite hidden markov model. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002; pp. 577–584.

77. Siva, P.; Xiang, T. Action detection in crowd. In Proceedings of the British Machine Vision Conference (BMVC), Aberystwyth, Wales, UK, 31

August–3 September 2010; pp. 1–11.

78. Li, B.; Yu, S.; Lu, Q. An improved k-nearest neighbor algorithm for text categorization. *arXiv* **2003**,arXiv:cs/0306099.

79. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 21 November 2012; pp. 1–6.

80. Shao, J.; Loy, C.C.; Kang, K.; Wang, X. Slicing convolutional neural network for crowd video understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5620–5628.

81. Wang, J.; Zhu, X.; Gong, S.; Li, W. Attribute recognition by joint recurrent learning of context and correlation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 531–540.

82. Lazaridis, L.; Dimou, A.; Daras, P. Abnormal behavior detection in crowded scenes using density heatmaps and optical flow. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 3–7 September 2018; pp. 2060–2064.

83. Xu, B.; Qiu, G. Crowd density estimation based on rich features and random projection forest. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), New York, NY, USA, 7–9 March 2016; pp. 1–8.

84. Boominathan, L.; Kruthiventi, S.S.; Babu, R.V. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2016; pp. 640–644.