WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Visual Head Counting and Verification Using Yolo and CNN Based Event Classification in Crowded Surveillance Environment

[1]S.P. Samyktha, [2]A. Asiya Mariyam, [3]S. Agnus, [4]B. Narmadha

[1]UG Scholar, Department of Artificial Intelligence & Data Science, Panimalar Engineering College, India
[2]UG Scholar, Department of Artificial Intelligence & Data Science, Panimalar Engineering College, India
[3]UG Scholar, Department of Artificial Intelligence & Data Science, Panimalar Engineering College, India
[4]Assistant Professor, Department of Artificial Intelligence & Data Science, Panimalar Engineering College, India
spsamyu516@gmail.com, asiyamariyam1618@gmail.com, agnus040204@gmail.com, narmadhab16@gmail.com

ABSTRACT

A convolutional neural network (CNN)-based individuals counter that categorizes a given image is proposed in this paper. A framed cube is classified to a particular occasion that shows individuals entering or leaving a specific area to assess people instantly count. A training program has been developed for the projected CNN, the cube of the input frame and its associated class label that depicts a particular event are produced using the suggested guidelines for counting. To lessen the issue of overfitting that may happen during the proposed CNN's training, data using foreground distribution, augment and post-class correction with event probability are applied. The experimental findings show that, despite computing the cumulative count by adding instantaneous people counts, whereas the benchmark methods had been optimized for doing so, the proposed method outperformed the benchmark methods in terms of F1 score and accuracy for the cumulative individuals counting outcomes by up to 9.0% and 15% respectively.

Additionally a verification procedure with the help of face recognition is employed. This compares with the database for similarities from the feature extracted images using CNN and reports for verification procedure. This will identify people's identity to avoid fake entry or exit in an event. This system can be mainly used in marketing prediction and also for event to avoid fake or overcrowd entry.

Keywords: Convolutional Neural Network (CNN), Person counter, framed cube, F1 score, accuracy

INTRODUCTION

In recent years, a variety of applications have used metrics that are based on vision extensively. Among them, a vision-based people counter is used frequently in a variety of applications, including video surveillance, urban design, administration of resources, and customer profiling. It calculates the number of persons entering and departing a particular region for a given image. For retail stores or restaurants in particular, the information on the number of people is quite helpful because it can be used to analyze client visit trends, which could enable effective management. Additionally, if data on the number of people at various retail establishments in a certain area is available, it can be used specifically for regional commercial analysis.

Consequently, a variety of vision-based individuals counters have been created. Most people counters count the number of people travelling through a target area using a line of interest (LOI) as well as region of interest (ROI) method. These methods for counting persons can be divided into two groups: segmentation using regression-based methods and detection with tracking-based methods. The persons count is extracted using tracking-based algorithms that identify people's locations and follow them. People positions have been extracted and tracked using unsupervised Bayesian clustering to detect individual entities.

The two-dimensional association of a bank of annular patterns and the retrieved foreground regions is used to identify heads [2]. Then, a Kalman filter is employed to monitor heads that have been spotted, and a LOI-based method is utilized to estimate the number of persons. However, in contexts where several persons overlap in the image or come and disappear, these tracking-based approaches for detection are quite susceptible to variations regarding illumination and occlusion. As a result, in such circumstances, the precision of people counting can considerably decline.

Foreground and mobility data are typically used in segmentation with regression techniques for crowd segmentation and taking people away counts. On the selected LOI, vectors (MVs) are retrieved via an optical flow-based method [7][8]. A temporal slice image (TSI) is then displayed. is produced via time-lapse stacking of LOI pixels, the shifting of blob is created by stacking segments, and each segment's size segment stacking is based on segment velocity, which can determined using the MV's size.

Regression is used to determine the total number of viewers for a specific video. The function using the captured blob's volume depicts a crowd [1]. Homogenous motions are divided into smaller parts utilizing a combination of motion models for dynamic textures. After that, each segment is used to extract a low-level feature. The quantity of persons are computed from the Bayesian Poisson Regression (BPR) employing the features extracted. A TSI is produced in by stacking LOIs over time with a predetermined line width for stacking, although FM considers a moving segment's speed as the line stacking's size. Then, using the TSI, a regression function is used to estimate the number of persons in a collection of overlapped sliding windows known as the temporal ROI (TROI). Using BPR and a local histogram of gradients with orientation and different characteristics, the number of people within each TROI is approximated. Utilizing a multiple-window-length TROI and an outlier-robust goal function, this method is enhanced.

For the evaluation of multiple individual counters, which were recorded in an environment during which the distance within the people along with camera is sufficiently large, those segmentation alongside regression-based methods successfully calculate the total number of people in the existing datasets. Although the movements of an arriving or departing person are typically significantly larger than those in the available datasets, these methods are disadvantageous for inside contexts like a tiny retail establishment where the camera and individuals are near each other. Additionally, brightness fluctuation and occlusion are common in these situations. Consequently, there may be a major decline in the accuracy regarding the present-day counters.

In our earlier studies, the flow volume analysis-based persons counter (FAPC), which can take into account the real-world surroundings of the people counting problem, was offered as a solution to these issues. To increase the accuracy of people counting in actual retail stores, our approach presented the multiple-touching section-based foreground evaluation and the maximum a-posteriori probability-based dilated motion estimate. Even if the FACP might deliver more accurate people counting than the current techniques, its effectiveness can still be enhanced. Additionally, because it concentrates on determining the cumulative estimate for each input sequence, the FACP was unable to produce a precise instantaneous population count. The FACP specifically used a multivariate linear regression algorithm on an aggregated counts for each test set, providing results that were typically less accurate than instantaneous counts but more accurate than cumulative counts.

Here is a people counter that uses convolutional neural networks (CNN) and can deliver the exceptional accuracy of immediate population compared to the current population counters are used in an atmosphere where individuals are counted inside. The proposed approach aims to derive persons with the input frames obtained from a wide number of cameras, count of retail establishments in a constrained space; as a result, it can be utilized for commercial analysis of the region [3]. A surveillance camera is mounted on the ceiling of a doorway over the heads of customers, and it is assumed that the input frame rate is extremely low (below 5 frames per minute), taking into account the input frame acquisition and the emission circumstances of a practical retail shop. Consequently, compared to the current datasets, the input photos would often have a narrower field of view and more pronounced human gestures.

EXISITING METHODS

[1] In sensing applications, population counts is essential. People detection and counting have been widely used employing IR-UWB radars, due to its excellent penetration and high-range resolution. The elimination of the signal's direct current (DC) element, bandpass filtering, and clutter signal elimination are the three fundamental processes in current signal processing techniques that use IR-UWB radar [1]. To choose effective peaks for counting, a manually determined environment-dependent threshold is constructed. However, the methods used to get clearer signals could also get rid of important information. A unique strategy utilizing convolutional neural networks, or CNNs, is suggested in this research.

[2] In this paper, we propose a novel convolutional neural network (CNN)-based approach for counting persons utilizing an IR-UWB radar with low radiation impulse. It is difficult to handle the individuals calculating task by simply identifying targets for each range bin due to the perpetually shifting signals resulting from a variety of human motion dimensions, juxtaposition and obstruction of signals, as well as the attenuation of signal's vitality along the distance and the angle. Thus, in order to carry out the counting task, we hope to uncover information about targets' patterns in the detecting zone, particularly their densities and types of pattern distributions.

[3] In many sensing applications, like smart cities and commercial malls, people counting is essential. In this study, we present a data-driven approach to count the number of randomly walking persons in an interior environment using a low-powered ultra-wideband impulse (UWB) radar. Clear clutter signals via UWB radar are processed using a pre-processing signal analysis technique. We looked into advanced convolutional neural networks (CNNs), which dynamically acquire the data to determine the total number of people in an indoor space, as opposed to the conventional counting approaches, which manually extracted features and learned from useful data patterns.

[4] In order to conduct video surveillance and provide anomaly alerts, it is crucial to count people in highly thick crowds. The paucity of training tests, severe occlusions, complex scenes, and variance in perspective make the challenge more difficult. Existing techniques either rely on supplementary face and human detectors or act as a stand-in by calculating crowd densities. The majority of them, like SIFT, HOG, and other hand-crafted features, are vulnerable to failure when training samples are insufficient or density increases. In this article, we suggest a powerful convolutional neural network (CNN) regression framework for counting individuals in photos of incredibly dense crowds.

[5] One of the most crucial components of artificial intelligence today and one that is actively growing is computer vision. Numerous works exist in this field. The applications for computer vision are numerous. For instance, it can be used in the medical field to diagnose an X-ray or magnetic resonance imaging image, in security systems to find intruders, in autonomous cars or robotic space navigation, etc [19]. However, object detection and picture recognition frequently coexist [5]. Typically, just a small portion of the original image is needed to recognize an object, whereas the remainder of the

image contains no helpful information. For this reason, we must locate this item before recognizing it in order to reduce calculation time. Numerous techniques and technologies exist.

[6] In many applications, counting people is a standard and fundamental computer activity. The majority of people counting approaches use image processing techniques and a sensing device, usually a camera, to follow pedestrians. However, there are numerous security and privacy concerns when tracking individuals who carry cameras in public spaces. Another interesting option is the passive infrared sensor (PIR), which can measure body temperature using infrared light. The signals of one PIR are insufficient to recognize the complicated situations of several persons, despite the fact that a single PIR can readily detect different conditions of one individual.

[7] We show that motion alone includes more information than has previously been realized for the general case, even while crowds of different subjects may provide application-specific clues to detect people. The main objective of the unsupervised Bayesian clustering technique described in this research is the discovery of individual entities. Simple picture features are tracked and probabilistically grouped into clusters that correspond to independently moving objects. Without the use of supervised learning or a subject-specific model, the number of clusters and the arrangement of the constituent features are decided. In the novel method, the sole probabilistic criteria for grouping are space-time closeness and trajectory coherence throughout picture space.

[8] The algorithmic implementation for a low-level head-detection method-based FPGA-based architecture for people counting is shown in this work. On the same chip, capturing and online processing are possible due to the hardware (HW) deployment of an FPGA. Various annular patterns are employed to process the image in parallel and find heads of various sizes. Utilizing reconfigurable hardware, preprocessing & extraction of edges are also completed [17][18]. The vision algorithms has been changed and adjusted for HW implementation, allowing the created system to take use of HW processing. It requires the smallest amount of Spartan3 space (1.5 M gates) and, while utilizing extremely affordable circuits, achieves real-time performance that is on par with that of more complex algorithms.

[9] In order to understand the patterns of collective behavior of pedestrians in crowded environments, a new Mixture framework for MDA is put forward in this study. Crowd dynamics as a whole are characterized by collective behaviors. The dynamic pedestrian-agent, that is a linearly dynamic system with preliminary and completion phases expressing a pedestrian's perception of the starting place and the destination, is what propels every person walking in the crowd according to the agent-based modeling [9]. Then, a collection of dynamic pedestrian-agents is used to model the entire crowd. MDA may replicate the crowd behaviors after the model has been unsupervised learned from actual data.

[10] Both security applications and people management applications can benefit greatly from real-time information on people flow. This study describes a four-module counting system that includes background extraction, head-shoulder component recognition, tracking, and trajectory analysis. First, a flexible components number selection technique for a combination of Gaussians model is given to lower computation costs and handle various difficult surveillance scenarios for foreground extraction. Second, because the head-shoulder region is less variable and less likely to be obscured from a downward-sloping view, pedestrians are identified by this region. Thirdly, the Kalman filter approaches and cost function are used to track each pedestrian over a series of frames in order to count individuals arriving or leaving the scene, the resulting trajectories are lastly examined.

PROPOSED METHODOLOGY

In order to determine the population of a particular region in an event like situation, the suggested method classifies people as they enter or leave using the CNN architecture. First, the procedures for creating training data and training CNN are described. The suggested CNN's architecture is then described. Further using face recognition the registered people can be recognized for an event for verification purpose[5][8]. This will identify people’s identity to avoid fake entry or exit in an event. This system can be mainly used in marketing prediction and also for event to avoid fake or overcrowd entry [20]. Further the incorporation of this system in the CCTV will result in efficient people count in a crowed environment using YOLO and CNN and need verification among them using a simple face detection technique at necessary situations.

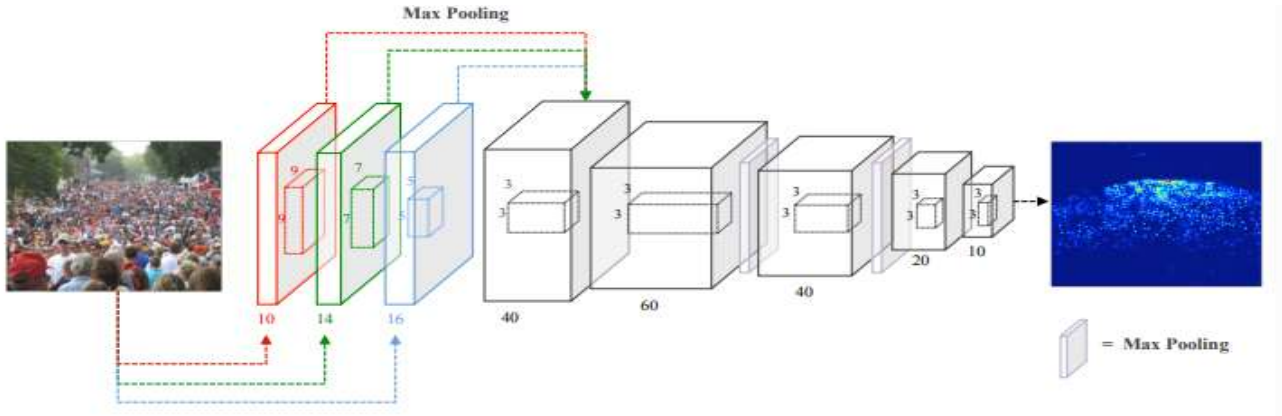


Fig.1. Overview of the proposed C-CNN architecture. The network incorporates three parallel filters of different size and colors which are merged to estimate crowd density estimation.

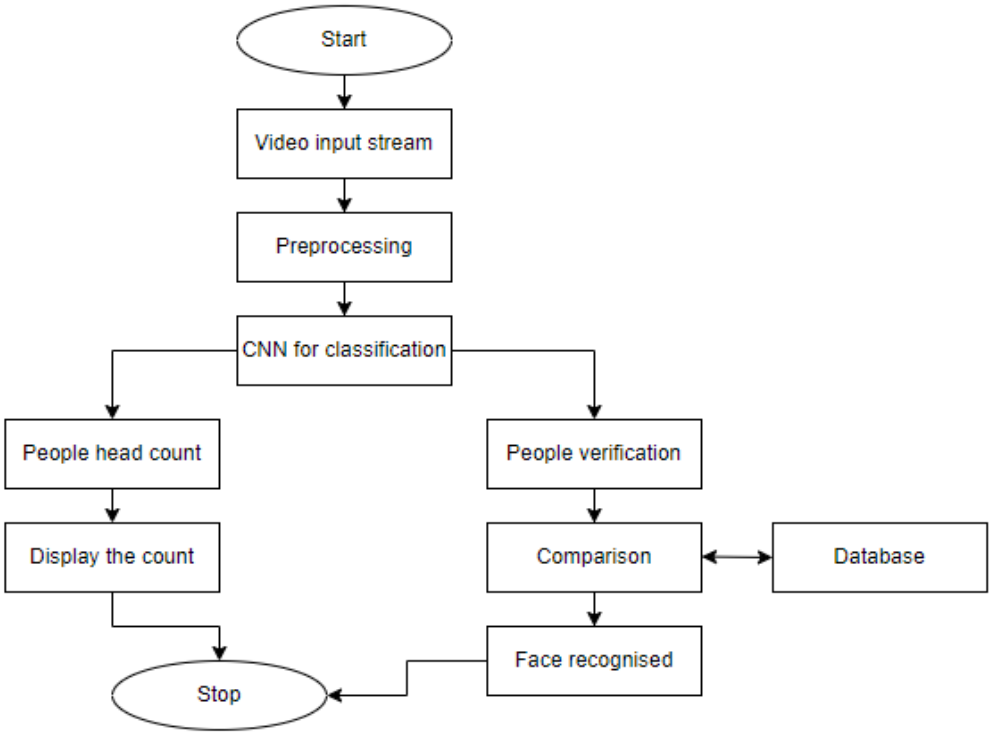


Fig.2. Architecture diagram for the proposed system

Head count:

To develop a quick cascade head detection featuring the histograms of oriented gradients (HOG) feature, we use the cascade Adaboost method in the initial off-line training step. Using a fresh dataset created from the detection outcomes of employing the cascade head detector to the original training dataset, we train a CNN model in the subsequent off-line training stage [9]. The next stage employs this CNN model as a characteristic 5 extractor. We develop a Support Vector Machine (SVM) classifier with head classification in the second stage by using the CNN feature extraction method on the dataset. We employ the learned SVM classifier to further identify the head after applying the cascading head detection system to the test picture to provide head recommendations. After post-processing the recognition findings, we total the results as the number of persons.

HEAD PROPOSAL DETECTION:

We must first provide CNN the object candidates in order to use CNN's significant advantage on classification jobs to solve the object detection problem. Adopting pre-existing region suggestion techniques would result in two issues for practical surveillance applications. First, the procedures for area proposals that are available off the shelf often take a lot of time, which would not meet the time requirement [15]. Second, these approaches often offer the top N possible object locations, with N typically not being less than 1000. The offered candidate locations are often significantly more than just actual items in most circumstances. However, because these algorithms only detect broad items and not individual objects, the majority of candidate areas are unrelated to relevant things. Thus, this tactic will significantly lengthen CNN's categorization time. In order to quickly find the candidate head regions, we use the cascade Adaboost approach in this research. As the feature representation, we choose for HOG [2[10]. Additionally, we examine features that are Harr-like and LBP and discover that the functionality of HOG does the task the best.

The HOG cascade was initially proposed in to greatly accelerate human detection. The primary concept is to integrate the HOG descriptors with the cascade classifiers technique in order to maximize their benefits. Unlike the original HOG descriptors, which used blocks of uniform size, this technique uses blocks that vary in size, placement, and aspect ratio. This approach employs the Adaboost algorithm to choose the right set of chunks to be involved in the cascade out of a vast number of alternative blocks. Although it worked up to 70 times quicker than the initially developed Dalal and Triggs algorithm, this technique can attain performance that is equivalent to it. In this study, we maintain a high recall rate by modifying associated variables of the cascade detector in order to make sure that the head area suggestions likely contain all genuine heads.

HEAD CLASSIFICATION BASED ON CNN:

True head areas and non-head regions, which together make up the dataset as displayed, are divided into two categories as the potential head regions generated by our cascade detector. The second component resembles actual heads to some extent, making it challenging to a cascade detector to identify their labels. Actually, the second section may be seen as a series of stark contrasts to the way things are now. The next objective is to determine how to accurately distinguish genuine head areas from those that are hard negative instances. We use CNN to do this assignment in light of its performance on classification problems. Instead of employing the CNN directly as a classifier, instead we simply utilize the CNN for feature extraction. A linear Support Vector Machine (SVM) classifier is then trained using the CNN features to do the classification. The fundamental tenet of this approach is that the SVM

classifier has strong classification performance when given data, and the CNN model can develop a very useful feature representation. As of now, the most popular method of collecting CNN features is to utilize a pre-built CNN that has been trained on a sizable dataset, like ImageNet. Actually, this is a common transfer learning approach and it can perform well if the test pictures are comparable to the CNN's original training dataset.

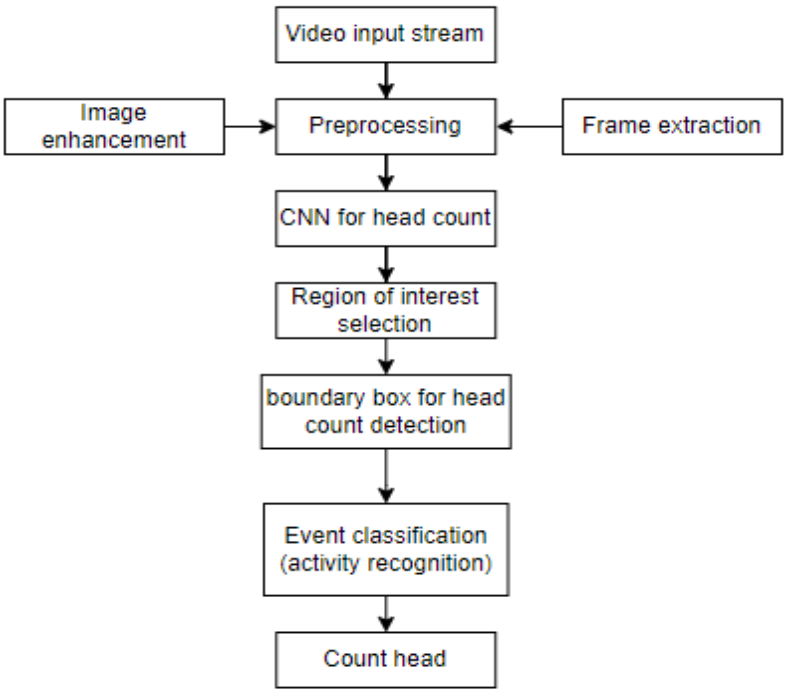


Fig.3. Workflow of the head count using CNN event classification

However, because the ImageNet dataset, which is often used to train CNNs, is distinct from our surveillance dataset, this transfer learning technique would fail to solve our problem. Thus, we must use the surveillance dataset for training a new CNN model. We need a strong CNN architecture to do this [4][15]. Unfortunately, there are two reasons why the widely utilized CNN architecture is inappropriate for our situation. It typically has a fairly deep structure on the one hand. For our small surveillance dataset, this would raise over-fitting issues, and this deep structure makes feature extraction time-consuming.

Face recognition:

The various stages of the suggested method's development are described in the framework's design. In this strategy, both bottom-level and top-level planning are included. The investigator will first gather the video evidence they have gathered so they may do facial detection on it. The cropped image then goes through the image enhancement procedure after the face detection approach. The resolution of the picture is much improved after this treatment [3][6]. The collected face will next have its features recognized and put through a comparison procedure with databases that have previously been saved in databases . Here, the footage video that was received directly from the CCTV is used as the input video. It goes through several steps, including face detection, pre-processing, facial feature extraction, and identification. Face detection involves extracting the face region from the input video, cropping the face from the extracted face region, and then using pre-processing algorithms to the cropped face picture to improve the image quality [12]. The face's features are then taken from the improved cropped image, and the recognition process compares them to the faces that have previously been saved in the database.

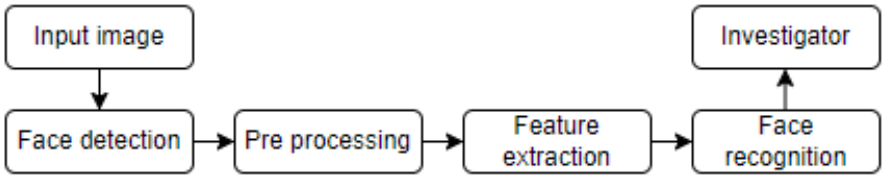


Fig.4. Overall flow of the face recognition for verification

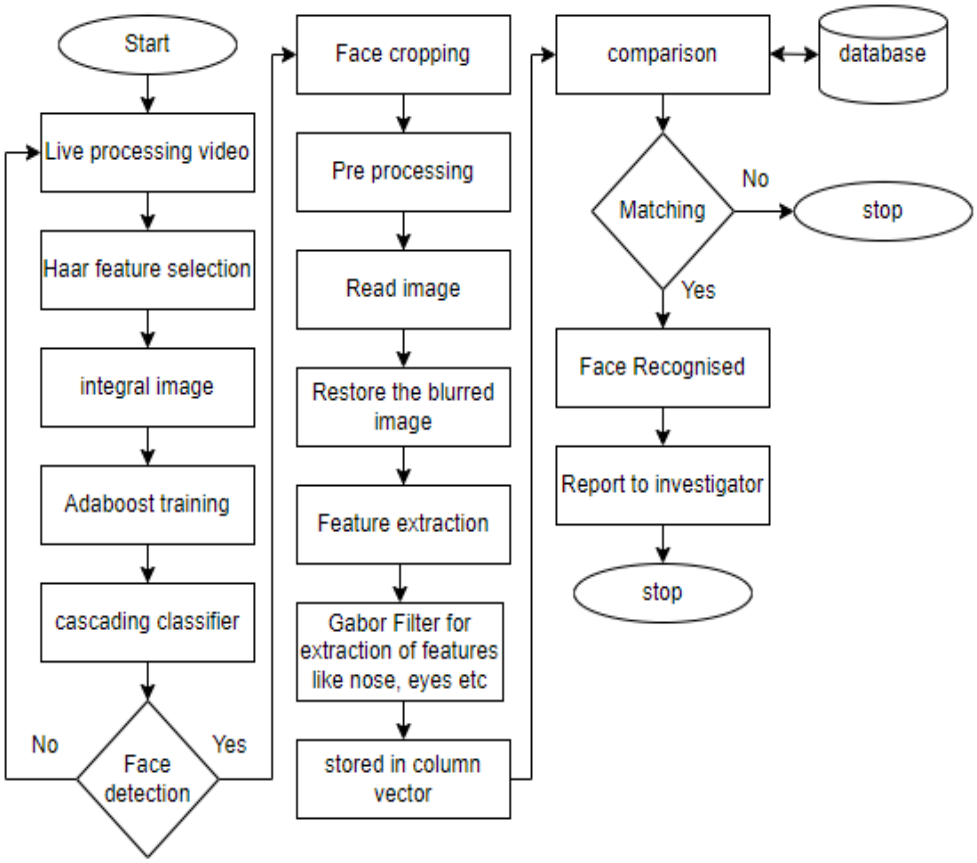


Fig.5. Internal flow of the face recognition for the verification

For optimal results, Face Recognition Algorithms always follow a set of organized processes. You can describe the stages as follows:

FACE DETECTION: For identifying faces in CCTV footage, the MATLAB computer vision system tool box is highly useful. The object detection mechanism and a unique detector are provided by the computer vision toolbox. The people detector makes it simple to identify a person's face. A trained support vector machine (SVM) classifier and Histogram of Oriented Gradients (HOG) features are two examples of histogram-based classifiers. The face detection method's output is always a cropped face picture.

PREPROCESSING: Noise reduction and distortion removal are applied to the cropped picture. In this stage, picture enhancement is accomplished using the wiener filter. By generating the de-blurred picture from the blurred photographs, this filter improves the quality of the image. The image processing toolbox in MATLAB is used to do this [1]. Workflow applications are offered by Image Processing Toolbox. This image processing toolkit may be used for 3D processing, picture segmentation, and noise reduction. The result in this case will be an image that has been upgraded in quality.

FEATURE EXTRACTION: Through this period, the numerous facial characteristics, such as the positioning of the eyes, nose, and ears, will be discernible. The geometrical extraction of features approach will be employed to conclude this stage. The MATLAB computer-vision toolbox is additionally employed for this stage [11][13]. The outcome of this particular step will be a collection of facial characteristics transformed from the improved input image.

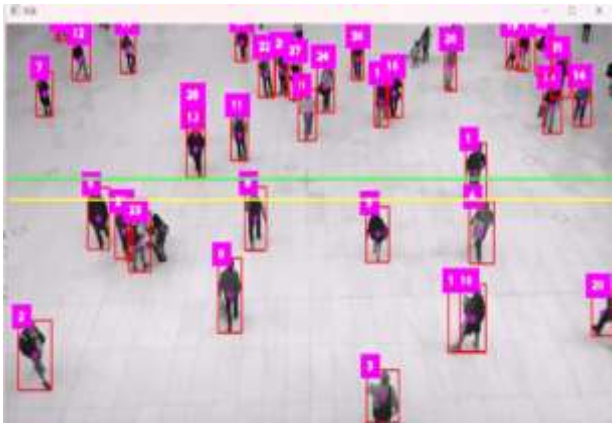
FACE RECOGNITION: At this level, the database correctly matched the face with the proper one. By comparing and matching the identified image with other images stored in the databases, the system will display the identification outcome.

DATABASE: The CCTV are used for this process which will store vast amount of data. This is the primary component for the entire process.

EXPERIMENTAL RESULTS

For head count:

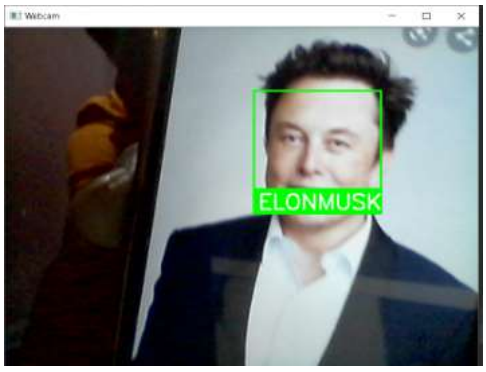
Input



Output

For face recognition:

Input



Output

CONCLUSION

In our proposed system, we suggest a crowded surveillance environment head detection-based people counting approach that combines the CNN and Adaboost algorithms. Our solution employs the cascading Adaboost algorithm to produce the head region suggestions rather than the generic object region proposal method. The subsequent categorization time might be cut in half using this method. By utilizing its powerful feature learning capabilities, we employ the CNN as an extractor of features rather than a classifier. The last classification assignment is then sent to a linear SVM that has been trained. Finally, we post-process the detection results to further reject false detection and the people count is obtained by counting the head detection results. Experimental results show that this method has good performance and outperforms the baseline methods of DPM, cascade Adaboost methods. Also the verification method involved in our system using face recognition will identify people's identity to avoid fake entry or exit in an event. This system can be mainly used in marketing prediction and also for event to avoid fake or overcrowd entry.

We will eventually expand on our existing approach to handle surveillance scenarios with several cameras. To enhance the performance of head detection, we can apply cutting-edge classifier approaches and head posture classification methods in particular. Additionally, we will incorporate tracking strategies into our methodology to manage the work of people-flow counts. For instance, tracking techniques may be used to record human motion data once we have the head location. Thus, we are able to count the movement of individuals in various directions.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all the fellow authors and professors for guiding and supporting us by providing the necessary resources and infrastructure during this research journey.

REFERENCES

[1] Zhang, Lu, Miaojing Shi, and Qiaobo Chen. "Crowd counting via scale-adaptive convolutional neural network." In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1113-1121. IEEE, 2018.

[2] Yan, Ran, Shengrong Gong, and Shan Zhong. "Crowd counting via scale-adaptive convolutional neural network in extremely dense crowd images." *International Journal of Computer Applications in Technology* 61, no. 4 (2019): 318-324.

[3] Bao, Runhan, and Zhaocheng Yang. "CNN-based regional people counting algorithm exploiting multi-scale range-time maps with an IR-UWB radar." *IEEE Sensors Journal* 21, no. 12 (2021): 13704-13713.

[4] Pham, C-T., V. S. Luong, D-K. Nguyen, H. H. T. Vu, and M. Le. "Convolutional neural network for people counting using UWB impulse radar." *Journal of Instrumentation* 16, no. 08 (2021): P08031.

[5] Wang, Chuan, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. "Deep people counting in extremely dense crowds." In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1299-1302. 2015.

[6] Kanatov, Maksat, and Lyazzat Atymtayeva. "Deep convolutional neural network based person detection and people counting system." *Advanced Engineering Technology and Application* 7, no. 3 (2018): 5-9.

[7] Yang, Xiuzhu, Wenfeng Yin, and Lin Zhang. "People counting based on CNN using IR-UWB radar." In *2017 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1-5. IEEE, 2017.

[8] Xu, Huazhong, Pei Lv, and Lei Meng. "A people counting system based on head-shoulder detection and tracking in surveillance video." In *2010 international conference on computer design and applications*, vol. 1, pp. V1-394. IEEE, 2016.

[9] Gao, Chenqiang, Pei Li, Yajun Zhang, Jiang Liu, and Lan Wang. "People counting based on head detection combining Adaboost and CNN in crowded surveillance environment." *Neurocomputing* 208 (2016): 108-116.

[10] Shi, Xiaowen, Xin Li, Caili Wu, Shuchen Kong, Jing Yang, and Liang He. "A real-time deep network for crowd counting." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2328-2332. IEEE, 2020.

[11] Boominathan, Lokesh, Srinivas SS Kruthiventi, and R. Venkatesh Babu. "Crowdnet: A deep convolutional network for dense crowd counting." In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 640-644. 2016.

[12] Babu Sam, Deepak, Shiv Surya, and R. Venkatesh Babu. "Switching convolutional neural network for crowd counting." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5744-5752. 2017.

[13] Lin, Htet Htet, and Kay Thi Win. "People counting system with C-deep feature in dense crowd views." In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp. 451-456. IEEE, 2018.

[14] Zhang, Yingying, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. "Single-image crowd counting via multi-column convolutional neural network." I

n *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589-597. 2016.

[15] Ding, Han, Jinsong Han, Alex X. Liu, Wei Xi, Jizhong Zhao, Panlong Yang, and Zhiping Jiang. "Counting human objects using backscattered radio frequency signals." *IEEE Transactions on Mobile Computing* 18, no. 5 (2018): 1054-1067.

[16] Yang, Xiuzhu, Wenfeng Yin, Lei Li, and Lin Zhang. "Dense people counting using IR-UWB radar with a hybrid feature extraction method." *IEEE Geoscience and Remote Sensing Letters* 16, no. 1 (2018): 30-34.

[17] Babu Sam, Deepak, Shiv Surya, and R. Venkatesh Babu. "Switching convolutional neural network for crowd counting." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5744-5752. 2017.

[18] Walach, Elad, and Lior Wolf. "Learning to count with cnn boosting." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 660-676. Springer International Publishing, 2016.

[19] Wu, Xingjiao, Baohan Xu, Yingbin Zheng, Hao Ye, Jing Yang, and Liang He. "Fast video crowd counting with a temporal aware network." *Neurocomputing* 403 (2020): 13-20.

[20] Sam, Deepak Babu, and R. Venkatesh Babu. "Top-down feedback for crowd counting convolutional neural network." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1. 2018.