Open Challenges for Crowd Density Estimation

Shaya A. Alshaya
Computer Science department
College of Sciences and Humanities at alGhat
Majmaah University, Saudi Arabia

***Abstract*—Nowadays, many emergency systems and surveillance systems are related to the management of the crowd. The supervision of a crowded area presents a great challenge especially when the size of the crowd is unknown. This issue presents a point of start to the field of the estimation of the crowd based on density or counts. The density of a crowded area is one of the important topics dealt with in many kinds of applications like surveillance, security, biology, traffic. In this paper, we try not only to present a deep review of the different approaches/techniques used in the previous works to estimate the size of the crowd but also to describe the different datasets used. A comparison of some related works based on the weakness and the strength features of each approach is highlighted to show the important research key related to the field of the estimation of the crowded area.**

***Keywords*—Crowd density; count density; deep learning; CNN; datasets; metrics**

I. Introduction

The surveillance system is widely used in our life daily. It is mounted in bank agency [1], traffic [2], mall [3], etc. Their uses are different from one application to another. This paper is focused in the use of the surveillance systems to estimate the size of the crowd.

Estimate crowd density aims to understand the behavior of a crowded scene and to analyze the comportment for better security, management, and safety. A computer vision technique is used for all systems. Some of them propose the computation on a single image, and the most compute frames from a video streaming. Crowd analysis is associated with multiple disciplinary research topic as computer vision, public safety, biology [4], physics [5], psychology [6].

Crowd density estimation can be classified into five research topics.

1) Disaster management: Systems based on the estimation of the crowd aims to supervise the behavior to avoid disasters in several situations as music concerts, sports events, political rallies, and public demonstrations.

2) Safety monitoring: Crowd analysis help in understanding behavior, congestion, anomaly, and event [7]. These analyses are applied for many video surveillance purposes such as shopping malls, airports, and sports events.

3) Design of public spaces: Crowd analysis improves the optimization of public spaces design to ensure more safety in crowded situations.

4) Intelligence gathering and analysis: Crowd analysis is used for interesting products, interesting places. It ensures intelligence in queuing systems. Therefore, analysis improves the knowledge of the system and helps in improvement or optimization strategies.

5) Forensic search: Crowd analysis determine a particular data in a crowded scene as detecting suspicious behavior or detecting suspects [8], [7].

These topics have encouraged researchers in different specialties to contribute and to improve the estimation of a crowded area via various methods and related tasks such as density estimation [9] counting[9], [10, 11], tracking [11], behavior analysis [12]. All these tasks can be extracted from a crowded scene and there can be applied for different applications. The challenge is increased when the scene is identified as a very high dense situation. Previous studies use a variety of techniques/methods like regression [13], clustering [14], and detection [15] to count or to estimate crowds. These approaches require standards dataset to estimate the performance of crowd density analysis.

This paper is distributed as follows: different datasets used by researchers to evaluate their approaches are described in Section 2. A review of crowd density estimation approaches is presented in Section 3. A comparison between previous approaches in crowd density estimation is performed in Section 4. Finally, a conclusion based on open challenges is presented in the last section.

II. Datasets

In vision processing systems, datasets represent an essential requirement to evaluate their proposed design. This section lists the different datasets used by the previous works to assess the estimation of the crowded area approaches. Some related works perform their own dataset but most of the studies use standard and universal datasets. We focus on this section to the standard dataset used in the field of the crowded zone.

- WorldExpo'10 dataset [16]: This dataset is characterized by their size. It is composed of a big number of scene performed for the count of the crowd. It is characterized by the number of prototypes (1132 video clips), the number of scenes (108), the resolution (576*720) which are bigger than other datasets. These videos are captured by more than one hundred cameras.
- UCSD dataset [17]: It includes 2000 frames extracted from video streaming. This dataset is limited in terms of scenes because all frames are done by one camera which means one scene. Frames have little resolution (158*238).

TABLE. I. CHARACTERISTICS OF EACH DATASET

Datasets		Number of frames	Resolution	Number of scenes	Number of frames/second	Minimum number of persons	Maximum number of persons	Total number of persons	Average crowd count
WorldExpo'10		4440000	576*720	108	50	1	253	199923	50.2
UCF_CC_50		50	-	50	1	94	4543	63974	1279.5
UCSD		2000	158*238	1	10	11	46	49885	24.9
ShanghaiTech	**Part A**	482	different	-	-	33	3139	241677	501.4
	Part B	716	768*1024	-	-	9	578	88488	123.6

- UCF_CC_50 [18]: This dataset includes a limited number of frames (50) but the number of labeled pedestrians is bigger than the UCSD dataset and it achieved about 63000.
- ShanghaiTech dataset [19]: this dataset introduces a large scale crowd. It includes 330165 people as the total number of labeled pedestrians. It includes 1198 images. These images define two groups: The Part A is grouped randomly from images stored in the Internet, and the part B is composed of images captured from Shanghai streets.
- Make3D [18]: the resolution of this dataset is 2272*1704. This dataset is adopted to learn features and it estimates the scene depth from a single frame. The Make3D dataset provides more than 1000 scenes composed of outdoor and indoor scenes.

In light of this brief review, we mention that each dataset could be applied for a particular case. Table I resumes characteristics of the most used datasets in literature.

III. EVALUATION METRICS

This section discusses the different metrics used to evaluate crowded systems. In the literature, there are four factors. The most two factors used by them are:

- The Mean Absolute Error (MAE) [20] computes the average of all absolute errors which is defined by theses following formula:

Absolute Error (AE) computes the error rate between the true value (x) and the measured value (xi) related to n frames.

$$AE = |X_i - X| \quad (1)$$

The MAE is the average of all absolute errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - X| \quad (2)$$

- The Mean Squared Error (MSE) [20] represents the average of all errors related to the distance between the regression line and the value. The regression line is the best line drawn by the measured data. The accuracy is higher when the MSE is smaller.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - X)^2 \quad (3)$$

Other studies use the following metrics to evaluate crowded systems:

- Mean Windowed Relative Absolute Errors MWRAE [21] computes the average of all errors related to the distance between the real counts and the estimated counts. This metric is defined by the following formula:

$$MWRAE = \frac{1}{n}\sum_{i=1}^{n}\frac{\|C_i - \widetilde{C_i}\|}{C_i} \times 100 \quad (4)$$

Where Ci is the real count of the crowd related to the ith test video stream, $\widetilde{C_i}$ is the estimated count of the crowd related to the ith test video stream, and the parameter n represents the total number of test stream.

- Root MSE (RMSE) [17] computes the averages of all errors related to the standard deviation. This metric defines the best line around data based on the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X)^2} \quad (5)$$

Where the xi is the measured value and the x is the true value of n frames.

IV. REVIEW OF CROWD DENSITY ESTIMATION METHODS

This study is based on a deep search on Web of science database since December 2017. The most significant keywords for this search are 'Crowd density estimation' that describes the scope of this paper. During the study collection, we set only papers written in English and dealt about the density/count estimation a crowd.

During the search, we use the combination of the following words: "Crowd", Density Estimation", Crowd Count" to find papers related to the scope. Logical operators are used between keywords. Only articles on journals and conferences are approved. Steps of the selection of the paper can be resumed as follow:

First step: This step aims to filter papers according to the title and the abstract.

Second step: Eliminate duplicate papers that use the same methods in the same dataset.

Third step: In this step, we approve papers agreed with necessary criteria: articles in English, known authors, reviewed paper, developing paper, and discussed paper.

In the literature, the crowded scene is seen either by the level of density or by counting the number of people.

S. Lin et al., [22], propose an intelligent algorithm based on SVM classifier to detect heads. The frame is proceeded by a processing phase to reduce noises. Then an extraction phase is performed by Haar wavelet transformation and normalization step. Finally, the matching phase ensured by the SVM classifier is chosen. The SVM aims to classify the extracted features belong to head class or not. The estimation accuracy is between 90-95% (about 125 persons in image). The experimentation shows that the camera position has to aligned to the optimal value of the angle (72.5 degrees). This angle is defined by the camera sensor position and the plane of the crowd. The method proposed by the authors supposes a unique size of all human heads and a uniform repartition of the crowd over the horizontal plane.

JH. Yin et al., [23], performs five methods to recognize the size of the crowd area. The first method removes the background based on the subtraction of the reference image and computes the occupied surface. The error associated with this method is about 15%. The second method computes the total perimeter of the busy area by applying the Edge detection algorithm. When the number of people has increased the accuracy of the algorithm is decreased. The error is about 23%. The third method combines the two-last method to improve estimation accuracy. The crowd density estimation error is decreased to 8%. These three methods still suffer from near-far effect especially. Persons how are near the camera occupied more area than other distant persons. Then authors propose the fourth method based on Geometric distortion to compensate for the near-far effect. The fifth method attempts to detect the movement without identifying objects in video streaming. This is done by the optical flow that is defined by the difference of the brightness from one image to the next. Based on Horn's optical flow algorithm, the motion is measured. These methods did not take into consideration the constraints of real-time execution.

CS. Regazzoni et al., [24], propose an estimation approach. They use temporal information of a sequence image. A means of a distributed Kalman filter network is performed. The proposed approach attempts to synchronize between multiple sensors based on modularity and data-fusion. The distributed Extended Kalman Filtering (DEKLF) algorithm implements both static models and status history. The first one is defined by some features of the edge function as the number of vertical edges, the number of edge points, the sum of the amplitudes of the maxima detected in the shape. The second one is defined by the depletion, the enhancement, and the steady conditions of the number of people. Algorithms are chosen to increase density accuracy and real-time exigence. The experimentation discusses the results according to a comparison between the proposed DEKF and the Bayesian belief network. The error is less than 20%.

AN. Marana et al., [25], use the Minkowski fractal dimension to estimate the count of people. This method verifies the case of a railway station. The edge detection is performed to the input image. Then a binarization step followed by a dilation method (enlarge the boundaries of regions) is applied to the image. Finally, the fractal dimension classifies the image according to the density into very high, high, moderate, low, and very low rubrics. The authors evaluate their method by comparing results with Minkowski methods and the Gray Level Dependence Matrix (GLDM) [26]. The last method is not able to distinguish between area with very high density and area with high density.

SY. Cho et al., [27] choose to apply the neural network as an intelligent algorithm to find an accurate result of the crowd's size. This method is based on background removal. The neural network is applied to identify if a mask belongs to black features, white features, and edge features. The case study is implemented for Railway station. This paper proposes a novel block diagram: a fast edge detection is proposed to ensure real-time. A binary step is applied instead of the Sobel filter. Then the edge algorithm is used. Then an estimation of the undesired region is performed by a crowd object extraction. This is done by removing the background. Finally, a Hybrid Global Learning (HGL) associated to a neural model is implemented. The HGL is performed by three algorithms. The first algorithm performs a hybrid of least squares and random search. Results prove that is the fast one with 2.02 min (CPU running time for learning) but the lowest estimation accuracy (90.72%). The second algorithm performs a hybrid of least squares and Simulated Annealing (SA). It obtains the best estimation accuracy (94.36%) but the worst speed (197.5 min). The third algorithm performs a hybrid of least squares and genetic algorithm (GA). It obtains 75.3 min in terms of time learning and 93.89 % for estimation accuracy.

C. Wang et al., [28] apply an end-to-end deep CNN regression to approximate people's number in a condensed crowd. The authors focus to decrease the influence of the ground by including negative samples to the training data. The truth counting of these samples is defined as zero. The proposed method enhances the estimation of counting persons. Results highlight a decrease of the error between absolute difference and the normalized absolute difference by 16.7% and 27% to mean respectively. The error rate is about 10 % which is still important. Nevertheless, this method is limited to 1300 persons per image.

F. Min, [29] presents an optimized method of the CNN named ConvNet to enhance the accuracy and the speed of the estimated crowd density. The author implements two stages on the cascade of CNN and he proposes to remove some network connections of the CNN design to speed up the computation. Experimentation is based on the PETS_2009 dataset. This method is limited to an image size of 42x40. Results show a decrease in the error rate to 3.2 %. Unfortunately, the author does not discuss the acceleration achieved by his method.

C. Zhang et al., [16] approximate the crowd's density/count especially for unseen scene by applying a deep CNN. The authors describe a data-driven method to finetune the trained CNN model for the target scene. They, also, built a novel dataset constituted with 108 frames which supports about 200,000 persons. A comparative study based on other datasets is done to show the reliability and the effectiveness of their method.

E. Wolf et al., [18] attempt to count persons by employing CNN. The addressed method focuses on layered boosting and selective samples. It aims to enhance accuracy and speed up the processing time. The authors achieve their goal by reducing

the mean absolute error from 20% to 35% and the training time is decreased by 50%.

Z. Zhao et al., [21] present a CNN- based method to compute the number of persons across a line-of-interest. The method uses pairs of videos as inputs and it performs the training with pixel-level supervision maps. The proposed enhancement let the CNN learn more about features by decomposing the training phase into two steps: (1) Estimate the crowd density map, and (2) Estimate crowd velocity map. This decomposition provides more accuracy to solve the original problem by starting to answer each step. The authors perform a new dataset based on pedestrian trajectory annotations to show the robustness of the method via introducing a novel metrics: Mean Windowed Absolute error (6%).

Y. Zhang et al., [19] try to estimate the crowd from an unique image by performing a Multi-column CNN architecture. The MCNN supports any size or resolution of the input image. The method uses filters with different sizes to let CNN learn the features of each column. Then a geometry-adaptive kernel is used to compute the true density map associated with the input image. A new dataset including 1198 images is introduced by authors to cover all the challenging situations. Experimentation shows that the mean absolute error is 1.07%.

C. Shang et al., [30] attempt to count the crowd directly from an input image using an end-to-end CNN. The method estimates the crowd based on both global and local features by applying a pre-trained CNN to the image. The recurrent network layers provide the local counting by mapping features. The local count reduces the training time, and the global count enhances the accuracy of the results obtained by the local regions. A comparative study based on many databases is discussed to demonstrate the effectiveness of their attempt.

T. Mundhenk et al., [31] apply the deep learning method to count the crowd related to cars. The authors perform a large contextual dataset to help drivers to choose the best target and avoid bottlenecks. The proposed method aggregates residual learning and inception-style layers. This solution represents a new way to counts objects instead of the base of the known method on density estimation and localization. The authors prove via their experimentation that results are more accurate and the processing time is faster.

L. Boominathan et al., [32] announce the "crowdnet" framework based on the deep CNN to count the density of the crowd. Crowdnet is performed by the combination of the deep and shallow applied to a static image. This aggregation provides effective results associated with semantic information and features. To improve accuracy, the authors propose to enlarge the trained dataset to exceed 100 samples. Results are discussed using UCF CC 50 dataset.

A. Vishwanath et al., [33] count the crowd by using both the end-to-end cascaded CNN and the density map estimation. The proposed idea by authors provides the estimation of the crowd by classifying count into groups. This method enables us to learn globally features that refine highly the density maps and decreases the error count. A comparative study is highlighted to prove the accuracy of the density maps with the minimum count error.

S. Deepak et al., [34] present a mapping method between crowd counting and their density. A multi-scale CNN is described to decrease the worst effects of some factors as inter-occlusion, the high similarity of appearance, and view-points. The method is based on the switching of the CNN according to independent regressors to enhance the accuracy and the estimation. The proposed switch between classifiers to select the best CNN regressor. Results show that the switch relays patch to, particularly column in CNN to identify the crowd density of the input image. The comparative study proves that the proposed method enhances the accuracy and the mean error is decreased to near 2%.

X. Yang et al., [35] present an emergency evacuation as a case study related to the crowd area. The authors perform a clustering algorithm to extract informed and uninformed walkers. The goal of their study is to find the optimized guide during evacuation. The density of the crowd constitutes important criteria to achieve their goal. The informed method with an exponent model attains an approved accuracy.

Z. Zhikang et al., [9] propose to count the crowd based on many structures. The authors announced their method named the Adaptative Capacity Multi-scale CNN. This method ensures the assignment of different capacities to different portions. This method focuses on important regions instead of the whole image to ensure optimized allocations. The proposed method is composed of a fine network, a coarse network, and a smooth network. The first one finds the region to be focused and produce the rough feature map. The second one extracts the region of interest into a fine feature map. The third one enhances results by aggregate the two studied features to decrease the effect of division. The proposed method is well validated according to five used datasets.

Z.Liping et al., [36] introduce a deep learning technique to compute the crowd's density in the case of non-uniform density and variations. The authors apply pooling operation to the density map to overcome the loss of the local spatial information. This pooling is performed by the use of dilated CNN to support details related to person position. This last feature is provided by global context guidance. The proposed method is proved by the use of many datasets.

X. Zeng et al., [37] attempt to decrease the problem of the scale variation related to the crowd's estimation. The authors propose to provide more accurate contextual information by using a deep scale purifier network. The described method encodes multiscale features. The proposed supports a frontend and a backend model. A cross scene evaluation is applied to the approach. Many datasets are used to evaluate the accuracy of the DSPNet method.

This brief review proves that most techniques are applied only for an image. The authors in [18], [21], [33], and [33] propose an attempt to treat video instead the image to estimate the density of the crowd. These attempts should be enhanced to support any inputs. Recent works adopt deep learning methods to compute the density. These attempts request a learning phase based on a dataset. The high accuracy is the strong point

of these methods but they suffer always from the increased time of processing. Datasets aim to evaluate the performance of methods proposed by researchers to estimate the crowd's size. When the evaluation is made by different datasets, results are more acceptable.

The real-time constraints are not well studied by the cited related works. The authors in [18], [21] attempt to propose methods with respect to the real-time constraints. This field requests to propose many hardware architectures to be implemented into a camera to estimate the density of the crowd.

This section has discussed some important studies related to the estimation of the density or the counts of crowds. The presented review lists many techniques based on video or image processing. Some methods extract the density according to the spatial information of the frame. These methods are accurate only in the case of the small size of crowds (inferior to 50). Other methods based on deep learning techniques show more accurate results especially in the case of the biggest size of crowds.

V. SYNTHESIS

This section discusses the most important studies to extract the benefits and limits of each work. Then, a comparison based on different results metrics is highlighted to show the accuracy. At the end of this section, the evolution of this field is presented according to the number of publications during the last five years.

Previous works should be presented according to their characteristics, strengths, and weaknesses. Table II describes many studies based on the nature of the input, application type, used approach, used dataset, benefits, limits, and real-time processing.

TABLE. II. OVERVIEW OF THE PRINCIPAL METHODS TO ESTIMATE/COUNT CROWD AREA

Authors	Inputs	Application	Approach	Dataset	Benefits	Limits	Real-Time
[28]	Image/different position	Pedestrians	End-to-end deep model CNN	UCFCC	Reduce the mean and the deviation of AD and NAD	Only 1300 per/image	NA
[29]	Image 42x40	Pedestrians	Optimized CNN (2 layers)	PETS-2009	Speed up processing and increase the correct rate	Complicate algorithm and limited dataset	NA
[16]	Image 158x238	Pedestrians	CNN	New dataset WorldExpo'10 UCSD UCF_CC_50	Applied for unseen scene	NA	NA
[18]	Image/Video 158x238	Bacterial cells, microscopy images	Gradient boosting with CNN (2 layers)	EXPO UCSD UCF50	Increasing accuracy	NA	Reduction by 50% in training time
[21]	Video 1280x720	Alley, street square	CNN with NVidia Titan GPU	New dataset	Decrease the time processing	A non-standard dataset is used	Yes, T=0.1s
[19]	Image/arbitrary camera	Pedestrians	Multi-column CNN	New dataset UCSD UCF_CC_50 WorldExpo'10	Improve estimation density	NA	NA
[30]	Image 640x480	Pedestrians	CNN based on local and global mapping	New dataset UCF_CC_50 WorldExpo'10	Decrease the time processing	NA	NA
[31]	Image 256x256	Cars	CNN with residual learning	New dataset	Increase accuracy	Time processing increased	
[32]	Image 225x225	Pedestrians	Deep learning and shallow	UCF_CC_50	Increase Accuracy	Insufficient number of training image	Yes
[33]	Image/Video	Pedestrians	End-to-end cascade CNN	ShanghaiTech UCF_Crowd_50	Increase Accuracy	NA	NA
[34]	Image/Video	Pedestrians	Switch-CNN	WorldExpo'10 UCSD ShanghaiTech UCF_CC_50	Increase Accuracy	NA	NA
[9]	Image	Pedestrians	ACM-CNN	WorldExpo'10 UCSD ShanghaiTech UCF_CC_50	Focus on important regions only	NA	NA
[36]	Image	Pedestrians	Dilated Convolution with Global Self-Attention	ShanghaiTech UCF_CC_50 UCSD	Decrease the loss of the low-level features	NA	NA
[37]	Image	Pedestrians	Deep scale purifier network	ShanghaiTech UCF_CC_50 UCF-QNRF	Reduce the loss of the contextual information	NA	NA

The following Tables III-VII compare the results of the related works according to the used datasets and the used metrics.

TABLE. III. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO WORLDEXPO'10 CROWD COUNTING DATASET

Methods	Average MAE	Average MSE
Z. Zhikang et al., [9]	8.56	-
LBP+ RR [16]	31	17.4
Crowd CNN [16]	14.1	15.5
Fine-tuned Crowd CNN [16]	12.9	14.9
Crowd CNN+RR [16]	10.7	14.3
Sam et al., [34]	9.4	-
Regression learning [38]	26.7	14.3
Ridge Regression [39]	16.5	14.1
Shang et al., [40]	11.7	-
Y. Zhang et al., [41]	11.6	-

TABLE. IV. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO UCSD DATASET

Methods	MAE	MSE	MWRAE
Z. Zhikang et al., [9]	**1.01**	**1.29**	-
C. Zhang et al., [16]	1.6	3.31	-
Walach et al., [17]	1.53	-	-
Zhuoyi et al., [21]	0.9	-	0.54
Sam et al., [34]	1.62	2.1	-
Z.Liping et al., [36]	1.08	1.44	-
Ridge Regression [39]	2.25	7.82	-
Y. Zhang et al., [41]	1.07	1.35	-
An et al., [42]	2.16	7.45	-
A. B. Chan et al., [43]	2.24	7.97	-
K. Chen et al., [44]	2.07	6.86	-
Lempitsky et al., [45]	1.7	-	-
Zhang et al., [46]	1.6	-	-
Pham et al., [47]	1.6	-	-
Ma et al., [48]	0.64	-	0.61

TABLE. V. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO UCF_CC_50 DATASET

Methods	Mean Absolute Error	Mean Squared Error
Z. Zhikang et al., [9]	291.6	337
C. Zhang et al., [16]	467.0	498.5
Crowd CNN+RR [16]	467	498.5
Walach et al., [17]	474.0	-
Sam et al., [34]	318.1	439.2
Z.Liping et al., [36]	257.0	343.9
X. Zeng et al., [37]	243.3	307.6
Shang et al., [40]	270.3	-
Y. Zhang et al., [41]	377.6	509.1
Rodriguez et al., [49]	655.7	697.8
Learning to count [45]	493.4	487.1
Lempitsky et al., [45]	493.4	487.1
Zhang et al., [46]	467.0	-
MS counting [50]	468	590.3
Idrees et al., [50]	419.5	541.6
Boominathan et al., [51]	452.5	-
Sindagi et al., [52]	322.8	397.9
Onoro-Rubio et al., [53]	465.7	371.8

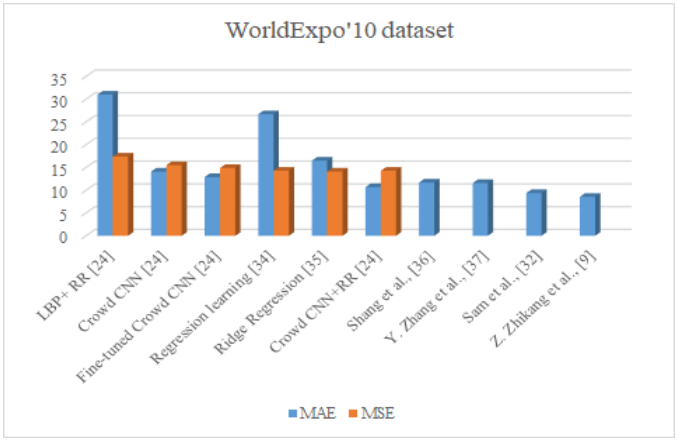
TABLE. VI. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO MAKE3D DATASET

Methods	Root Mean Squared
Walach et al., [17]	13.89
Saxena et al., [54]	16.7
Li et al., [55]	15.2
Karch et al., [56]	15.1
F. Liu et al., [57]	12.89
M. Liu et al., [58]	12.6

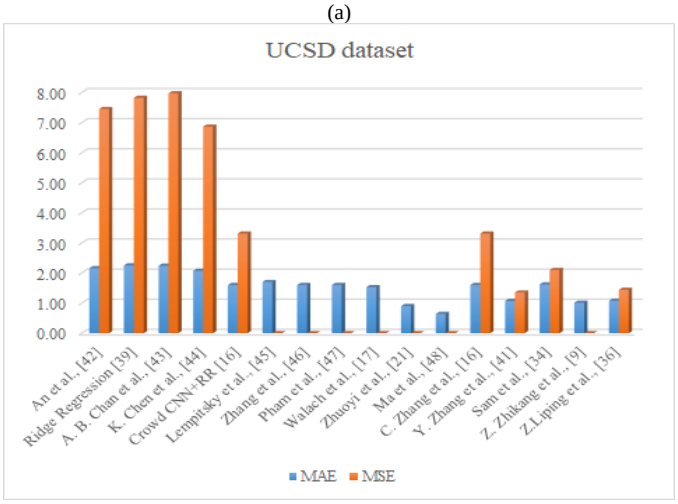
TABLE. VII. COMPARATIVE STUDIES BASED ON FOUND RESULTS RELATED TO SHANGHAITECH (PART A) DATASET

Methods	MAE	MSE
Z. Zhikang et al., [9]	17.5	103.5
LBP+ RR [16]	59.1	303.2
C. Zhang et al., [16]	49.8	336.1
Sam et al., [34]	21.6	135.0
Z.Liping et al., [36]	65.6	107.2
X. Zeng et al., [37]	68.2	107.8
Y. Zhang et al., [41]	26.4	173.2
Sindagi et al., [52]	20.0	152.4

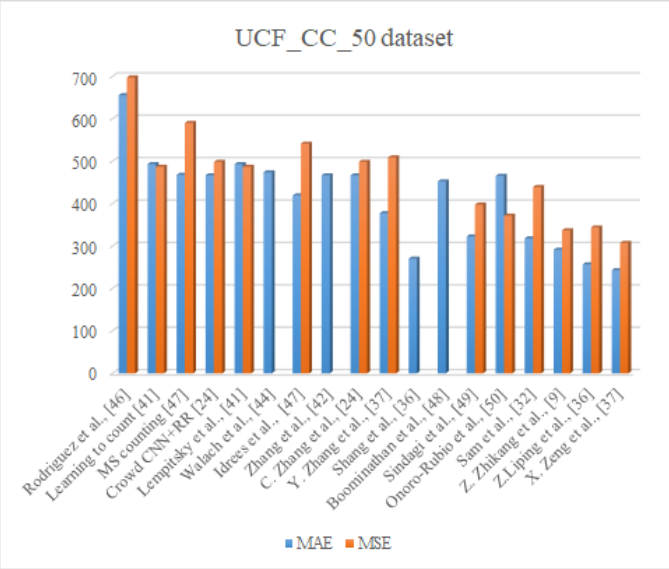
Fig. 1 highlights the difference between related works according to the MAE and MSE.



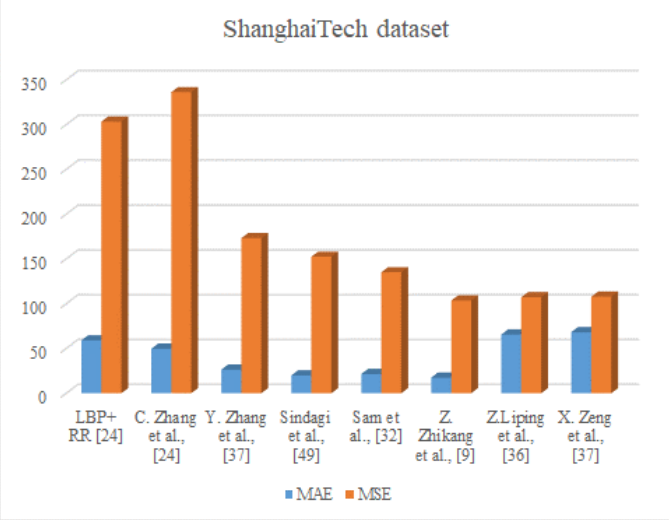
(a)



(b)



(c)



(d)

Fig. 1. A Comparison between Related Works According to different Datasets.

VI. Conclusions

The estimation of the density of the crowded area is still a challenge for researchers. In this paper, we have presented the different research axes related to the crowded zone especially in terms of the datasets, metrics, and approaches.

Datasets should be improved by increasing the number of scenes and achieve over than half-million pedestrians. These enhancements are required to improve results found in the evaluation phase.

Approaches still request improvement by decreasing the error between the real number of crowd and the estimated counts. Other methods would be applied to support the movement of the camera and the fusion of data received from many sources. In addition, the Real-Time constraint has to be the future of research work related to this domain.

The estimation of the density of crowd area is employed in a different type of applications as the estimation of pedestrians, crowded car traffic, the crowd in malls, and bacterial cell microscopy.

Besides, this domain continues to be an open challenge according to the number of the article published in the Web of Science database. Publishing is boosting from 85 articles in 2015 to achieve 148 articles published in 2019. The IEEE Xplore database shows that the number of published papers between 2018 and 2019 is increased by 25%. In the Science direct database, about 20% of the evolution of published papers are highlighted since 2018.

These statistics prove not only the importance of the domain but also the continuity of the challenge facing the crowd's estimation.

According to this deep study, the crowd size's estimation still requests enhancement in accuracy and real-time constraints.

References

[1] M. Ben Ayed, S. Elkosantini, and M. Abid, "An Automated Surveillance System Based on Multi-Processor and GPU Architecture," (in English), Engineering Technology & Applied Science Research, Article vol. 7, no. 6, pp. 2319-2323, Dec 2017.

[2] K. A. Eldrandaly, M. Abdel-Basset, and L. Abdel-Fatah, "PTZ-Surveillance coverage based on artificial intelligence for smart cities," (in English), International Journal of Information Management, Article vol. 49, pp. 520-532, Dec 2019.

[3] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazan, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," (in English), Expert Systems with Applications, Article vol. 42, no. 21, pp. 7991-8005, Nov 2015.

[4] J. J. Feng and Y. He, "Collective motion of bacteria and their dynamic assembly behavior," (in English), Science China-Materials, Article; Proceedings Paper vol. 60, no. 11, pp. 1079-1092, Nov 2017.

[5] J. C. Chen et al., "A Cascaded Convolutional Neural Network or Age Estimation of Unconstrained Faces," in 2016 Ieee 8th International Conference on Biometrics Theory, Applications and Systems(International Conference on Biometrics Theory Applications and Systems, New York: Ieee, 2016.

[6] A. Sieben, J. Schumann, and A. Seyfried, "Collective phenomena in crowds-Where pedestrian dynamics need social psychology," (in English), Plos One, Article vol. 12, no. 6, p. 19, Jun 2017, Art. no. e0177328.

[7] M. B. Ayed, S. Elkosantini, S. A. Alshaya, and M. Abid, "Suspicious Behavior Recognition Based on Face Features," IEEE Access, vol. 7, pp. 149952-149958, 2019.

[8] M. Ben Ayed, SabeurElkosantini, and M. Abid, "An Automated Surveillance System based on Multi-Processor System-on-Chip and Hardware Accelerator," (in English), International Journal of Advanced Computer Science and Applications, Article vol. 8, no. 9, pp. 59-66, Sep 2017.

[9] Z. K. Zou, Y. Cheng, X. Y. Qu, S. L. Ji, X. X. Guo, and P. Zhou, "Attend to count: Crowd counting with adaptive capacity multi-scale CNNs," (in English), Neurocomputing, Article vol. 367, pp. 75-83, Nov 2019.

[10] A. Mahmood and S. Al-Maadeed, "Action recognition in poor-quality spectator crowd videos using head distribution-based person segmentation," (in English), Machine Vision and Applications, Article vol. 30, no. 6, pp. 1083-1096, Sep 2019.

[11] H. H. Chen, B. Guo, Z. W. Yu, and Q. Han, "CrowdTracking: Real-Time Vehicle Tracking Through Mobile Crowdsensing," (in English), Ieee Internet of Things Journal, Article vol. 6, no. 5, pp. 7570-7583, Oct 2019.

[12] Y. Ma, E. W. Lee, Z. A. Hu, M. Shi, and R. K. Yuen, "An Intelligence-Based Approach for Prediction of Microscopic Pedestrian Walking

Behavior," (in English), Ieee Transactions on Intelligent Transportation Systems, Article vol. 20, no. 10, pp. 3964-3980, Oct 2019.

[13] X. L. Wei, J. P. Du, M. Y. Liang, and L. F. Ye, "Boosting deep attribute learning via support vector regression for fast moving crowd counting," (in English), Pattern Recognition Letters, Article vol. 119, pp. 12-23, Mar 2019.

[14] A. S. Rao, J. Gubbi, S. Marusic, and M. Palaniswami, "Estimation of crowd density by clustering motion cues," (in English), Visual Computer, Article vol. 31, no. 11, pp. 1533-1552, Nov 2015.

[15] K. Aziz, D. Merad, R. Iguernaissi, P. Drap, and B. Fertil, "Head detection based on skeleton graph method for counting people in crowded environments," (in English), Journal of Electronic Imaging, Article vol. 25, no. 1, p. 14, Jan 2016, Art. no. 013012.

[16] C. Zhang, H. S. Li, X. G. Wang, X. K. Yang, and Ieee, "Cross-scene Crowd Counting via Deep Convolutional Neural Networks," in 2015 Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2015, pp. 833-841.

[17] E. Walach and L. Wolf, "Learning to count with cnn boosting," in European Conference on Computer Vision, 2016, pp. 660-676: Springer.

[18] E. Walach and L. Wolf, "Learning to Count with CNN Boosting," in Computer Vision - Eccv 2016, Pt Ii, vol. 9906, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. (Lecture Notes in Computer Science, Cham: Springer International Publishing Ag, 2016, pp. 660-676.

[19] Y. Y. Zhang, D. S. Zhou, S. Q. Chen, S. H. Gao, Y. Ma, and Ieee, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in 2016 Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2016, pp. 589-597.

[20] S. Jachner, G. Van den Boogaart, and T. Petzoldt, "Statistical methods for the qualitative assessment of dynamic models with time delay (R Package qualV)," Journal of Statistical Software, vol. 22, no. 8, pp. 1-30, 2007.

[21] Z. Y. Zhao, H. S. Li, R. Zhao, and X. G. Wang, "Crossing-Line Crowd Counting with Two-Phase Deep Neural Networks," in Computer Vision - Eccv 2016, Pt Viii, vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. (Lecture Notes in Computer Science, Cham: Springer International Publishing Ag, 2016, pp. 712-726.

[22] S. F. Lin, J. Y. Chen, and H. X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," (in English), Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans, Article vol. 31, no. 6, pp. 645-654, Nov 2001.

[23] A. C. Davies, J. H. Yin, and S. A. Velastin, "CROWD MONITORING USING IMAGE-PROCESSING," (in English), Electronics & Communication Engineering Journal, Article vol. 7, no. 1, pp. 37-47, Feb 1995.

[24] C. S. Regazzoni and A. Tesei, "Distributed data fusion for real-time crowding estimation," (in English), Signal Processing, Article vol. 53, no. 1, pp. 47-63, Aug 1996.

[25] A. N. Marana, L. D. Costa, R. A. Lotufo, S. A. Velastin, Ieee, and Ieee, "Estimating crowd density with Minkowski fractal dimension," in Icassp '99: 1999 Ieee International Conference on Acoustics, Speech, and Signal Processing, Proceedings Vols I-Vi(International Conference on Acoustics Speech and Signal Processing (ICASSP), New York: Ieee, 1999, pp. 3521-3524.

[26] R. M. Haralick, "STATISTICAL AND STRUCTURAL APPROACHES TO TEXTURE," (in English), Proceedings of the Ieee, Review vol. 67, no. 5, pp. 786-804, 1979.

[27] S. Y. Cho, T. W. S. Chow, and C. T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," (in English), Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, Article vol. 29, no. 4, pp. 535-541, Aug 1999.

[28] C. Wang, H. Zhang, L. Yang, S. Liu, X. C. Cao, and Acm, Deep People Counting in Extremely Dense Crowds (Mm'15: Proceedings of the 2015 Acm Multimedia Conference). New York: Assoc Computing Machinery, 2015, pp. 1299-1302.

[29] M. Fu, P. Xu, X. D. Li, Q. H. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," (in English), Engineering Applications of Artificial Intelligence, Article vol. 43, pp. 81-88, Aug 2015.

[30] C. Shang, H. Z. Ai, B. Bai, and Ieee, "END-TO-END CROWD COUNTING VIA JOINT LEARNING LOCAL AND GLOBAL COUNT," in 2016 Ieee International Conference on Image Processing(IEEE International Conference on Image Processing ICIP, New York: Ieee, 2016, pp. 1215-1219.

[31] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning," in Computer Vision - Eccv 2016, Pt Iii, vol. 9907, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. (Lecture Notes in Computer Science, Cham: Springer Int Publishing Ag, 2016, pp. 785-800.

[32] L. Boominathan, S. S. S. Kruthiventi, R. V. Babu, and Acm, CrowdNet: A Deep Convolutional Network for Dense Crowd Counting (Mm'16: Proceedings of the 2016 Acm Multimedia Conference). New York: Assoc Computing Machinery, 2016, pp. 640-644.

[33] V. A. Sindagi, V. M. Patel, and Ieee, CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting (2017 14th Ieee International Conference on Advanced Video and Signal Based Surveillance). New York: Ieee, 2017.

[34] D. B. Sam, S. Surya, R. V. Babu, and Ieee, "Switching Convolutional Neural Network for Crowd Counting," in 30th Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2017, pp. 4031-4039.

[35] X. X. Yang, X. L. Yang, Q. L. Wang, Y. L. Kang, and F. Q. Pan, "Guide optimization in pedestrian emergency evacuation," (in English), Applied Mathematics and Computation, Article vol. 365, p. 12, Jan 2020, Art. no. Unsp 124711.

[36] L. Zhu, C. Li, B. Wang, K. Yuan, and Z. Yang, "DCGSA: A global self-attention network with dilated convolution for crowd density map generating," Neurocomputing, 2019.

[37] X. Zeng, Y. Wu, S. Hu, R. Wang, and Y. Ye, "DSPNet: Deep scale purifier network for dense crowd counting," Expert Systems with Applications, vol. 141, p. 112977, 2020.

[38] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 2685-2688: IEEE.

[39] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in BMVC, 2012, vol. 1, no. 2, p. 3.

[40] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1215-1219: IEEE.

[41] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "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.

[42] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-7: IEEE.

[43] A. B. Chan, Z.-S. J. 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, 2008, pp. 1-7: IEEE.

[44] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2467-2474.

[45] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in Advances in neural information processing systems, 2010, pp. 1324-1332.

[46] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 833-841.

[47] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "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, 2015, pp. 3253-3261.

[48] Z. Ma and A. B. Chan, "Crossing the line: Crowd counting by integer programming with local features," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2539-2546.

[49] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in 2011 International Conference on Computer Vision, 2011, pp. 2423-2430: IEEE.

[50] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2547-2554.

[51] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 640-644: ACM.

[52] 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), 2017, pp. 1-6: IEEE.

[53] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in European Conference on Computer Vision, 2016, pp. 615-629: Springer.

[54] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in Advances in neural information processing systems, 2006, pp. 1161-1168.

[55] C. Li, A. Kowdle, A. Saxena, and T. Chen, "Towards holistic scene understanding: Feedback enabled cascaded classification models," in Advances in Neural Information Processing Systems, 2010, pp. 1351-1359.

[56] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 11, pp. 2144-2158, 2014.

[57] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 2024-2039, 2015.

[58] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 716-723.