International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 6855-6864

Crowd Density Estimation Using Image Processing: A Survey

Sneha.P.K, Rabichith, Sri Nithya S, Surekha Borra

Department of ECE, K.S. Institute of Technology, Bangalore-560109, India.

Abstract

People counting is a crucial subject in video surveillance application. Factors such as severe occlusions, scene perspective distortions in real time application make this task a bit more challenging. The use of Infra Red (IR) sensors and Channel State Information (CSI) of the WIFI network, which are the classical methods, give the count but have their own range constraints and its limited applicability to controlled environment. Video-surveillance systems are one of the advanced technologies used to estimate the density of people in a place for security reasons and to obtain the human statistics. The vision based techniques works well when people are in motion and when a high resolution image with clear background are available. This paper presents the state of the art of such image processing algorithms which are used for crowd estimation and their related applications.

Keywords: Convolutional layers; Crowd estimation; Fourier analysis; Image processing Neural network; Occlusion; Video surveillance

INTRODUCTION

People counting, monitoring and managing plays a vital role in preventing disasters like stampede in pilgrimages, concert, stadium, etc. Detection, tracking and description of a crowd density are important in the video surveillance systems, where they provide more instantaneous data and thereby provide an aid for pre detection and automatic detection of unusual scenarios using image processing algorithms and computer vision techniques [46]. There are many methods employed for the purpose of detecting the crowd such as Fourier analysis, segmentation, feature extraction [5, 6], pixel counting etc. in which eliminating the background can create problems in situations where camera setup is disturbed [6]. Image features like Scale Invariant Feature Transform (SIFT)[35], Histogram of Oriented Gradient (HOG) [36] etc, are not robust to problems related to occlusion and scale variations. Hence there is a huge demand for improving the crowd management system [9, 10].

Regazzoni and Tesei [2, 3] developed counting systems based-on CCTVs. These techniques measure the area occupied by the foreground pixels (used for counting the density) by eliminating the image background. Lin et al. [4] used a method based on the detection of head contour by Haar wavelet transform, followed by an estimation of the crowd size. Also, Granular Computing (GrC), a coming up computing approach of information processing [7], initiates human cognitive process by simulating absorption at different granularities. Epsilon-Support Value Regression (SVR) fusion-based approach is one such method which is an indicator of the chances of any potential crowd disaster. [11] Recently, deep CNN (ConvNet) has celebrated a huge success in many research topics under large-scale image recognition, object detection [13] and segmentation [14]. Marana introduced in texture-based features which is based on gray level co-occurrence matrix (GLCM) and self-organizing neural network [2] for classification of the crowd density.

Also as proposed in [39] the detection as well as counting of individuals is not limited, but is based on a group, that is the number of persons in a moving crowd is calculated as a whole. In one of the existing system it talks about using local crowd parameters likely the segment and texture features obtained from background subtraction for the calculation the number of pedestrians in a density of crowd. Albiol *et al* [51] proposed corner features from which the computation of motion takes place for clustering them into foreground features and background features. The authors implement the ratio of the moving points number and fixed/static points and obtain the number of moving persons. The image processing algorithms consider the region of foreground pixels and remove the background by various subtraction methods.

There has been a huge amount of research done in the field of human recognition, and it functions properly in situations where the crowd is less. Various existing CCTV's capture low resolution and hence many high resolution cameras which are cost effective were introduced. But since its necessary for a 24/7 uninterrupted functioning it generally makes it difficult to promise a good picture quality (contrast, brightness), and because of more number of cameras being vigilant simultaneously, the total time for processing has to be kept low (that is real-time considerations). As an extra problem, the higher the density of the crowd, the lesser is the number of pixels telling about a single person making it very difficult to use standard individual detectors such as histograms of oriented gradients. By keeping the above drawbacks in mind, approaches using image processing and regression algorithms have been given in the past.

This paper presents the state of the art of such image processing algorithms which are used for crowd estimation and their related applications. Section 2 gives different approaches to crowd estimation and the state of the art, Section 3 briefs the applications, Section 4 provides the challenges and Section 5 concludes the paper.

APPROACHES AND STATEOF THE ART

1. METHOD OF CROWD ESTIMATION IN SMALL CROWDS:

It is an easier [31] method for accounting the count in large crowds. The first and foremost step involved in processing is to detect the foreground blobs with the help of background subtraction. The method used is based on Multiple Gaussians [44]. In order to obtain homogenous blobs, tuning of algorithms and Morphological dilation of blobs [45] is done to fill the consistent gaps because of unavoidable statistical short-comings of algorithms.

The camera calibrations are achieved by Tsai algorithm which gives a set of parameters that can be used to project. The blob projection will result in a map plane called Ground plane [31]. When an object is present far from the camera, its projected size becomes bigger and becomes evident as the angle of the camera decreases. This can be reduced by placing the ground plan at an approximate height of the person and take the intersected area of those projections along with a parallel plane. The later plane is called as the Head plane. When blobs are constructed the spaces are filled by the morphological dilation. Hence the points lost due to the initial ground-plane projection are restored by the growing region, and the blob dimensions being affected much. The head plane should be adjusted practically, considering the two following facts. 1) Objects which are below Head Plane are removed by the double projection. 2) Only that HPH that is very manages with the first projected area issues at dissimilar distances partially.

The count is obtained from intersected area of those two projections. Further operations are performed on every blob and added to get the count for the present frame.

The total count estimation is then given by subjecting it to sudden oscillations. Then it was heuristically smoothened by using simplest Finite Impulse Response Filter (FIR), namely the Simple Moving Average (SMA) filter . As the background is estimated at first the count is always 0 initially. The outcome gives estimation accuracy requirements which are around 20%.

2. PIXEL COUNTING:

The pixel counting system [32] takes the geometric correction into consideration which is generally ignored by other methods. It gives the relation for the ground plane and geometric correction. The performance of this method relies on the foreground segmentation result [48]. Assuming certain conditions the authors formulated a certain rule of proportionality given as: N (persons) = a*N (pixels).

If the crowd density is moderate and occlusion is not a factor of importance, the above relationship still hold good but with a constant fixed term and few other parameters:

N (persons) = a*N(pixels)+b.

Another important parameter to be considered is the perspective distortion which can be corrected by using geometric correction. It first computes the scale for the lowest level plane (GC) and then discusses its applicability to the estimation.

After application of robust segmentation algorithm mask determination is done as the segmentation results in false foreground regions. The Region of Interest (ROI) is accounted manually and the mask is designed. It can also be found by using accumulation of the foreground pixels. The authors also developed a method which is adaptive to time variation and can be used for unusual crowd.

3. MOTION ESTIMATION AND MOTIONLESS DETECTION METHOD:

Samia Bouchafa *et al.* [33] specify the problems and solutions for motion estimation and also developed a motionless detection method to handle three difficulties: real time constraint, deformable objects and occlusion. This method is an optical technique that suits the crowd monitoring. The block matching technology is the basis for motion detection and estimation. This method uses three techniques as shown below:

Matching techniques: This method divides the image into blocks and compares them between two consecutive pictures according to a similitude criterion. A similitude function was chosen based on "add the differences in absolute value". The size of the block provides good results at the cost of sensitivity. Hence to reduce computational complexity they chose a smaller block size. The matching is done only along the edges. Hence they could not succeed in improving the computational time required to precede images on-line.

Frequential techniques: This method suits for spatial time related surfaces which is a constant phase to obtain the two motion components [49]. A set of Gabor filters are convolved with the image and using this technique they get the displacement vectors. The disadvantage here is that it is time consuming.

Differential techniques*:* This method is based on the assumption that the brightness of a moving point is constant with respect to time. The proposed system implements the modified Horn and Schunk [50] method in which the velocity vectors have a very small change between the consecutive images. A global crowd motion direction is obtained by using these results. To process the segmentation in good context two filters namely spatial filtering and temporal filtering are used.

In the motionless detection a module finds all those places where the motion is in action. By eliminating the respective areas and filtering the results stationery persons/objects can be located. The computation takes place only when there is no existence of motion. As a result, the stop duration gets a delay at every occlusion. To prevent the occlusion rate is recorded for a given frequency and the duration is corrected using the information. The accuracy of this system is found to be around 93%.

4. GRANULAR COMPUTING BASED IMAGE SEGMENTATION:

This paper puts forward a crowd segmentation based framework using granular computing (GrCS) to validate the issue of crowd segmentation to be analyzed at different hierarchy of granular, and to map issues to small problems. It shows it can make same pixels into proper atomic structure granules by dissolving the correlation in the pixel granules.

In GrC-based crowd segmentation (GrCS), in which granules are the fundamental thing, each step follows a non-identical level of granularity. This is done to obtain the capability of people to detect at various granular levels with the aim of mapping issues to manageable small problems. An extended version of Local Binary Patterns (LBP) operator called uniform patterns [23] is used to confront with variance in gyration of captured microstructures.

The granularity skeleton structures are logically atomic regions in the frame that gives the natural separated areas in between different structures of humans and background. The main aim of the atomic regions is to possess a pixel total process flexible to various crowd scenes. Hence this will be the best group diverse structure in the scene for robust crowd segmentation. This method mainly performs analysis in large crowd image by adapting principles of GrC to segmentation of crowd problem at different granular levels [4].

5. PARALLEL VIRTUAL MACHINE (PVM**):**

This technique estimates the crowd density in real time which is done on the basis of crowd image textures [29]. In this method input pictures are segmented in the form of classes of crowd density. The classification is then processed using a filtering technique (low pass) depending on the previous images from the incoming image sequences.

The first stage involved in this method is the categorization of every pixel obtained from the sequence of images into one of the previously recognized texture class. This process of classifying each pixel is carried out based on the method of Self Organized Maps (SOM)[42].The SOM uses the feature vectors of the textures taken from co-existed matrices[43]. As the classification of various pixels takes a lot of time in the real time environment, the method is extended using a distributed algorithm called Parallel Virtual Machine (PVM) [29].

The steps involved in this algorithm are:

- Initially, the master processer cuts the input image into n pieces called fragments (where n refers to the count of slave nodes in the bundle).
- Then each of the broken image fragments is put to the slave processor.
- The function of the slave processer is to perform the classification of texture of the image fragments by adapting a sequential algorithm.
- Further the slave processor sends the allocated pieces to the master.
- Lastly, the master arranges each and every fragment into a terminal texture-segmented image.

Generally the images of the are texture-segmented and texture histograms are calculated as a feature vector, followed by SOM neural network to carry out the classification of the input images into their crowd density classes. The method in the proposed paper [29] is experimented on a sequence of approximately 10,000 images captured sequentially from a tape that is recorded in a region of the airport. The technique was tested on different classes like the VH class, VL class and the H class, and the best results were shown for the VH class(95% accuracy).

6. SVR MODEL:

This method estimates the crowd density and number of the people in images. Due to alterations in the crowd images, the density may vary across the vision's field. To overcome this problem, images are separated into small patches of same dimensions called patches and this number of patches is determined. The information is extracted from different sources in terms of head count, confidences and absolute errors from the given patch. Further, to improve the count accuracy, the authors introduced cascade training of head images, with selection of bounding boxes covering all the positions and orientation of the human heads.

This method is a fusion of three systems: head detection, Fourier analysis and feature extraction. Head detection is done using HOG-based feature descriptor. This method is used to differentiate the local object and outline it by using edge detection and intensity gradients. The image is then segmented into tiny spatial regions called patches for which 1-D histogram of gradient directions or edge orientations of the pixels are calculated. The human head appears as small dots in high density crowd images. To overcome this problem, Fourier analysis of the image is done which is accurate in detecting human heads. Fourier transform of all the patches is obtained to get information about positions and large changes in the intensity values. The high frequencies in these patches are filtered out by Butterworth low pass filter. The target size is chosen as a trade-off between better results and increased time of detections [17, 18]. Fourier analysis is accurate for crowded patches but not for non-crowded patches. To solve this problem, the confidences of the patch is calculated and combined with the results of the determined Fourier model. The interests points in the images are then detected using SIFT features. The descriptors of these points frame the base strategy which classifies crowd and non-crowd objects. This also uses K-means clustering technique. Once the number of patches is obtained, the e-SVR model is done and is used to train the patches on their ground truths. They model is tested on crowd databases ([16], [15]).

7. FULLY CONVOLUTIONAL NETWORKS FOR DENSE CROWD COUNTING:

Boominathan et al. proposed [30] a method for estimation of crowd density by making use of still images. A deep with a

mixture of shallow, convolutional networks is used for the calculation of density route of a crowd for a particular image. The technique is evaluated for a standard dataset [36] and features are obtained. The system employed different networks as given below:

NETWORK ARCHITECTURE:

The targets are taken from different viewpoints which results in wide variety of mind set and variations in the scale. For a situation, where people are very much near to camera the head blob is taken. The proper detection needs a model to simultaneously work at high semantic level and also to determine the low line head patters. This method achieves the operation by the amalgamation of deep and shallow convolutional/twisted networks. Both of which are briefed in the following lines of the survey.

DEEP NETWORKS:

The deep networks help in capturing the high level semantics necessary for the crowd counting which uses architectural design similar to VGG-16 [43] network design. This VGG-16 system design was initially used for classifying the objects. The filters are extremely good at generic caption and have found their application in saliency prediction, object segmentation [37] etc. This approach is based on the classical energy levels of the VGG network by adjusting the filters in calculating the number. To overcome the problems involved in image classification, in which only one discontinuous label is given for the whole image, pixel level fore-sights are obtained by eliminating the full connected layers involved in VGG design. The VGG network consists of five pool layers of pace two each and therefore the features obtained have a spatial resolution of about 1/32 times that of the incoming image. The design [37] of VGG prototype have set the pool layer pace as 4 to 1 and have removed the 5^{th} max pooling part. Due to this the input resolutions are at 1/8 times the fore-sights. This approach also makes use of the concept of holes.

SHALLOW NETWORKS:

Shallow networks are used in detecting the low level head blob patterns that are present in the crowd images. This is done based on shallow convolutional networks. This arrangement is designed as shallow at the depth of only three convolutional layers as it does not need the high level semantics to be captured. All of these 3 layers have 24 filters with a dimension of 5x5.They make use of pooling layers to generate the spatial resolution of the network's estimation. The average pooling is used instead of max pools to make sure that there is no loss of count in the shallow networks.

COMBINATION OF DEEP AND SHALLOW NETWORKS:

The results obtained using both deep and shallow networks are combined together [30] with 1/8 times resolution of the incoming image and are processed by 1x1 convolution level. The results thus obtained are further down-sampled to the input image dimensions in order to obtain the final probability of crowd density which is done using bilinear interpolation. The probable density map is used for the calculation of the total count by using the summation function.

The deep convolutional network is trained by the Deeplab [37], a framework of Caffe deep learning version. The network was trained at a rate of 1e-7 using Stochastic Gradient Descent (SGD) optimization and momentum of 0:9.

8. CONVOLUTIONAL NEURAL NETWORK AND MARKOV RANDOM FIELD:

In this method, the characteristics borrowed from deep convolutional networks are used for recognition of images, detection of objects and segmentation of images. [21]. Convolutional Neural Network-Markov Random Field (CNN-MRF) is used for calculation of the human count in the scene. First the picture is separated into patches one above the other (overlapping) and then deep CNN is used to obtain the features from the overlapped and highly correlated patches. The number of people in nearing patches is same and may fluctuate drastically at some locations due to vehicles, trees, etc in the image. To smooth the counting result and to make the counts same as the ground truths the Markov random field (MRF) is used on the local patches. This technique is implemented on MATLAB using MatConvNet [20]. The CNN-MRF based method is particularly used to get the count of people in an image from different places.

9. DEEP DETECTION FRAMEWORK:

This method uses depth images to perform head detection on collected frames by an overhead vertical Kinect sensor [6]. It considers the number and quality of Region Proposal Network (RPN) positive anchors on the performance of faster R-CNN and proposes a solution. Kinect sensors evaluate in hardware, the depth map of a scene. They have adopted Faster R-CNN models and an RPN-alone model to detect heads on depth images.

Generally, there are two possibilities of solving the problem of positioning in object detection. One is the traversal method like a sliding-window detector [26] and the other is the gradual approaching process. Traversal programs develop multi-object and multi-category detection over the entire image. A small range of regression programs aim at reducing the amount of calculation and improve the efficiency and precision. While classical R-CNN methods are based on the colour or edges of the image.

A region proposal network used is a class-agnostic detector. An image is fed as a network input and a group of region proposals are obtained as outputs. RPN is like the sliding-window method. A microscopic network slides over the convolutional feature map to generate proposals. Many approaches are presented for addressing multiple scales. The first approach is to build features maps from blocks of images and to run the classifier at multiple scales. The second

approach is to run multiple filter scales on the feature map. The RPN uses pyramids of multiple reference boxes called which are different from previous methods to handle the multi-scale problem. The original Faster R-CNN uses VGG-16 [28] model as the back bone network, which is considered as the baseline. The default RPN in Faster R-CNN [27] uses a foreground ratio of 0.5 in the RPN. After computing the classification loss of a batch, if in image there are slight more positive anchors they are filled with negative anchors. The number of positive anchors in a batch is often less than negative anchors, and the default foreground ratio setting causes the positive and negative anchors in the batch to be unbalanced. The network is biased towards negative samples as they are dominating and would harm the performance. The deep neutral network detection method enables robust and efficient detection.

10. DEEP CONVOLUTIONAL NEURAL NETWORK:

Deep convolutional neural network (ConvNet) [3] is widely used in large-scale image recognition, object recognition and segmentation. The estimation of people densities by deep ConvNets directly extracts the image features and maps features to crowd density in different levels: low, medium and high. The human total for each range and the number of ranges itself may depend on the application and specific characteristics of the field. [12]

The steps involved in this method are as follows:

1) Convolutional layers are convolved with input image or feature maps with linear filter. The resultant characteristic maps represent the response of each filter.

2) Pooling layers are non-linear with down-sampling layers with maximum (or average values) obtained in every sub-region of input image or feature maps. The efficiency of translation is increased and the count of network parameters is reduced.

3) Non-linear activations on neurons which are the inputs are applied by activation layers. Common activations are sigmoid function, hyperbolic tangent function, etc.

4) Outputs are computed by fully-connected layers by connecting them to every characteristic map elements of the prior layer. [22]

The neural network is trained using frame samples in the train subset, which are categorized into Very-low, Low, Medium, High and Very-high based on the number of persons in the image. By quantizing the estimated crowd density, the output of the neural network is divided into 5 levels of crowd density. The performance is evaluated by the classification accuracy of the test subset. In this deep ConvNets method a new crowd dataset of subway scenes with over 160K pictures is used to evaluate the accuracy of crowd density estimation method. Experimental results for this method provide the best accuracy of 91.73% on average and this proposed method can perform better for practical applications [3].

Kang Han [2] evaluated an approach on UCF and Shanghaitec crowd totaling datasets. Result count is obtained by calculation of the sum of all non-overlapping patches. The proposed method was implemented on MATLAB using MATLAB tool box with CNN for Computer Vision applications. The high crowd density estimation [1] was tested on their own dataset of commercial dell laptop to check the accuracy and efficiency of the approach. Careful fine-tuning of the model parameters such as size of the clusters and thresholds for noise suppression gave convincing results. The absolute error of around 10 people per patch and around 415 per image is an improvement when compared to previous models. To reduce the dependency of neighboring patches and to improve the overall performance of the system smoothness constraints can be implemented on nearby patches and the time constraints can be bound for the estimation of the patches across the used crowd scenes and videos.

Shiliang Pu et al. [3] used both the googlenet and VGGnet which are fine turned with pre-trained models from image net. The input images are re-sized to 224x224, and 5 output layer channels are considered depending upon the index levels of crowd density.

In small crowd people count estimation [31] experiments were carried on the sequences taken from the PETS dataset. The images that differ in various parameters like illumination, dimensions, noise etc are used in order to examine the response to various values. The results had errors as scattered people in low density crowd had major gaps between them and hence it was counted. The algorithm is complied over different sequences at different times which gave accurate results.

Marana [29] used 9892 crowd images for real time human density estimation. Pixel counting method is carried out on a window from where co-occurrence matrices were obtained. The requirement was reached for real time crowd estimation since it took 1.025 seconds to process.

In another system, for motion and motionless detection [33] different situations were considered like an empty environment, crowded environment with stationery people and stationery people with different occlusion rate. But the ability to give false detection rate reduced due to the continuous running of the system. In some situations a person was not counted due to two reasons: low contrast and poor ability of the process. This method gives an accurate information for the performance of false detection. Author Ruihua et al. [32] investigated the geometric correction to study and talk about the perspective distortion. The transform obtained is generally used for geometric correction and is applicable only for the ground plane and not any human. But in this method they proved that applying the geometric corrections to humans leads to a linear relationship with the number of pixels. This can be applicable for effective crowd density estimation.

The above methods give accurate results but with its own disadvantages. Dynamic models for different filters, crowd sets for different scenarios and other factors can be integrated to get an effective estimation algorithm achieving a specific application.

Table 1: Comparison of Head Detection Methods

METHODS	ACCURACY	ADVANTAGES	DISADVANTAGES
SVR MODEL	~70%	Very low normalized mean and standard deviation values.	There is no proper time constraint taken into consideration during the estimation.
Convolution neural network and Markov random field	~79.1%	Due to the overlapped patches separated strategies, the nearing original counts are highly correlated.	For a crowd exceeding more than thousands this method shows high error rate.
Deep convolution neural network	~82.66-91.73%	The accuracy of cross-scenes is better evaluated using 160K density annotated images	This method doesn't employ rough people counting function.
Granular computing	92.6%	This method is effective in organizing blocks of same pixels into a batch to cope with perspective distortion, varying crowd, and scattered background	Granulated sight of various levels of granular is restricted when crowd scenes are below par
Deep detection framework	~70%	Runs in real-time at a frame rate of about 110 frames per second.	Cannot handle the condition where some objects are nearer to the sensor than the head.
Parallel Virtual Machine algorithm	73.89%	Real-time automatic human density estimation	Miss classifications are expected
Convolution networks	~60%	Effectively captures the high level semantic information as well as low-level features.	Requires large amount of training data
Estimation in small density crowds	~75%	Head blobs and other such information regarding the projections.	Camera with high calibration is required to capture small crowds.
Pixel counting	90%	The expression for the geometric correction of the ground plane is derived.	Automatic calibration is absent.
Motion estimation and motionless detection method	93%	Both motion based and motionless detection is done	The working speed is not optimized.

APPLICATIONS

Huang et al. proposed a crowd density estimation method to be implemented at different places like bus stands, subway platforms etc. The robustness of the method is high and hence it can be used for the wide area surveillance in subway stations.[52]

Ruihua et al.[32] proposed a system of public surveillance that goes beyond the basic CCTV approach. This method implements pixel counting for crowd density estimation. In the video captured, the foreground segment is separated from the background and geometric correction is applied wherein every feature of the segment is normalized before pixel counting in order to estimate the density of the crowd. This method finds its applicability in any open ground, localized region where light condition varies and also the at places where non moving objects exist.

Samia Bouchafa et al. [33] proposed a system where crowd estimation is performed by breaking down the captured video into two parts: the motion and the motionless. To process the motion parts, methods such as differential techniques and displacement vectors filtering are used to compare frames of the video and to match the extracted features. The motionless part of the scene is segregated by updating a reference image dynamically. A separate model is used to map part of the frame where motion is occurring. This method finds its application in airport baggage claim, securing corridors for any threats in general. Further this system can be used not only to estimate the density of the crowd but also to anticipate the crowd gatherings.

Lauro Snidaro et al. [53] proposed a system that focus on indoor crowd detection using smart sensor nodes to track the movements and to extract features required for event detection purposes. The sensors are categorized as nodes and they are the main features of the system. This system finds application in maintaining a log of access in a targeted building, and thereby making a value addition to automated security

systems. The system also provides a scope for future research on increasing the target range into outdoor environments.

Zhang et al. [54] defined the difficulty in detecting moving objects in a dynamic background example in an elevator. The frame of reference is hard to establish and hence the detection becomes difficult. Hence, as a solution to this problem the author proposed the use of model specific direction filter in which certain features are extracted from the frame of the video captured. For crowd counting, the contour of the head is the most suitable feature then by using this as a locus, a low level filtering is applied to approximate the position of the head or in other words the person itself. This system finds application in environments where a dynamic frame of reference can be used in maintaining access records of a building and detection of persons in a crowd.

Chen et al. [55] proposed an automatic person counting system specifically in a bus environment. The first step of the process would be to capture images at a steady rate. The image is then divided into manageable blocks and each block is separated based on its motion vectors. If the magnitude of the motion vectors crosses an established threshold, then the direction of the object in the frame is decided. This method is proposed specifically for a bus but with further work application scope can be extended to supermarkets, business exchanges etc.

Grantham et al. [56] aimed at performing human counting in a complicated scene which defines a scene where mobile and non-mobile individuals exist along with high contrast scene. The first step is converting the frame into a binarized image. The foreground image is then extracted by performing background subtraction. To minimize the processing steps that follow, the foreground image is applied through a threshold to obtain a count of the number of people in the scene. To locate the individuals the expectation maximization method can be applied even in a low resolution image. This method finds application in complicated environments like crowded crosswalks, railway station platforms etc.

CHALLENGES

There are several problems [57,58] involved in the detection of the head and in estimating the count. These problems reduce the accuracy drastically which is an important parameter. These challenges have to be resolved for an efficient detection and few of them are shown below:

1. The pixel counting method is employed for estimation of face problems related to perspective distortions. These distortions occur as far objects in the picture seem smaller than near pictures in the image. To overcome perspective distortions geometric correction can be used, in which all the objects are bought to the same scale. However these corrections too have not given accurate formulations.
2. Background subtraction techniques are used to extract the foreground features from the images in order to classify the objects based on their size, shape, colour etc. These subtraction techniques are complex and difficult to apply. Existing techniques have given considerable efficiencies but more techniques need to be developed to obtain 100% accuracy.
3. Blob extraction is a principle step involved in the detection of the head, and the shape of the blob projection should be precise for further processing steps to take place. The shapes account for the way in which the human packages appear and these packages provide different local densities. For an effective detection all the above parameters are very important.
4. Varying scales are one of the problems posed during the detection in highly dense crowds. These problems can be faced by training the augmented images. This helps in guiding the CNN to learn variations in scales. However the datasets available for training contain very few images and in deep learning techniques many images are required and in such cases multi-scale data augmentation can be considered.
5. Occlusion is another common problem that occurs during the detection and counting of the heads. Occlusion refers to the overlapping scenarios of the head. The differentiation of the two heads is necessary and several techniques have been employed to overcome these problems. Deep neural networks, pixel counting methods, and several algorithms can be employed to avoid the same. The existing techniques have been successful in clearing occlusion problems to an extent.
6. Time constraint is another challenge that has to be faced during the detection. The techniques involved in processing are many which include image acquisition, noise removal, image enhancement and feature extraction and the time required for performing these steps is more. Hence for real time implementation time is an important factor.
7. Robustness against camera shaking needs to be achieved as the steps involved in detection can be performed only when the image obtained is static and clear. In real time due to the positioning of the camera and other factors, the images acquired maybe not be clear. Techniques have to be developed that can provide effective results inspite of camera shaking.
8. Noise is another unavoidable factor that comes with the image. Many filtering techniques are used for the removal of noise. Median filters and low pass filters are most commonly used. Noise occurs at every step of the detection and hence after every process noise removal is mandatory for a precise detection and estimation.
9. Background clutter is another problem that has to be removed for the estimation of the heads present in an image. Existing methods have been successful in removing the clutters and in one of the methods which involve granular segmentation of the pixels the problem is best faced.
10. False positives and false negatives are an important factor that has to be considered at the output obtained.

False classifications lead to inappropriate results and thereby the accuracy of the process decreases. Due to unavoidable circumstances certain objects will be considered as a head like a helmet or any round object when shape based detection is being employed. These problems are very difficult to overcome and many methods are still under process to provide acceptable results.

11. The existing algorithms work only for a specific place for example the algorithm used for detection in a playground is not effective for the detection in a shopping mall. The algorithms are place specific and generalized methods which gives 100% accurate results for any place is yet to be developed.
12. The feature extraction methods which yield several parameters like the size, shape, colour of the object varies largely from one another. Suitable measures have to be taken to effectively differentiate the necessary parameters.

CONCLUSION

This paper presented the crowd density estimation methods which have provided the most satisfactory results. Marana and team has explained about the real time crowd estimation based on texture features and similarly the other authors have provided their ways of dealing the small and tight crowds. The characteristics obtained from the CNN model trained showed a strong capacity to count the crowd, and principles of GrC are used to gestate crowd segmentation issue at different granular levels and other such algorithms of estimation. Further, these methods can be revised to form new algorithms with multiple advantages and can be implemented for a specific application like monitoring the crowd in shopping malls, in uncontrolled environments like bus stands and railway stations thereby preventing congestions and provide comfort.

REFERENCES

[1] Rohit, Chauhan, Vandit, Santosh Kumar, and Sanjay Kumar Singh,2016 "Human count estimation in high density crowd images and videos." In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on, pp. 343-347.

[2] Han, Kang, Wanggen Wan, Haiyan Yao, and Li Hou.,2017 "Image Crowd Counting Using Convolutional Neural Network and Markov Random Field." arXiv preprint arXiv:1706.03686.

[3] Pu, Shiliang, Tao Song, Yuan Zhang, and Di Xie.,2017 "Estimation of crowd density in surveillance scenes based on deep convolutional neural network." Procedia Computer Science111: 154-159.

[4] Kok, Ven Jyn, and Chee Seng Chan.,2017 "Grcs: Granular computing-based crowd segmentation." IEEE transactions on cybernetics 47, no. 5: 1157-1168

[5] Senst, Tobias, Volker Eiselein, Ivo Keller, and Thomas Sikora.,2014 "Crowd analysis in non-static cameras using feature tracking and multi-person density." In Image Processing (ICIP), 2014 IEEE International Conference on, pp. 6041-6045.

[6] Song, Diping, Yu Qiao, and Alessandro Corbetta.,2017 "Depth driven people counting using deep region proposal network." In Information and Automation (ICIA), 2017 IEEE International Conference on, pp. 416-421.

[7] Pedrycz, Witold,ed.,2001 Granular computing: an emerging paradigm. Vol. 70. Springer Science & Business Media.

[8] Moravec, Hans.,1988 Mind children: The future of robot and human intelligence. Harvard University Press.

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

[10] Seiler, A., G. Evensen, J. A. Skjervheim, J. Hove, and J. Vabø.,2011 "Using the EnKF for history matching and uncertainty quantification of complex reservoir models." Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty: 247-271.

[11] Idrees, Haroon, Imran Saleemi, Cody Seibert, and Mubarak Shah.,2013 "Multi-source multi-scale counting in extremely dense crowd images." In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 2547-2554.

[12] Marana, A. NSAV, Sergio A. Velastin, L. da F. Costa, and R. A. Lotufo.,1998 "Automatic estimation of crowd density using texture." Safety Science 28, no. 3 : 165-175.

[13] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik.,2014 "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587.

[14] Long, Jonathan, Evan Shelhamer, and Trevor Darrell.,2015 "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.

[15] www.prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html accessed on 20-6-2016

[16] www.crcv.ucf.edu/data/crowd_counting.php accessed on 30-6-2016.

[17] Rodriguez, Mikel, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert.,2011 "Density-aware person detection and tracking in crowds." In Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2423-2430.

[18] Xu, Tianchun, Xiaohui Chen, Guo Wei, and Weidong Wang.,2016 "Crowd counting using accumulated HOG." In Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on, pp. 1877-1881.

[19] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.,2016 "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

[20] Vedaldi, Andrea, and Karel Lenc.,2015 "Matconvnet: Convolutional neural networks for matlab." In Proceedings of the 23rd ACM international conference on Multimedia, pp. 689-692. ACM,

[21] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.,2016 "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

[22] Jia, Yangqing, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell.,2014 "Caffe: Convolutional architecture for fast feature embedding." In Proceedings of the 22nd ACM international conference on Multimedia, pp. 675-678.

[23] Ojala, Timo, Matti Pietikainen, and Topi Maenpaa.,2002 "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence 24, no. 7: 971-987.

[24] Fradi, Hajer, and Jean-Luc Dugelay.,2013 "Crowd density map estimation based on feature tracks." In Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, pp. 040-045.

[25] Chan, Antoni B., Zhang-Sheng John Liang, and Nuno Vasconcelos.,2008 "Privacy preserving crowd monitoring: Counting people without people models or tracking." In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1-7.

[26] Dalal, Navneet, and Bill Triggs.,2005 "Histograms of oriented gradients for human detection." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893.

[27] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun.,2015 "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, pp. 91-99.

[28] Simonyan, Karen, and Andrew Zisserman.,2014 "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

[29] Marana, Aparecido Nilceu, Marcos Antonio Cavenaghi, Roberta Spolon Ulson, and F. L. Drumond.,2005 "Real-time crowd density estimation using images." In International Symposium on Visual Computing, pp. 355-362. Springer, Berlin, Heidelberg.

[30] Boominathan, Lokesh, Srinivas SS Kruthiventi, and R. Venkatesh Babu.,2016 "Crowdnet: A deep convolutional network for dense crowd counting." In Proceedings of the 2016 ACM on Multimedia Conference, pp. 640-644. ACM.

[31] Morerio, Pietro, Lucio Marcenaro, and Carlo S. Regazzoni.,2012 "People count estimation in small crowds." In Advanced video and signal-based surveillance (AVSS), 2012 IEEE Ninth International Conference on, pp. 476-480.

[32] Ma, Ruihua, Liyuan Li, Weimin Huang, and Qi Tian.,2004 "On pixel count based crowd density estimation for visual surveillance." In Cybernetics and Intelligent Systems, 2004 IEEE Conference on, vol. 1, pp. 170-173.

[33] Bouchafa, Samia, Didier Aubert, and Salah Bouzar.,1997 "Crowd motion estimation and motionless detection in subway corridors by image processing." In Intelligent Transportation System, 1997. ITSC'97., IEEE Conference on, pp. 332-337.

[34] Lowe, David G.,2004"Distinctive image features from scale-invariant keypoints." International journal of computer vision60, no. 2: 91-110.

[35] Dalal, Navneet, and Bill Triggs.,2005 "Histograms of oriented gradients for human detection." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893.

[36] Idrees, Haroon, Imran Saleemi, Cody Seibert, and Mubarak Shah.,2013 "Multi-source multi-scale counting in extremely dense crowd images." In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 2547-2554.

[37] Liang-Chieh, Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille.,2015 "Semantic image segmentation with deep convolutional nets and fully connected crfs." In International Conference on Learning Representations.

[38] Kilambi, Prahlad, Evan Ribnick, Ajay J. Joshi, Osama Masoud, and Nikolaos Papanikolopoulos.,2008 "Estimating pedestrian

counts in groups." Computer Vision and Image Understanding 110, no. 1: 43-59.

[39] Regazzoni, Carlo S., and Alessandra Tesei.,1996 "Distributed data fusion for real-time crowding estimation." Signal Processing53, no. 1: 47-63.

[40] Tesei, A., and C. S. Regazzoni.,1994 "Local density evaluation and tracking of multiple objects from complex image sequences." In Industrial Electronics, Control and Instrumentation, 1994. IECON'94., 20th International Conference on, vol. 2, pp. 744-748.

[41] Kohonen, Teuvo.,1990 "The self-organizing map." Proceedings of the IEEE 78, no. 9: 1464-1480.

[42] Haralick, Robert M.,1979 "Statistical and structural approaches to texture." Proceedings of the IEEE 67, no. 5: 786-804.

[43] Simonyan, Karen, and Andrew Zisserman.,2014 "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

[44] Ferrando, Silvia, Gianluca Gera, and Carlo Regazzoni.,2006 "Classification of unattended and stolen objects in video-surveillance system." In Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on, pp. 21-21.

[45] Serra, Jean.,1983 Image analysis and mathematical morphology. Academic Press, Inc.

[46] Collins, Robert T., Alan J. Lipton, Hironobu Fujiyoshi, and Takeo Kanade.,2001 "Algorithms for cooperative multisensor surveillance." Proceedings of the IEEE 89, no. 10: 1456-1477.

[47] Lv, Fengjun, Tao Zhao, and Ramakant Nevatia.,2002 "Self-calibration of a camera from video of a walking human." In Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 1, pp. 562-567.

[48] Li, Liyuan, Weimin Huang, Irene YH Gu, and Qi Tian.,2003 "Foreground object detection from videos containing complex background." In Proceedings of the eleventh ACM international conference on Multimedia, pp. 2-10. ACM.

[49] Fleet, David J., and Allan D. Jepson.,1990 "Computation of component image velocity from local phase information." International journal of computer vision 5, no. 1: 77-104.

[50] Horn, Berthold KP, and Brian G. Schunck.,1981 "Determining optical flow." Artificial intelligence 17, no. 1-3: 185-203.

[51] Albiol, Antonio, Maria Julia Silla, Alberto Albiol, and Jose Manuel Mossi.,2009 "Video analysis using corner motion statistics." In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 31-38.

[52] Lijun, Cao, and Huang Kaiqi.,2013 "Video-based crowd density estimation and prediction system for wide-area surveillance." China Communications 10, no. 5: 79-88.

[53] Snidaro, Lauro, Christian Micheloni, and Cristian Chiavedale.,2005 "Video security for ambient intelligence." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans35, no. 1: 133-144.

[54] Zhang, Xiaowei, and Graham Sexton.,1997 "Automatic human head location for pedestrian counting." 535-540.

[55] Chen, Chao-Ho, Yin-Chan Chang, Tsong-Yi Chen, and Da-Jinn Wang.,2008 "People counting system for getting in/out of a bus based on video processing." In Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on, vol. 3, pp. 565-569.

[56] Hou, Ya-Li, and Grantham KH Pang.,2011 "People counting and human detection in a challenging situation." IEEE transactions on systems, man, and cybernetics-part a: systems and humans 41, no. 1: 24-33.

[57] Surekha, Borra, Kanchan Jayant Nazare, S. Viswanadha Raju, and Nilanjan Dey. "Attendance recording system using partial face recognition algorithm." In *Intelligent techniques in signal processing for multimedia security*, pp. 293-319. Springer, Cham, 2017.

[58] Jayant, Nazare Kanchan, and Surekha Borra. "Attendance management system using hybrid face recognition techniques." In *Advances in signal processing (CASP), Conference on*, pp. 412-417. IEEE, 2016.