Defence Science Journal, Vol. 75, No. 2, March 2025, pp. 167-178, DOI : 10.14429/dsj.19142

Crowd Counting in Still Images for RoboSoldier: A Survey

Tendai Padenga[#,*] and Vidhi Khanduja[$]

[#]*Department of Computer Science, Gujarat Technological University, Gujarat – 382 424, India*
[$]*Department of Computer Science, Delhi University, Delhi – 110 007, India*
[*]*E-mail: tepadenga@gmail.com*

ABSTRACT

Part of the ultimate key capabilities for a RoboSoldier will lie in its ability to count the total number of people within a crowd with pin-drop precision. Crowd Counting (CC) relates to the estimation of the existing number of objects within a still image or video frame. Crowd counting has been successfully used in a wide range of applications from gathering business intelligence, such as consumer shopping patterns and ensuring normal operating conditions, to urban planning for crowd safety and stability through traffic monitoring. However, crowd counting has a fair range of its challenges like, high cluttering, varying illumination, severe occlusion, Perspective Distortion (PD), Irregular Object Distribution (IOD), non-object scales, and Location Information Loss (LIL). The emergence of Fast Learning Algorithms (FLA) for Deep Belief Networks (DBN) and Deep Learning (DL) techniques triggered groundbreaking research interests due to their inherent capabilities of overcoming shortfalls of traditional algorithms which were based on hand-designed feature extraction techniques. However, crowd counting is now sliced into two clear and distinct techniques namely traditional and Convolutional Neural Network (CNN). Traditional techniques can be broadly classified into those that count by regression and those that count by using density estimation techniques. The more recent and advanced is CNN whose architecture has data training capabilities drilling down to the base of as many network layers as available. Categorisation of these layers has been used to extensively offer researchers sound edifice in designing new and powerful distributed control algorithms as well as monitoring algorithms applicable to a wide range of crowd counting applications concerning public gatherings, military defense, wildlife census, etc.

Keywords: RoboSoldier; Location information loss; Deep Learning; Fast learning Algorithms

NOMENCLATURE

ACM	: Association of computing machinery
ACSCP	: Adversarial cross-scale consistency pursuit network
CC	: Crowd counting
CNN	: Convolutional neural network
DBN	: Deep belief network
DE	: Density estimation
DL	: Deep learning
DM	: Density map
DSRM	: Deep spatial regression model
DWT	: Discrete wavelet transform
EDM	: Electron density map
FCN	: Fully convolutional network
FCNN	: Fully convolutional neural network
FLA	: Fast learning algorithms
GAN	: Generic adversarial networks
GLCM	: Gray level co-occurrence metrics
GTD	: Gradient temporal difference
HOG	: Histogram of oriented gradients
IEEE	: Institute of electrical and electronics engineers
IOD	: Irregular object distribution
LBP	: Local binary pattern
LIL	: Location information loss
LSTM	: Long short-term memory
MCNN	: Multicolumn convolutional neural network
MMCNN	: Multicolumn multitask convolutional neural network
NN	: Neural networks
PD	: Perspective distortion
RESNET	: Residual network
RJMCMC carlo	:Reversible-jump markov chain monte
SCNN	: Skip-connection convolutional neural network
SVM	: Support vector machine
VGG	: Visual geometry group

Received : 18 May 2023, Revised : 15 April 2024
Accepted : 31 December 2024, Online published : 24 March 2025

1. INTRODUCTION

Crowds, in general, are synonymous with stampedes, which in turn generates a charged environment posing the following spectrum of threats e.g. fighting, rioting, violent protests, mass panic, and excitement[1]. However, considering the crowd size and applying corresponding effective measures, tools, and techniques a lot of these tragedies may be avoided.

Crowd counting has become an area of major concern due to its critical role in town planning, anomaly detection, defense, and healthcare amongst a host of others.

Crowd counting aims to achieve a near estimate object count within a prescribed still image. These objects may be people, cars, cells, and drones depending on the crowd in question. Its main objective is to lead to crowd size estimations which can be used to determine the respective distribution pattern of people throughout a public space[3].

1.1 Crowd-Counting Techniques Challenges

- High cluttering constitutes of objects in nonuniform arrangement and very close to each other.
- Varying illumination varying light illumination compromises on levels of accuracy.
- Severe occlusion occurs when a group of objects become very close leading to eventual merging, where it becomes difficult to identify individual objects[5,6].
- Perspective Distortion is a result of varying differing camera angles either tilting or up-down movements[7].
- Irregular object distribution results from a varying density distribution pattern within the same image space[8].
- Nonuniform object scales is also another case of differing perspectives, in essence, those objects that are closer appear larger in direct comparison to those farther apart.
- Location Information Loss and the total number of available people within an image share a complementary relationship, thereby making it a very important research area in crowd counting. Unfortunately, comprehensive research concentration has been paid on the network structure designs at the expense of the fundamental problem caused by Density Maps and that is "Location Information Loss"[9].

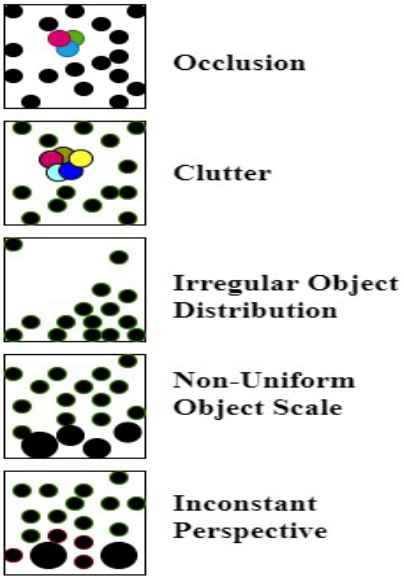

Figure 1. Unique challenges in crowd counting[5].

The generation of a density map for an exact head location or position (x_i,y_i) within a still image, through a normalised standard Gaussian kernel function $G_{\sigma_i}x,y$ as generated within the neighbourhood of its corresponding, can take the following alternative expression: where α becomes the normalization factor such that:

$$\sum G_{\sigma_i}x,y=1 \tag{1}$$

σ_i becomes the emergent variance of the Gaussian kernel of the i^{th} head position, and taking into design consideration the traditional Density Mapping models then becomes a constant. To then convert the corresponding marked points into a well-defined density function, it then translates the standard normalized Gaussian kernel function G_{σ_i} summing up different positions. The resulting density function $F(x,y)$ can now find its expression as follows:

$$M\left(x_i,y_i\right)=\max\left\{G_{\sigma_i}\left(x,y\right),M\left(x_{i-1},y_{i-1}\right)\right\} \tag{2}$$

$$F\left(x,y\right)=M\left(x_N,y_N\right) \tag{3}$$

where, $M(x_i,y_i)$ becomes a representation of the density function housing the i^{th} head positions and where N now becomes the estimate representation of people within the i^{th} image. However, each head sampling position is achieved using 3D scenarios.

Perspective distortion then results in generation of varying head sizes[9].

The perspective distortion problem can completely be solved using Density Maps that implements the adaptive geometric Gaussian kernel that in turn draws from previous findings; by making use of the available average distance between the k nearest heads and the corresponding i^{th} head position. Since $G_{\sigma_i}\left(x,y\right)$ is a normalised function, whose individual positions are either of Gaussian kernel function nature or adaptive geometric Gaussian kernel function nature whose sum either way will be equal to 1. The density function $F(x,y)$ which is the absolute sum of all pixels representing the total count of people within a particular image frame can then be applied to solve the perspective distortion problem. However, the addition operation then raises the possibilities of inevitable challenges by creating potential false peaks in certain instances, which then lead to the loss of position information.

Non maximum suppression, focuses on extrapolating extreme points thus maximising local searching. The presence and interference of false peaks in DM leads to the detection of several false positions when non-maximum suppression methods are employed. To calculate the total number of people, the regularized Gaussian kernel is used, though this results in the loss of location information. On the contrary, applying non-maximum suppression techniques on PCM, yields better results since each person's head corresponds to a peak thus that extreme point set out as P can be calculated as follows:

$$P=\bigcup_{i=1}^{W}\bigcup_{j=1}^{H}\left\{\arg\max\left(F\left(x_i,y_j\right),\delta_4\right)>\vartheta\right. \tag{4}$$

where, $F(x_i,y_j)$ denotes the $(i,j)^{th}$ pixel in PCM with the size of (W,H), δ_4 represents the four-pixel neighbourhoods, ϑ is the confidence, and argmax denotes the subscript to get the maximum value. The individual PCM pixels compare themselves with their corresponding four domains. The highest point within the domains becomes the local maximum meaning that particular head position P is a local maximum as well as greater than the set confidence.

1.2 Datasets

Currently, the mainstream crowd count dataset includes JHU Crowd dataset, NWPU Crowd dataset, ShanghaiTech, and UCF_CC_50. In the framework, we performed experiments on the above four datasets, each of which is described as Table 1.

1.3 Training Details

Here is the training detailed information on the ResNet-DC. It is important to note using a pre-trained ResNet, makes it easy to achieve convergence with ResNet-DC, quickly.

1.4 Label Normalize

Guerroum[60], *et al.* points out how a small regression DM value affects the performance of an entire network. Bearing the same effect on the PCM, we set the amplification factor on the PCM to 10. Applying the amplification factor when the PCM value is too small, makes it very easy for the network to predict the wrong peak value. On the other hand, a higher PCM value, will make it difficult for network convergence.

1.5 Data Augment

Al-Thani[53], *et al.* established that cropping may lead to the loss of critical global information, thus introducing the need to only horizontally flip once the original image to achieve twice the image.

1.6 Loss Function

Most significant research work[51,53-57] have made use of the mean square loss (MSE) for error evaluation. We have also used the MSE loss function as defined below:

$$L_{mse}(\theta)=\frac{1}{2N}\left|F(I_i;\theta)-G_i\right|^2 \tag{5}$$

where, θ represents the parameters that ResNet-DC needs to learn, N represents the number of pictures, $F(I_i;\theta)$ represents PCM predicted by the i^{th} input image, and G_i represents the ground-truth PCM of the i^{th} input image I.

We have used a convolution *kernel* = [[−1, −1, −1], [−1, 9, −1], [−1, −1, −1]] of size 3 × 3 in order to convolve with PCM to get the relationship between adjacent pixels. This loss can then be defined as:

$$L_{\ker}(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\left|F(I_i;\theta)*\ker nel-G_i*\ker nel\right|^2 \tag{6}$$

By using the kernel so as to convolve with PCM we attain the difference value between the centre point, and its subsequent eight neighbourhoods, only then do we proceed to calculate the territorial mean square error. The total loss L=(θ) can be calculated as follows:

$$L(\theta)=L_{mse}(\theta)+L_{\ker}(\theta) \tag{7}$$

1.7 Learning Setting

By the set standards of transfer learning[30] to accelerate the model convergence, we have to apply an end-to-end straightforward training structure to the ResNet-DC, coupled with a well-trained ResNet-18 and supported by a fine-tuned Backend. The upsampling, draws its initial values come from a Gaussian initialisation with 0.01 standard deviation. Using the Adam optimisation algorithm, the learning rate is 5*e* -5, and the weight decay rate is 1*e* -4. The input image is regularised (mean and variance on the Imagenet dataset) and then trained on the dataset to predict PCM.

1.8 Evaluation Metrics

The existing methods [14,19,25] presents the mean square absolute error (MAE) and mean squared error (MSE) for use to evaluate the performance of crowd counting, and they are of equal importance,

Mean Square Absolute Error (MAE) is regarded as the accuracy of the prediction,

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|c_i-c_i\right| \tag{8}$$

Mean Squared Error (MSE) is regarded as the generalisation ability of the model

$$MSE=\sqrt{\frac{1}{N}\sum_{I=1}^{N}(c_i-c_i)} \tag{9}$$

where, N is the number of pictures, c_i is the number of people in the i^{th} picture, and $\widehat{c}_i$ is the number of people predicted in the i^{th} picture,

2. MOTIVATION

Motivated by the above-mentioned challenges around crowd counting in still images and its subsequent application on RoboSoldier, we decided to comprehensively review the existing work done, a survey of literature.

2.1 Crowd-Counting Categorisation

It helps researchers into further exploiting and directly manipulating any crowd counting branch and obtaining the maximum possible output.

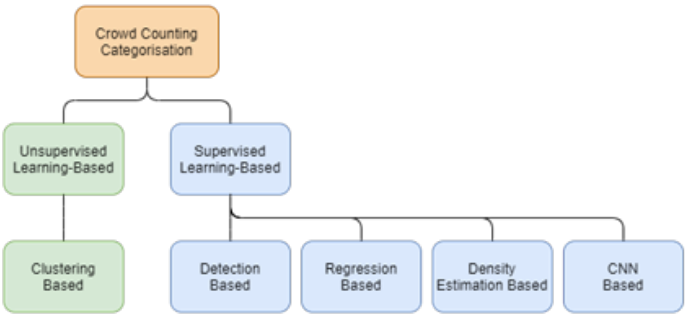

Figure 2. Crowd-counting techniques[10].

Early work in crowd counting methods were centred on estimating human count through either detection or regression-based frameworks. However, the inability for traditional crowd counting models to consider potential information within existing data sets led to the emergence of more descriptive-centred crowd counting models[1-3].

To summarise, this paper we specifically reviewed the literature against the following two broad themes:

- Traditional Crowd Counting techniques
- CNN-Crowd Counting techniques

2.2 The Crowd Counting Argument

Significant achievements in error reduction and advances in precision counting have achieved and outlined how crowd

counting systems must respond when faced with severe occlusion, scale variation, extreme weather, and high clutter as well non-uniform crowd distributions amongst other challenges as highlighted as follows:

- They must be considerably accurate and
- They must be robust

2.3 Methodology

The Fig. 3 depicts a 4-step model of how we have structured our literature review.

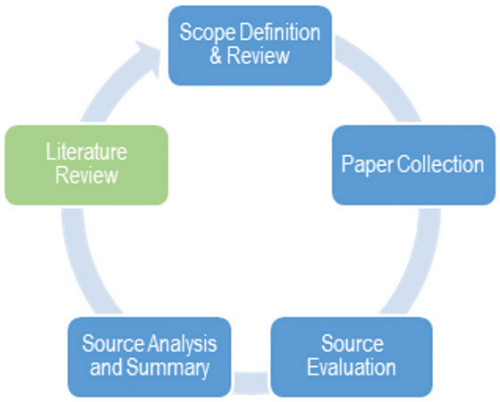

Figure 3. Flow diagram of systematic survey of literature.

3. IMPLEMENTATION

3.1 Defining the Scope of Review

A comprehensive survey of literature through the following Journals, IEEE Computer Science, ACM, Elsevier, Springer and Google Scholar as the search engine of choice was created for the study

3.2 Evaluating sources

Only 60 papers are appropriate literature were analysed for the research.

3.3 Paper Collection

Only the papers that appeared in journals relevant to the intended objective of the survey of literature were studied. Out of the 968 references collected only 60 are clearly aligned with the topic of the research, as shown in Fig. 4.

3.4 Analysing and Summarising Sources

In going through each paper and analysing it thoroughly considering its independent contributions and summarising them involved the simmering and rigour process of paper evaluation.

3.5 Organisation of the Survey of Literature

The survey of literature is structured as follows:

- Traditional crowd-counting methods;
- Discussion of complete, detailed and operational crowd-counting mechanisms.
- Features-based CNN-CROWD COUNTING techniques categorisation
- Discussion on applications of CNN-CC techniques

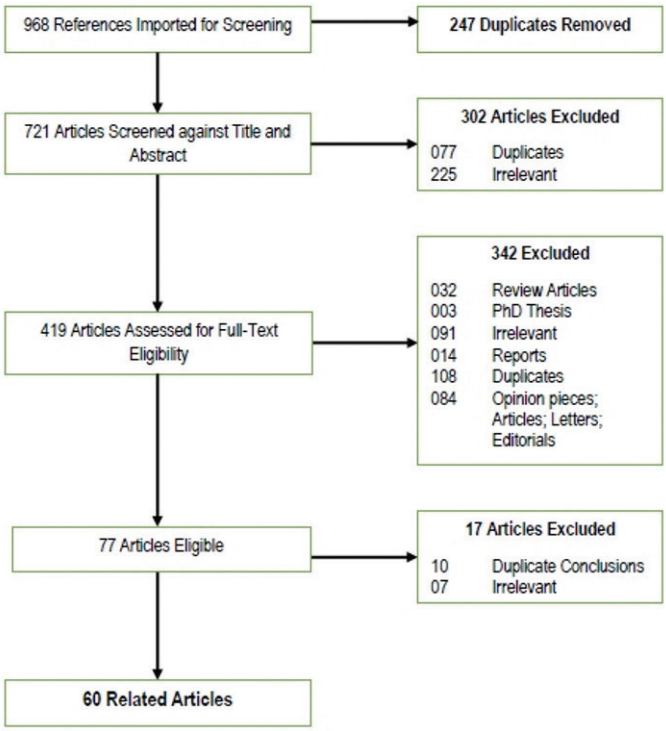

Figure 4. Primsa flow diagram.

4. LITERATURE REVIEW

4.1 Traditional Crowd Counting

Solving the crowd counting problem using traditional counting methods involves research around carefully designed features[9]. These methods are subdivided as follows:

4.1.1 Supervised Crowd Counting

Gathered input data, are defined and then labelled, the role of the machine becomes that of determining the hidden pattern or objective function only[5, 10, 11, 12].

4.1.2 Unsupervised Crowd Counting

The input data, its associated labels are both unknown, the role of the machine becomes that of categorising and labelling the raw input data prior to the determination of the objective function[13-14].

4.2 Counting by Detection

Computing image information is done by abstraction and harnesses point based local decisions to gather knowledge about the point's existing features. This approach acquires each individual's approximate location within the scene by utilising either the face or human detectors, head and or available segmentation algorithms[15].

- **Monolithic detection**[12] here a set of trained images are used to assist in detecting the complete individual this is ideal for low density crowds.
- **Part based detection**[16] employs the part-based approach by classifying an individual based on a particular body part.
- **Histogram of Oriented Gradients (HOG)**[17] this is an integration of a set of carefully designed features head detection,

- HOG, Fourier analysis, and associated points of interest brought together to overcome the prime single feature approach disadvantages.
- **Shape matching detection** approach apart from simply detecting the individual's location in the image it further conjointly detects the individual's exact pose within the same image or frame.
- **Crowd segmentation**[18] approach works through the estimation of potential spatial arrangements of pedestrians in a scene and from the resulting observations explains the existing image features.
- **The human 3D model** constitutes of four flexible ellipsoids coupled by adjustable parameters aimed at achieving an estimate optimal solution, by harnessing the Reversible-Jump Markov Chain Monte Carlo (RJMCMC)[15].

4.3 Counting by Regression

Counting by regression[19] is composed of Regression modelling and Low-level feature extraction. Their explicit task is to encode lower level information in the form of edges, textures, and gradient features and make an estimate headcount within a given still image[20-21].

- **Counting by global regression**[8] this utilises clustering or detection methods and when properly trained and configured at the low-level features it's appropriate for counting in densely crowded scenes.
- **Regression-based crowd-density estimation:** The extraction of the foreground area and edge low-level features from the still image frame, will result in the mapping between both the actual head count and the estimated head count applying a linear-regression model approach[20].
- **Shape and part-based detectors** here the low-level feature extraction component as well as the regression modelling component both assist in the absolute defining of the counting by regression pipeline approach[21-22].
- **Blob based algorithms** are easily implemented without any training requirements, and perimeter-area ratio blob-based holistic features appear to yield better results
- **Blob regression technique**[14] uses the sum total of group estimates to achieve crowd size by simply applying regression techniques to each and every available blob within the image resulting in segmented group counts.
- **Grid of sub-regions** related approaches are widely used in the detection of binary classifier related local abnormalities used in the classification of discrete density levels or alternatively in cell based crowd count crowds[23].

The techniques mentioned above are more inclined to the global nature of the scene. However, there are also local features and textures as listed follows: 1.Gray Level Co-Occurrence Metrics (GLCM); 2. HOG; and 3. Local Binary Pattern (LBP).

These are implemented so as to learn and improve the relative mapping between existing low-level features and the ground reality crowd count to be achieved after having extracted both the local and the global features across a wide range of regression based methods like linear regression[24], Gaussian[25], and Ridge Regression[26], and Neural Networks (NNs)[27].

4.4 Counting by Density Estimation

This is used to solve the data dismissal problem which is predominantly caused by global regression based counting[28]. By making use of spatially available information and applying a density estimation approach to it, this technique makes it possible to overcome the occlusion and clutter problem.

In an effort to yield the count of people in a particular region Lin[28]*, et al.* applied density integration over the same region resulting in what they referred to as fractional crowd density estimation. Inorder achieve the fractional density at each individual pixel they made use of the linear model with Multivariate Linear Predictive Spectral Analysis capabilities.

- **Holistic Approaches**[7] predict crowd size estimates by making use of available global image features and are often referred to as "mapping-based" approaches.
- **Textural approaches**[8] are primarily built around the concept that the lower the crowd density the courser the textures and conversely the high the crowd density the finer the textures.
- **Grey Level Co-occurrence Matrix (GLCM)** was developed and still evolves around existing statistical models to predict and estimate crowd densities[24].
- **2D Discrete Wavelet Transform (DWT)** this was developed to operate on the strength of textural features extraction[29].
- **Edge features approach:** In order to detect body parts like legs and arms proposed edge feature tools like vertical edges have been used in the segmentation of foregrounds using existing background modelling techniques[30-31].

5. CONVOLUTIONAL NEURAL NETWORK

A Convolutional Neural Network (CNN) is an approach based around image retrieval rules of deep learning algorithmic nature assigned to all the possible aspects contained within an image of chosen categorisation from those surrounding it[33]. Within this "CNN" is an inherently hidden set of mathematical processes and principles that are aptly referred to as the convolution[34]. A rather unique specific architecture is used in the implementation of the Network which has data training capabilities drilling down to the base of as many network layers as available[35].

5.1 CNN Architecture

5.1.1 Convolution Layer

In order to be able to perform visual image analysis the convolution layer or the core layer of the CNN architecture has generally been employed for these tasks. The convolutional layer is primarily applied when there is need for the generation of output maps, through the simple operation of bias addition and activation application operation in the order of non-linear activation.

5.1.2 Pooling Layer/Subsampling Layer

The pooling layer is sandwiched between existing convolution layers and there are primarily three pooling

methods[36] namely max pooling, mean pooling, and stochastic pooling. Max-pooling is more potent and enjoys more benefits because it curbs over fitting by decreasing existing size dimensions.

5.1.3 Fully Connected Layer

Due to the conceptual similarity in the arrangement of its layers it has been since referred to as the traditional neural networks building block, thus elevating high-level filtered images consisting of an array of complementing convolutional layers alongside pooling layers built around a fully connected layer.

5.2 Crowd Counting Using CNN

It has become more practical and efficient to solve crowd counting problems using CNN approaches due to the rapid advancement of research work in the development of application tools around the deep learning concepts. CNN methods are developed and built around two clear-cut branches namely: Methods based on density maps and Methods based on detection. To perform analysis on crowds as well as to construct 3D body parts a whole new set of differing types of CNN are needed and they include those based on LSTM algorithms. These set of new techniques proved that when applied they greatly aided crowd counting through regression, detection, density-estimation and clustering [37].

5.3 Categorisation of CNN-Crowd Counting Techniques

This is key in creating a level of sound understanding among the researchers and empowering them into designing both novel distributed control algorithms as well as monitoring algorithms applicable on a wide range of crowd-counting applications concerning public gatherings, military defence etc[32-33].

5.4 Network-CNN-Crowd Counting Techniques

Through this technique an existing network undergoes modifications with relation to its layers or columns. This may involve the addition of any existing module to cater for segmentations, classifications, and surveillance finally altering the behaviour of the original network and these are referred to as Network-CNN-CROWD COUNTING. Unfortunately, due to their architectural complexities these have proved to be not computationally sound even with relation to crowd counting in real-time.

5.5 Basic CNN Crowd Counting Techniques

This is where all Crowd-counting techniques bearing the fundamental CNN architecture can be found. These are regarded as pioneering methods of deep-learning for applications on density estimations which are applicable for obtaining a real time crowd count owing to their simple network Architecture[38].

A Fully Convolutional Neural Network (FCNN) model for attaining a denser and diverse environment through increased data training sets was introduced[39-40].

The ultimate use of a basic CNN architecture on a wide range of indoor and outdoor counting applications was proposed by Sam[42], *et al.* Through this ensemble approach, if adequate training is applied to every network purely on the strength of improving on the previous errors, it can easily lead to a fine-tuned network.

Techniques that fall under this distinct subcategory tend to focus more on density estimation, and not on the crowd counting. However, their oversimplified architecture leads their underperformance in areas of high occlusion coupled with varying perspective occurrences. However, by removing redundant samples, the practical speed of density estimation within these techniques may be improved further. Also, the iterative reduction of errors in different network layers, may also reduce the error-rate probability.

5.6 Context-CNN-Crowd Counting Techniques

This includes all techniques available and used for crowd counting which implement contextual information and global contextual information within existing images for near-accurate counting improvements. Image contextual information simply refers to nearby pixels and their underlying relationships or neighbouring information within a targeted and acceptable area for comprehensive improvements. These are considered very useful in obtaining pure density levels as well as the distribution pattern within images of varying densities[41].

The proposal of an everyday object count after working on an idea involving associative subitizing simply meaning that the human ability to be able to respond with quick count estimates and or cost assessments in cases involving smaller object counts could be harnessed otherwise[42-43]. Dalal[13], *et al.* went on to propose the attention model aimed at detecting the head location where head location becomes synonymous with high probability. Alternatively, suppression of the nonhead regions was achieve through multiscale feature branches implementation. Walach[44], *et al.* hybrid approach, relates to possible CNN combinations and a dilated convolution, which is the use of expanded kernels to replace pooling, in an effort to improve the density map quality.

Wang[45], *et al.* split an image into varying small and overlapping patches and built a still image-based crowd counting CNN–Markov random field algorithm. The extracted features from both the fully connected Neural Networks as well as the overlapping patches, were implemented in the patch count regression operation.

The accurate counting task to estimate object count using a density-adaptation-based network[39], was proposed based on training one dataset over a generic framework and then later port another training dataset to it. Computing of the density level was through a selective training of various datasets over a network. This constitutes a three-tier architecture of networks: where the network for density adaptation is applied in identifying both low densities and high densities, whilst the remaining are counting networks.

Use of a spatial transformer counting algorithm whilst at the same time solving both scale and rotation variations was proposed by Liu[46], *et al.* under the deep recurrent spatially aware networks classification.

5.7 Scale-CNN-CC Techniques

These are Basic CNN Crowd Counting approaches that

were continuously improved over time, to address issues surrounding robustness improvements and accuracy. Scale variation refers to the varying resolutions resulting from varying perspectives. These are critical techniques in the enhancement of attaining accuracy within areas of high congestion and high occlusion. The goal for crowd counting is made comparatively easier by extracting multiscale patches from input images.

Wang[47], *et al.* made a proposal for a geometric aware based crowd density estimation technique and a corresponding and explicit model in dealing with effects of perspective distortion was further proposed by Wu[48], *et al.* through their research on the idea of stacked pooling that managed to reduce the computational cost. By using stacked pooling to extract scale information instead of multiscale kernel pooling it makes it more appropriate to solving real-time applications. In further research Xiong[49], *et al.* proposed use of an image pyramid in dealing with issues related to scale-varying within an image. Without altering the existing filter size; accuracy in crowd counting was achieved by simply feeding the images that had been down sampled back into the network.

High level semantics that is relating to face and body, including low level features and the accurate estimation of crowd density within a scale varying spectrum can be effectively captured by the unique application of a set of both deep networks and shallow networks. The use of a single column multiscale method with an array of cost-effective measures applicable under real-time conditions were developed[50]. Scale related features for use over dense crowd counting were extracted through the single column network with a multiscale blob approach.

A unique combination built on a gating and a multiple-expert CNN combination was developed by Kumagai[24], *et al.* The role of the gating CNN is to feed the expert CNN with its input patch which in turn gets the overall algorithm very robust for larger changes in appearance. Loh[34], *et al.* presented two novel methods of addressing both variations in scale and crowd appearances within images. First, a congested multivariant scale region-based Hydra CNN, without geometric information and secondly, a counting based CNN with image appearance mapping to density maps.

Enhanced quality density maps can be achieved through transpose convolution. Kumagai, *et al.* highlighted the use of Generic Adversarial Networks (GANs) for crowd counting, by drawing from its success in solving image translation problems. In generating their density maps they used GANs to translate both the image and the corresponding patches. A comparison between the original Gradient Temporal Difference (GTD) with its correspondingly generated map was done thus ascertaining the best possible density map resolution. For this work they employed the regulariser Adversarial Cross-Scale Consistency Pursuit network (ACSCP) with the capabilities of maintaining the whole image or parent and four patches or child. This relationship works well in reducing losses in counting which would have been previously triggered through averaging. In applying the adversarial loss method, loss minimisation calculations can be done for the distance running between the still parent image and the concatenated image density maps.

5.8 Multitask-CNN-Crowd Counting Techniques

Apart from accounting for crowd counting they account for classification tasks, uncertainty estimation tasks, segmentation tasks as well as crowd-behaviour analysis. Lempitsky, *et al.* succeeded in using the ConvNet architecture in the counting of penguins. Owing to a wide environmentally varying scale coupled with occlusion a proposal for a multitask learning technique was made. Idrees, *et al.* working on the three common challenges namely density estimation, localization and crowd counting went on to investigate the multitask technique. Their work ultimately established the unique relationship from the counting activity to localization and density estimation.

Zhu[50], *et al.* through direct extraction of features from a deep and shallow FCN went on to concatenate it with the summation of two deconvolutional layers and they postulated another deep and correspondingly shallow FCN, whose final output image appears more similar to the original image. Apart from the modelling of the absolute visual properties only, Huang, *et al.* went on to propose a body structure aware semantic scene CNN-based crowd counting method. Here they used decomposition techniques to reduce the problem of crowd counting to a much simpler multitask problem. However, these multitasks resulted in the loss quality semantic-feature information. Al Thani[53], *et al.* proposed the Multicolumn Multitask Convolutional Neural Network (MMCNN) which drastically overcame the scale variation problem in images. They applied the multicolumn method by effecting three major changes: First, utilising both upsampling and down sampling techniques in the extraction of multiscale features. Second, accounting for the losses suffered due to down-sampling by using deconvolution. Third, the reduced loss per scale and this made the features more discriminative. Liu[33], *et al.* increased the training data in order to improve on accuracy through the proposed self-supervised method. Multiscale sampling was performed on cropped image patches borrowed from actual images in order to further improve on accuracy.

5.9 Image-View-CNN-Crowd Counting Techniques

It is important to analyse the input image, be it from a perspective or aerial dimension and then design an appropriate network with improved accuracy. These techniques are applicable in a wide of areas like medical imaging, drone surveillance, and people counting. Image-view-CNN-CROWD COUNTING can be addressed in two subcategories owing to camera properties like, angle, tilt, and position.

5.10 Aerial-View-CNN-Crowd Counting Techniques

All the approaches and techniques used in network designing using the aerial view-based input image are described here. They have a broad range of applications from commerce, healthcare to military, etc. Khan[23], *et al.* did complete and significant research work leading to the development of a working method on cell counting within a growing human embryo. To perform cell counting in microscopic range an end-to-end deep CNN was used.

5.11 Perspective-CNN-Crowd Counting Approaches or Techniques

Here we are looking at those techniques that mainly

help in the design of networks according to their perspective-view-based input images. These techniques yield better results when applied into a perspective varying scenario whose scale variations differ significantly. These are highly effective in densely populated crowd scenarios associated with several perspectives like shopping malls.

Xiong[49], *et al.* proposed for a perspective information-based adaptive CNN. In this regard considering the available perspective information the weights bearing around the convolutional filter are implemented with respect to the existing image scene. Kang[21], *et al.* made a subtle proposal for a perspective-embedded deconvolution network which modelled the differing sizes of the people in relation to perspective distortion. To obtain the GTD, a Gaussian function was applied using location-aware capabilities of differing kernel parameters for each of the dots.

Using the deconvolution network or alternatively using the upsampling process for embedding perspective information within images applies the structured information at every level leading to more accurate high-density maps. This technique also reduces complexity by removing the duplicate cyclic parameters. A pretrained network classifier was implemented to do the high-level features extraction. The object mapping count is achieved through the implementation of the FCNN on the newly extracted object features. In order to be able to switch between prior learned visual domains like people counting, wildlife census, embryo cell counting, and traffic use of parameter subsets of interdomain sharing is done. To tackle varied perspectives, density variations and scales the application Interdomain switching techniques is important. Multi objective tasking driven switchable training techniques and the perspective map driven perspective-aware CNN model are also correspondingly important tools.

This is a very important layer tasked with the role of combining the differing density maps generated from different scale feature maps. A combination of these density maps and the perspective maps will produce the estimated count. The varying perspectives challenge and working solution to the same have been availed by Yao*, et al.* through the proposition of a CNN and LSTM methods based Deep Spatial Regression Model (DSRM). The summation of the available local counts generated the final working count.

5.12 Tube-CNN-NN Techniques

The techniques used to train the CNN are differentiated by the approaches they apply in the CNN training process, for example, complete images or cropped patches-based training. These approaches are best at either network accuracy prediction or improving the density map quality. The whole image-based training method achieves network computational cost minimisation by heavily compromising the accuracy of the network. On the contrary, though the patch-based training increases the network accuracy, unfortunately it also increases the cost computations significantly. However, regardless of their advantages or disadvantages thereof, these techniques remain very valuable in commercial, medical imaging, and military applications.

5.13 Patch-Based-CNN-Crowd Counting Approaches or Techniques

By running the sliding window algorithm over a chosen test image built out of cropped patches the CNN method can be trained. These usual scenarios are helpful in applications where high quality density maps are demanded without compromise, for example in cancer diagnosis. The main objective remains as the design and development, of higher quality density maps decorated with corresponding higher computational cost.

The use of a much smaller network in estimating the object count within a given receptive field is done as opposed to whole image crowd count estimation. The crowd count overestimation scenarios in sparse areas characterised by regression-based techniques and on the other hand, the crowd count underestimation in densely populated areas characterised by detection-based techniques gave birth to the detection and density-estimation (DecideNet) method.

The skip connection technique which is generally used for crowd-counting scenarios, led to the optimisation technique for information flowing through varying sets of convolution and deconvolution layers. The Convolution layers work on detecting the edges and colours. Though it was discovered later that this did not contribute to the network performance improvement. However, through the application of the Gated U-Net in the determination of the amount of information passing through the final or fully connected layer for a more accurate feature-selection process was proposed. In line with the conceptualisation of this idea they proposed the Digcrowd method in working with high-density images as well as images with highly differing perspectives. The image is split into two: the far view and the near view regions, where counting is by performed detection.

The use of head detection techniques has been done in identifying varying human head sizes within images. The image division into several patches allows the Support Vector Machine (SVM) classifier to effectively classify crowded patches and non-crowded patches alike. In establishing the head size, regression techniques have to be applied on every image patch. Only after establishing the head size, can the total number of heads in the patch be calculated through the division of the patch area with the head size.

Table 1. Summary of the four datasets

	A	B	C	D	E
1	1280	16,795	13.1	320*640	-
2	2000	63,325	31.2	640*480	-
3	482	241,677	501.4	Different	968*589
4	716	88,488	123.6	1024*768	-
5	50	63,974	1279.5	Different	902*653
Key					
1	JHU crowd				
2	NWPU crowd				
3	Shanghai Tech Part A				
4	Shanghai Tech Part B				
5	UCF_CC_50				
A	Datasets images				
B	Count				
C	Avg. density				
D	Resolution				
E	Avg. resolution				

A count-net approach or technique was proposed by Chan, *et al.*[7] and applies background filtering on the head portion. Here crowd counting extracts the features and performs the classification simultaneously. Zhang, *et al.* calculated density estimation using the geometry adaptive kernel through a novel patch-based Multicolumn CNN (MCNN) crowd counting technique. Management of the scale of varying heads or objects was achieved through the harnessing of receptive field size variances in each CNN column, though the ED map quality decreases with each aggregation. Wang, *et al.* worked and presented the Skip connection CNN (SCNN) technique for crowd counting. In extracting features of differing sizes this network used the four multiscale units approach. The existence of a plethora of multiscale units lead to the adoption by cropping two patches from each input image of an augmentation strategy.

Sam[42], *et al.* worked on a switching CNN approach which considered three regressors trained on the following image patches: Low, medium and high density. In order to address any density variation issues the use of a classifier was applied in order to channel the input patch to a more corresponding regressor.

5.14 Whole-Image-CNN-Crowd Counting Approaches and Techniques

Here the execution of whole image-based inference operations is done. These computationally cheaper techniques are key in performing real-time applications. Here there is no negative sampling.

Applying a deeply simulated learning algorithm; a fruit-counting CNN-based technique was developed. Using synthetic data of 24000 images of variably sized tomatoes, the complete network was trained by using a training technique based on harnessing whole images. A more complex Inception-ResNet architecture was implemented for tomato counting.

The combination of both spatial and context information on semantic information and locality-aware features produced a more discriminative power within image representation[58].

The mapping of pixel space into a comprehensive semantic feature map using the CNN architecture. In using locality aware features both local and contextual information were appropriately harnessed. The use of ResnetCrowd in investigating density level classification, crowd counting, and violent behaviour detection resulted in the development of a multi-objective technique. In addressing both scale variation and high-density problems within images a Fully Convolutional Network (FCN) approach was developed.

Without changing the CNN receptive filter size field and by simply feeding a scaled-down version in the network more contrite results were achieved. They had to compute the average of all the down-sampled images so as to obtain the final count. In learning of both crowd density and crowd classification Sindagi[43], *et al.* worked on a multitasking cascaded CNN network[59].

Test result shows in Table 1, Table 2, Table 3, Table 4, Table 5 and Table 6.

Table 2. Counting performance of the different methods on JHU crowd

Methods	Type	Position	JHU crowd	
			MAE	MSE
MCNN	DM	NO	2.24	3.35
FCNCC	DM	NO	1.74	2.43
ResNet-14	DM	NO	1.48	2.22
DR-ResNet	DM	NO	1.39	2.00
ResNet-DC	DM	NO	1.36	2.02

Table 3. Counting performance of the different methods on NWPU crowd

Methods	Type	Position	NWPU	
			MAE	MSE
CNNLSTM	DM	No	2.24	8.5
ASA	DM	No	2.3	3.0
MGF	Detection	Yes	1.89	7.29
ResNet-DC	DM	No	2.33	2.89

7. CONCLUSION

As part of a series of research papers aimed at equipping RoboSoldier with the ability to give a near crowd count estimates, this paper, therefore, presents a comprehensive summary of both traditional crowd counting approaches as well as convoluted neural network approaches. In doing so the survey of literature outlined in this paper presents in its entirety an overall and more compelling research platform within which advanced crowd-counting research can draw its anchor from.

This crowd counting survey also highlights a current potential weakness within the crowd counting domain that there tends to be slightly more research bias towards crowd counting network structure designs at the expense of one of the key fundamental crowd counting problems, that of Location Information Loss. This therefore leaves much more coverage and research space for Location Information Loss/Recovery in crowd counting.

Table 4. Counting performance of the different methods on shanghai tech

Methods	Type	Position	Shanghai Tech			
			Part A		Part B	
			MAE	MSE	MAE	MSE
MCNN	DM	No	110.2	173.2	26.4	41.3
Switching CNN	DM	No	90.4	135.0	21.6	33.4
MNCS	DM	No	86.6	129.7	19.3	35.3
ASA	DM	No	83.9	133.3	18.6	31.1
Sang, *et al.*	DM	No	75.84	124.9	11.0	18.6
ResNet-DC	DM	No	79.85	131.2	10.8	18.6

Table 5. CNN-cc comparative study table

Datasets	Technique	CNN-CC	Architecture	Features
UCF	Wang[46], *et al.*	Basic	FCN	End-to-end deep CNN regression model
	Zhang[7], *et al.*	Context	AM-CNN	Attention model for head detection
	Li Min[26], *et al.*	Context	CSRNet	Dilated convolution & multiscale contextual information
	Han[15], *et al.*	Context	ResNet	Combination of correlation and MRF
	Liu[30], *et al.*	Context	LRN	Spatially aware network
	Zhu[50], *et al.*	Scale	ACSCP	GANs-based network, novel regularizer
	Idrees[19], *et al.*	Multitask	DenseNet	Multitasking with loss optimization
	Shi[49], *et al.*	Perspective	PACNN	Perspective-aware weighting layer
	Hu[17], *et al.*	Perspective	DSRM with ResNet	General model based on CNN and LSTM
	Zhang[7], *et al.*	Patch-based	Count-net	Aggregated framework
	Lempitsky[25], *et al.*	Whole-image	FCN	Multiscale averaging to handle scale variation
USCD	Zhu[50], *et al.*	Basic	FlowNet	Joint learning of crowd density and velocity
	Walach[44], *et al.*	Basic	Boosting Net	Two supervisory signals: crowd count and crowd density
	Han[15], *et al.*	Scale	VGG network	Image pyramid to deal with scale variation
UCF, USCD	Hu[17], *et al.*	Basic	ConvNets	Gradient boosting and selective sampling, and elimination of low-quality training samples
	Lempitsky[25], *et al.*	Scale	CCNN, Hydra	CCNN for mapping the appearance of image patch to its density map;
	Huang[18], *et al.*	Multitask	Multi-column	Body structure-aware methods
	Liu[33], *et al.*	Multitask	VGG-16	Self-supervised tasking
	Han[15], *et al.*	Perspective	ACNN	Incorporating side information in CNN by using adaptive convolutional layers
Mall	Kumagai[24], *et al.*	Scale	MoC-CNN	Integration of multiple CNNs
	Juneja[20], *et al.*	Multitask	MMCNN	Multicolumn multitask CNN focusing on drastic scale variation
	Liu[31], *et al.*	Patch-based	DecideNet	Detection and density-estimation network
	Xiang[49], *et al.*	Patch-based	Multi-scale network	Depth-information-based method STB
	Sheng[28], *et al.*	Whole-image	W-VLAD encoding	Pixel-level semantic-feature map, learning locality-aware features
Time-lapse image sequences	Basalamah[4], *et al.*	Aerial-view	Krizhevsky	Automatic approach to select a region of interest by computing a bounding box
BBBC005	Hernnandez[16], *et al.*	Aerial-view	VGG-Style NN	Feature pyramid network
RPE	Biswas[5], *et al.*	Aerial-view	VGG-net	Two convolutional regression networks

Table 6. Counting performance of the different methods on ucf_cc_50

Methods	Type	Position	UCF_CC_50	
			MAE	MSE
Faster R-CNN	Detection	Yes	592.09	672.19
MCNN	DM	No	377.6	509.1
Switching CNN	DM	No	318.1	439.2
MNCS	DM	No	306.7	396.3
DA-Net	DM	No	290.8	326.5
SD-CNN	Detection	Yes	235.74	345.6
ResNet-DC	DM	No	286.3	415.0

REFERENCES

1. Abbott, F.T.; Johnson, A.H.; Prior, S. & Steiner, D.D. Integrated Biological Warfare Technology Platform (IBWTP). Intelligent software supporting situation awareness, response, and operations. Tech. Rep. Quantum Leap Innovations Inc Newark De, 2007, pp 95.
2. Arteta, C.; Lempitsky, V. & Zisserman, A. Counting in the wild. *In* European Conference on Computer Vision. 2016, pp. 483–498.
3. Babu, S.; Deepak, S.S. & Babu, R.V. Switching convolutional neural network for crowd counting. Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition. 2017, pp. 5744–5752.
4. Basalamah, S.; Khan, S.D. & Ullah, H. Scale-driven convolutional neural network model for people counting and localization in crowd scenes. *IEEE Access,* **7**, 2019, 71576–71584.
5. Biswas, M.; Kuppili, V.; Saba, L.; Edla, D.R.; Suri, H.S.; Cuadrado-Godia, E.; Laird, J.R.; Marinhoe, R.T.; Sanches, J.M.; Nicolaides, A. & Suri, J.S. State-of-the-art

review on deep learning in medical imaging. Front Biosci (Landmark Ed), 2019, **24**, 392–426.
6. Chan, Antoni B. & Nuno, V. Bayesian poisson regression for crowd counting. *In* IEEE 12th International Conference On Computer Vision. 2009, pp 545–551.
7. Chan, A.B.; Zhang-Sheng, J.L. & Nuno, V. Privacy preserving crowd monitoring: Counting people without people models or tracking. *In* IEEE Conference On Computer Vision And Pattern Recognition. 2008, pp 1–7.
8. Chattopadhyay, P.; Vedantam, R.; Selvaraju, R.R.; Batra, D. & Parikh, D. Counting everyday objects in everyday scenes. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp 1135–1144.
9. Chen, K.; Loy, C.C.; Gong. S. & Xiang, T. Feature mining for localised crowd counting.*Bmvc,* 2012, **3**.
10. Chen, Y.; Yang, L.; Zhang, D.; Zhang, K.; Chen, B. & Du, S. Region-aware network: Model human's Top-Down visual perception mechanism for crowd counting. *Neural Networks,* 2022, **148**, 219–231.
11. Cheng, Z.; Qin, L.; Huang, Q.; Yan, S. & Tian, Q. Recognizing human group action by layered model with multiple cues. *Neurocomputing*, 2014, **136**, 124–135.
12. Chéron, G.; Ivan, L. & Cordelia, S. P-cnn: Pose-based cnn features for action recognition. Proceedings of the IEEE International Conference on Computer Vision. 2015, pp. 3218–3226.
13. Dalal, Navneet & Bill, T. Histograms of oriented gradients for human detection. *In* IEEE Computer Society Conference On Computer Vision And Pattern Recognition (CVPR'05). 2005, pp 886–893.
14. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D. & Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transa on Pattern Anal. and Machine Intelli.*, 2009, **32**, 1627–1645.
15. Han, K.; Wan, W.; Yao, H. & Hou, L. Image crowd counting using convolutional neural network and Markov random field. *J. Adv. Comput. Intell. and Intell. Inf.,* 2017, **21**, 632–638.
16. Hernández, Carlos X.; Mohammad, M.S. & Vijay, S.P. Using deep learning for segmentation and counting within microscopy data. *arXiv preprint arXiv,* 2018, **180***, 10548*.
17. Hu, Y.; Chang, H.; Nian, F.; Wang, Y. & Li, T. Dense crowd counting from still images with convolutional neural networks. *J. Visual Commun. and Image Rep.,* 2016, **38**, 530–539.
18. Huang, S.; Li, X.; Cheng, Z.-Q.; Zhang, Z. & Hauptmann, A. Stacked pooling: Improving crowd counting by boosting scale invariance. *arXiv preprint arXiv,* 2018, **1808***,* 07456.
19. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N. & Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. Proceedings of the European Conference On Computer Vision *(ECCV),* 2018, pp. 532–546.
20. Khanduja, V. & Juneja, S. Defaulter prediction for assessment of credit risks using machine learning algorithms. *In* 4th International Conference on Electronics, Communication And Aerospace Technology (ICECA), 2020, pp. 1139–1144.
21. Kang, Di; Debarun, Dhar & Antoni B. Chan. Crowd counting by adapting convolutional neural networks with side information. *arXiv preprint arXiv:1611.06748* , 2016.
22. Kannan, P.G.; Venkatagiri, S.P.; Chan, M.C.; Ananda, A.L. & Peh, L.-S. Low-cost crowd counting using audio tones. Proceedings of the 10th ACM Conference On Embedded Network Sensor Systems. 2012, pp. 155–168.
23. Khan, Aisha; Stephen, Gould & Mathieu Salzmann. Deep convolutional neural networks for human embryonic cell counting. *In* European Conference On Computer Vision. 2016, pp. 339–348.
24. Kumagai, Shohei; Kazuhiro, H. & Takio, K. Mixture of counting CNNS: Adaptive integration of cnns specialized to specific appearance for crowd counting.*arXiv preprint arXiv,* 2017, **1703***,* 09393.
25. Lempitsky, Victor & Andrew, Zisserman. Learning to count objects in images. *Adv. Neural Inf. Proces. Syst.*, 2010, **23**, 1324–1332.
26. Li, M.; Zhang, Z.; Huang, L. & Tan, T. Estimating the number of people in crowded scenes by mid-based foreground segmentation and head-shoulder detection. *In* 19th International Conference on Pattern Recognition, 2008, pp. 1–4.
27. Li, W.; Fu, H.; Yu, L. & Cracknell, A. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing,* 2017, **9**(22).
28. Lin, Sheng-Fuu; Jaw-Yeh, Chen & Hung-Xin, Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Trans on Systems, Man, and Cyber-Part A: Systems and Humans*, 2001, **31**, 645–654.
29. Lin, Zhe & Larry, S.D. Shape-based human detection and segmentation via hierarchical part-template matching. *IEEE Trans On Pattern Anal and Machine Intell,* 2010, **32**, 604–618.
30. Liu, J.; Gao, C.; Meng. D. & Hauptmann, A.G. Decidenet: Counting varying density crowds through attention guided detection and density estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 5197–5206.
31. Liu, L.; Wang, H.; Li, G.; Ouyang, W. & Lin, L. Crowd counting using deep recurrent spatial-aware network. *arXiv preprint arXiv:1807.00601,* 2018.
32. Liu, W.; Lis, K.; Salzmann, M. & Fua P. Geometric and physical constraints for head plane crowd density estimation in videos. *CoRR abs/1803.08805,* 2018.
33. Liu, Z.; Chen, Y.; Chen, B.; Zhu, L.; Wu, D. & Shen, G. Crowd counting method based on convolutional neural network with global density feature. *IEEE Access,* 2019, **7**, 88789–88798.
34. Loh, Yuen Peng & Chee, S.C. Getting to know low-light images with the exclusively dark dataset. *Comput. Vision and Image Understand.,* 2019, **178**, 30–42.
35. Mousas, Christos & Christos-Nikolaos, A. Learning motion features for example-based finger motion estimation for virtual characters. *3D Research,* 2017, **8**, 1–12.

36. Paragios, Nikos & Visvanathan, R. A MRF-based approach for real-time subway monitoring. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR 2001,* 2001.
37. Rahnemoonfar, M. & Clay, S. Deep count: Fruit counting based on deep simulated learning, *Sensors,* 2017, **17**, 905.
38. Razzak, M.I.; Saeeda, N. & Ahmad, Z. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps,* 2018, 323–350.
39. Rekabdar, Banafsheh; Christos Mousas & Bidyut Gupta. Generative adversarial network with policy gradient for text summarization. *In* IEEE 13th International Conference on Semantic Computing (ICSC), 2019, pp. 204–207.
40. Ryan, D.; Denman, S.; Sridharan, S. & Fookes, C. An evaluation of crowd counting methods, features and regression models. *Comput. Vision and Image Underst,* 2015, **130**, 1–17.
41. Sabzmeydani, Payam & Greg Mori. Detecting pedestrians by learning shapelet features. *In* IEEE Conference on Computer Vision and Pattern Recog*.,* 2007, 1–8.
42. Sam, D.B.; Sajjan, N.N.; Maurya, H. & Babu, R.V. Almost unsupervised learning for dense crowd counting. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8868–8875.
43. Sindagi, Vishwanath A. & Vishal, M.P. A survey of recent advances in CNN-based single image crowd counting and density estimation. *Pattern Recognition Letters,* 2018, **107**, 3–16.
44. Walach, Elad & Lior, Wolf. Learning to count with cnn boosting. *In Eu. Conference on Computer Vision,* 2016, pp. 660–676.
45. Wang, F.; Liu, K.; Long, F.; Sang, N.; Xia, X. & Sang, J. Joint CNN and Transformer Network via weakly supervised Learning for efficient crowd counting. *arXiv preprint arXiv:2203.06388*, 2022.
46. Wang, Li.; Shao, W.; Lu, Y.; Ye, H.; Pu, J. & Zheng, Y. Crowd counting with density adaption networks. *arXiv preprint arXiv:1806.10040,* 2018.
47. Wang, M. & Xiaogang, W. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. *CVPR,* 2011, pp. 3401–3408.
48. Wu, Bo & Ramakant Nevatia. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. *In* 10th IEEE International Conference on Computer Vision (ICCV'05)*,* 2005, **1**, pp. 90–97.
49. Xiong, F.; Xingjian, S. & Dit-Yan, Y. Spatiotemporal modelling for crowd counting in videos. Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 5151–5159.
50. Zhu, J.; Fan, F.; & Bo, S. People counting and pedestrian flow statistics based on convolutional neural network and recurrent neural network. *In* 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2018, pp. 993–998.
51. Baharun, N.; Razi, N.F.M.; Masrom, S.; Yusri, N.A.M & Rahman, A.S.A. Auto modelling for machine learning: A comparison implementation between rapid miner and python. *Int. J. Emerging Technol. Adv. Engin.*, 2022, **12**(5), 15-27.
52. Choudhary, D & Malasri, S. Machine learning techniques for estimating amount of coolant required in shipping of temperature sensitive products. *Int. J. Emerg. Technol. and Adv. Engin.*, 2020, **10**(10), 67-70.
53. Al-Thani, M.G. & Yang, D. Machine learning for the prediction of returned checks closing status. *Int. J. Emerg. Technol. and Adv. Engin.*, 2021, **11**(6), 19-26.
54. Rahman, A.S.A.; Masrom, S.; Rahman, R.A. & Ibrahim, R. Rapid software framework for the implementation of machine learning classification models. *Int. J. Emerg. Technol. and Adv. Engin.*, 2021, **11**(8), 8-18.
55. Huamaní, E.L. & Ocares-Cunyarachi, L. Use of artificial intelligence for face detection with face mask in real-time to control the entrance to an entity. *Int. J. Emerg. Technol. and Adv. Engin*., 2021, **11**(11), 68-75.
56. Huamaní, E.L. & Ocares-Cunyarachi, L. Analysis and prediction of recorded COVID-19 infections in the constitutional departments of Peru using specialised machine learning techniques. *Int. J. Emerg. Technol. and Adv. Engin.*, 2021, **11**(11), 39-47.
57. Kcomt-Ponce, E.J.; Huamaní, E.L. & Delgado, A. Implementation of machine learning in health management to improve the process of medical appointments in Perú. *Int. J. Emerg. Technol. and Adv. Engin.*, 2022, **12**(2), 74-85.
58. Perez-Siguas, R.; Matta-Solis, H.; Matta-Solis, E.; Matta-Zamudio, L. & Remuzgo-Artezano, A. Application proposal for gastritis diagnosis and treatment applying machine learning. *Int. J. Emerg. Technol. Adv. Engin.*, 2022, **12**(4), 32-38.
59. Agustono, I.; Asrol, M.; Budiman, A.S.; Djuana, E. & Gunawan, F.E. State of charge prediction of lead acid battery using transformer neural network for solar smart dome 4.0. *Int. J. Emerg. Technol. Adv. Engin.*, 2020, **12**(10), 1-10.
60. Guerroum, M.; Zegrari, M.; Masmoudi, M.; Berquedich, M. & Elmahjoub, A.A. Machine learning technics for remaining useful life prediction using diagnosis data: A case study. *Int. J. Emerg. Technol. Adv. Engin.,* 2022, **12**(10), 122-135

CONTRIBUTORS

Mr Tendai Padenga is the Founder Executive Dean of The School of Information Sciences and Technology at The Harare Institute of Technology, Zimbabwe.
He has supervised the entire research work.

Dr Vidhi Khanduja is an Assistant Professor in the Department of Computer Science at Hansraj College, Delhi University.
Her areas of interest include: Database protection, Information security, evolutionary algorithms, multimedia.
She has carried out the experimental work and put all the ideas in writing this paper.