International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 11, Issue 5, May 2022 ISSN 2319 – 4847

Exploration on Crowd Detection Techniques

Mrs. Ashvini Ladekar[1], Kushal Choudhary[2], Darshan Phaldesai[3], Saurabh Agrawal[4], Mihir Gadhe[5]

[1]Assisstant Professor, IT dept PCCoE, Pune

[2]Third Year Student, IT dept PCCoE, Pune

[3]Third Year Student, IT dept PCCoE, Pune

[4]Third Year Student, IT dept PCCoE, Pune

[5]Third Year Student, IT dept PCCoE, Pune

ABSTRACT

Due to an ever-increasing population, crowd density estimation is an extremely significant factor that must be considered in today's times, and people counting is a critical subject in video surveillance applications. When a large group of people gathers, there is a risk of loss of life, property, and other things. Human action identification, crowd anomaly detection, and behavior analysis are some of the most popular disciplines of video processing study. This is where crowd density estimate, which is based on image and video processing, can help organizers of public events, railway security, college campuses, and other places maintain track of crowd density. Various image processing and video processing techniques are employed in this project effort to estimate the number of persons from a given video footage or image.

Keywords: Crowd detection, deep learning, image processing, video processing.

1. INTRODUCTION

When the entire population of a certain area exceeds the capacity, the area gets congested. Various accidents may occur as a result of the throng. Extreme crowds cause people to lose control and transform the location into a tragedy. Miscreants frequently use such crowds to carry out inhumane acts such as pestering women. There have been numerous major tragedies as a result of overcrowding.

The demand for crowd analysis systems for controlling dense crowds is increasing in order to minimize potentially dangerous situations caused by crowds. This entails determining the total number of persons in the crowd as well as the density of the crowd in various regions of the area. Certain possible accidents may be readily averted by providing advance notice if a certain area's population density exceeded the safe limit. This could also help with the area's overall administration and infrastructure.

Many deep learning (DL) studies have yielded positive results, focusing largely on three data types: pictures, audio, and text. Image classification, audio recognition, regression problems, pattern recognition, and text sentiment classification are all common uses of these data fields. The process is made more difficult by factors like as strong occlusions and scene perspective distortions in real-time applications. The use of infrared (IR) sensors and the WIFI network's Channel State Information (CSI), which are the traditional ways, provide the count, but they have their own range limits and limited applicability to regulated environments. Video surveillance systems are one of the most modern methods for estimating the density of people in a given area for security and obtaining human statistics. When people are moving and a high-resolution image with a clear background is available, vision-based approaches function well.

Despite having rich data, video data is difficult to evaluate and handle due to its high file sizes and complexity. After several AI algorithms were created for Image processing for various purposes, particularly in the last ten years, research on video processing using AI gained interest. Video data is a popular choice among users of many platforms such as Twitter, YouTube, Facebook, and others, and it is also the fastest-growing data format today.

2. MOTIVATION

In subsequent surveys, we discovered that a missing component is a combined review of several up-to-date image processing functionalities. Anomaly detection, anomalous human activity recognition, multi-object tracking, and behaviour analysis are only a few of the features covered in the aforementioned surveys. None of the polls combine the results of different types of research into a single survey.

As a result, we are motivated to discuss recent advances in image and video processing methods for a variety of functions, including motion detection, object detection, human action recognition, object tracking, image classification, and so on, as well as deep learning techniques for performing these functions.

3. RELATED WORK

3.1 Image detection using SVM and Motion Estimation

3.1.1 Using SVM and ANN

The output is typically a vector or multi-vector after image processing, picture segmentation, and feature extraction. There is a lot of depiction space and subspace. The component vector for each sub-space would be eliminated. The ANN receives this feature vector as input.

SVM is used to determine the ideal weight. The support vector mechanism must be trained first, and the SVM parameter must be changed to fit the training data and the problem. The support vector mechanism brings together all of the artificial neutral networks that have been classified. The study presents a process for detailed classification that takes less time to install and process [1].

3.1.1 Motion Estimation and Motionless Detection Method

Samia Bouchafa et al. define the issues and solutions for motion estimation, as well as a motionless detection approach to deal with three challenges: real-time constraints, deformable objects, and occlusion. This method employs an optical technology that is well suited to crowd surveillance. Motion detection and estimate are based on block matching technology. As indicated below, this strategy employs three techniques:

• Matching techniques: This method separates the image into blocks and compares them using a similitude criterion between two consecutive images. "Add the differences in absolute value" was used to select a similitude function. The block's size yields good results at the expense of sensitivity. As a result, they picked a smaller block size to reduce computing complexity. Only the edges of the pieces are matched. As a result, they were unable to reduce the computing time required to prepare photographs for online viewing.

• Frequential techniques: This method is suitable for obtaining the two motion components from spatial time related surfaces with a constant phase. A set of Gabor filters is convolved with the image, and the displacement vectors are obtained using this method. The downside is that it takes a long time.

• Differential techniques: This approach is based on the premise that a moving point's brightness remains constant throughout time. The suggested system uses a modified Horn and Schunk approach, in which the velocity vectors change only little between successive photos. These data are used to derive a global crowd motion direction. Two filters, spatial filtering and temporal filtering, are employed to process the segmentation in an appropriate context.

3.2 Image Processing using Deep Convolutional Network for Crowd Detection and Parallel Virtual Machine

3.2.1 Using Deep Convolutional Network

In large-scale image identification, object recognition, and segmentation, deep convolutional neural networks (ConvNet) are commonly employed. Deep ConvNets estimate population densities by extracting picture features directly and mapping them to crowd density on three levels: low, medium, and high. The overall number of people in each range, as well as the number of ranges, may vary depending on the application and the field's specific features [2].

The steps involved in this method are as follows:

• With a linear filter, convolutional layers are convolved with input image or feature maps. The distinctive maps that result

depict each filter's reaction.

• Down-sampling layers with maximum (or average) values obtained in every sub-region of the input image or feature maps are non-linear with pooling layers. The translation efficiency is improved, and the number of network parameters is lowered.

• Activation layers apply non-linear activations to neurons that serve as inputs. Sigmoid function, hyperbolic tangent function, and other activations are common.

• Fully-connected layers compute outputs by linking them to every characteristic map element of the previous layer.

The neural network is trained using frame samples from the train subset, which are divided into five categories based on the amount of people in the image: Very-low, Low, Medium, High, and Very-high. The output of the neural network is separated into 5 levels of crowd density by quantifying the estimated crowd density. The test subset's classification accuracy is used to assess performance. A new crowd dataset of subway scenes with over 160K photos is utilised to evaluate the accuracy of the crowd density estimate approach in this deep ConvNets method. This method's experimental findings show that it has the best accuracy of 91.73% on average, and that it can perform better in real applications [2].

3.2.2 Using Parallel Virtual Machine

This technique uses crowd image textures to estimate the density of the crowd in real time. In this method, input photos are divided into population density classes. Following that, the categorization is handled using a low pass filtering algorithm based on the prior images from the incoming image sequences. The initial step in this method is to classify each pixel extracted from a sequence of photos into one of the previously identified texture classes. The approach of Self Organized Maps is used in this process of identifying each pixel (SOM). The method is expanded using a distributed algorithm called Parallel Virtual Machine because the classification of various pixels takes a long time in a real-time context (PVM) [3].

The steps involved in this algorithm are:

• The master processer divides the input image into n fragments at first (where n refers to the count of slave nodes).

• The slave processor is then given each of the fragmented image fragments.

• The slave processer's job is to use a sequential method to classify the texture of image fragments.

• In addition, the slave processor delivers the assigned parts to the master processor.

• Finally, the master assembles all of the fragments into a single texture-segmented image.

3.3 Video Processing using Deep Learning Methods

3.3.1 Using CNN Approach

In a previous work, a CNN model was trained using large-scale videos including near to 500 sports lessons. This approach includes a architecture that doesn’t uses permanent motion information from video. It also has modules for categorizing movies that comprise a context stream for low-resolution image modelling and a fovea stream (for high-resolution image processing) [4]. The author employs an ImageNet-based pre-trained model to categorise unusual occurrences captured by the security camera. The cost of training a large CNN model for video processing is reduced using this method.

Simonyan and Zisserman suggested 2-stream model as a universal deep learning approach using 2 CNN streams. Two-stream is designed with two separate layers, one for recording spatial information using a single frame and the other for storing temporal information using optical flow. In a two-stream CNN, regular images and optical flow images are mixed as input. To achieve high throughput, these two different networks were connected utilising a late fusion approach.

DeepSORT is a widely used CNN-based elegant object tracking technique. The author used the Basic Online and Realtime Tracking strategy for multiple item tracking, which focuses on simple, realistic methods. The author employed a single conventional hypothesis tracking approach that included recursive Kalman filtering1 and frame-by-frame data association.

As a result of CNNs' significant success in image recognition tasks, the authors used gait energy image, a popular picture-based gait representation. Data from the GEI was transmitted into the GEINet. GEINet performs brilliantly on the OU-ISIR large population dataset. For vehicle detection, the authors presented a nine-layer CNN based on a popular video processing programme. Background subtraction, object detection, obstacle detection for self-driving cars, anomaly

detection in crowded settings, lane marking, and monitoring wild animals are among the applications for which CNN-based techniques have been proven to be a potential option [5].

3.3.2 Using DNN Approach

A deep neural network is a sort of neural network that contains more than two layers and is considered a more complicated form of neural network. Large volumes of higher-dimensional data can be handled using DNN-based video processing techniques. A robust deep neural network-based Multivariate Gaussian Fully Convolution Adversarial Autoencoder (MGFC-AAE) model was presented to address the requirement for video anomaly detection and localization.

Pashchenko et al. used a DNN-based transportation system model to recognize the urgent case. Amosov et al. developed a DNN-based video processing approach for determining the classification probabilities for each video fragment, as well as detecting and recognizing normal and abnormal situations. Similarly, road sign identification and lane detection activities are necessary for road analysis in automated driving. To detect abnormalities, Luo et al. applied DNN-based video processing in addition to sparse coding. The purpose of this strategy is to develop a vocabulary that can be used to evaluate a wide range of common occurrences with few reconstruction flaws [5].

3.3.3 Using Hybrid Approach

The hybrid approach exhibits the use of a range of deep learning algorithms for video processing. This strategy has been utilized in several studies. A constant gradient flow in an LSTM makes it easier to back-propagate than a standard recurrent neural network. Because it prevents gradients from popping or vanishing, LSTM is also more stable. The purpose is to partition the video data into equal parts, extract short snippets from each component of the movie, classify each and every part using a multi-stream CNN network, and then agree to get a score for the full video [5].

For video processing on real-time yoga posture detection using deep learning, a hybrid deep learning model utilizing CNN is recommended. In this way, the CNN algorithm gathers many posture characteristics before using LSTM features to create real temporal predictions. A novel transformer network uses the attention mechanism in deep learning to outperform when combined with spatio-temporal based models like CNN for human activity identification. The linked network works on attention processes independently, with a focus on hands and faces, which are critical for accurate human action recognition tasks [6].

Because it illustrates the intelligent monitoring of equipment used during surgery in the operating room, healthcare is a particularly demanding use of the video process.

3.4 Detecting Crowd Behavior with Video Processing

Video processing is a 3 step process. The first step of Feature Detection and temporal filtering. In the step, the characterizing visual aspects of crowd flow are observed.

Multiple different parameters are considered when features of characterized. for example:

1) Brightness
2) Small motion
3) Spatial coherence ,i.e., neighboring points move in the same direction [9].

Feature trackers are use to provide a "feature description" of the object. This description once extracted from provided sample image can be used to separate said object from other objects and track its position and velocities. For the recognition to be reliable, the detection should be able to perform well in any changes in scale, noise and illumination.

3.4.1 Scale Invariant feature tracker (SIFT)

SIFT is a robust tracker and is able to detect objects even among clutter and partial occlusion. It is invariant to uniform scaling, orientation, illumination changes and partially invariant to affine distortion. First, the image is transformed into a large collection of feature vectors. Low-contrast candidate points and edge response points are rejected. This helps for stable matching and recognition. Next step includes indexing and storing SIFT kets for matching against new images. After this, clusters are identified based on consistent interpretation of each feature in the feature set and final features are extracted after weighing features against their sample scale. This tracker also includes a verification step [10].

3.4.2 KLT feature tracker

Kanade-Lucas-Tomasi feature tracker is one of the many feature descriptors like SIFT or SURF used for identifying and tracking local visual features. KLT performs well at minimal image distance, resists tweaking parameters, and requires little computer power.This method uses gradient weighted approximation applied over translation of objects to perform feature tracking. It also includes an additional stage of verification so that features are tracked correctly.

The tracker produces a collection of frames along with a feature points in adjacent frames. A vector of the said features is stored in memory along with their velocities. The motion points are weighted using a probability grid.

The second step is image segmentation and blob extraction. Using binary masks, the RGB frames are filtered using a 3x3 kernel gaussian blur filter and then passed through multiple segmentation techniques which aggregate homogeneous regions globally. This results in the development of a grayscale intensity edges. Then this edge image is passed to an algorithm with vectorizes it and then a triangular tessellation is generated over it.

The last step is Crowd behavior detection. The data collected undergoes statistical analysis to detect crowd activity in the observed scene. A 3D-Grid builds a grayscale activity map taking a set of binary blob images as input.

Using this 3D map, number of objects can be enumerated and characteristics defined. This way crowds can be identified [9].

4. PROPOSED WORK

Crowd Identification and Tracking systems mostly include a single optic device/ camera to capture the images required. While this method is cost effective and can be used map large areas, this method is not able to monitor and track objects in 3D space. The proposed system is combination and modification to many simple 2D tracking systems used together to identify objects in 3D space for better accuracy and positional insight. Input from multiple cameras can be fed to the algorithm to generate a realtime 3D map of an area and then objects can identified in the said map. This should provide us with more accurate distance tracking in said objects.

5. CONCLUSIONS AND FURTHER SCOPE

The approaches for estimating crowd density that have produced the best results were presented in this paper.

The features gained from the CNN model training shown a good capacity to count the crowd, and GrC principles are utilised to gestate crowd segmentation issues at various granular levels, as well as other estimate algorithms. Furthermore, these methods can be altered to create new algorithms with many benefits that can be deployed for a specific application such as crowd monitoring in shopping malls, in uncontrolled situations such as bus stops and train stations, preventing congestion and providing comfort [7].

References

[1] 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

[2] Pu, Shiliang, Tao Song, Yuan Zhang, and Di Xie.,2017 "Estimation of crowd density in surveillance scenes based on deep convolutional neural network."

[3] Sabrina Haque, Muhammad Sheikh Sadi, Md. Milon Islam, Md Erfanul Haque Rafi, "Real – Time Crowd Detection to Prevent Stampede", DOI: 10.1007/978-981-13-7564-4_56.

[4] Sneha.P.K, Rabichith, Sri Nithya S, Surekha Borra Department of ECE, K.S. Institute of Technology, Bangalore-560109, India., "Crowd Density Estimation Using Image Processing", ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 6855-6864.

[5] Vijeta Sharma (Member, IEEE), Manjari Gupta, Ajai Kumar and Deepti Mishra (Senior Member, IEEE), "Video Processing Using Deep Learning Techniques: A Systematic Literature Review", Digital Object Identifier - 10.1109/ACCESS.2021.3118541.

[6] K Ragavan, K Venkatalakshmi, K Vijayalakshmi, "A case study of key frame extraction in Video Processing", ISSN: 2566 -932X, Issue 4, July 2020.

[7] D. Mohanapriya, K.Rajalakshmi, N.Geetha, N.Gayathri and Dr.K.Mahesh, "Video Processing - Challenges and Future Research", Ph.D Research Scholar, Department of Computer Applications, Alagappa University, Karaikudi, UGC Approval no.:40934, CASS-ISSN:2581-6403.

[8] Neetish Kumar, Dr. Deepa Raj, "Video Processing and its Applications", Department of Computer Science, BBA University, Lucknow, India, ISSN 2278-6856, Volume 6, Issue 4, July - August 2017.

[9] Xie, S., Zhang, X. & Cai, J. Video crowd detection and abnormal behavior model detection based on machine learning method. Neural Computer & Applications 31, 175–184 (2019) https://doi.org/10.

[10] Choudhary, Shivali; Ojha, Nitish; Singh, Vrijendra (2017). [IEEE 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) - Madurai, India (2017.6.15-2017.6.16)] 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) - Real-time crowd behavior detection using SIFT feature extraction technique in video sequences, (), 936–940, doi:10.1109/ICCONS.2017.8250602