applied sciences

MDPI

Article

A Systematic Deep Learning Based Overhead Tracking and Counting System Using RGB-D Remote Cameras

Munkhjargal Gochoo [1], Syeda Amna Rizwan [2], Yazeed Yasin Ghadi [3], Ahmad Jalal [2] and Kibum Kim [4,*]

[1] Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain 15551, United Arab Emirates; mgochoo@uaeu.ac.ae
[2] Department of Computer Science, Air University, Islamabad 44000, Pakistan; 190211@students.au.edu.pk (S.A.R.); ahmadjalal@mail.au.edu.pk (A.J.)
[3] Department of Computer Science and Software Engineering, Al Ain University, Abu Dhabi 122612, United Arab Emirates; Yazeed.ghadi@aau.ac.ae
[4] Department of Human-Computer Interaction, Hanyang University, Ansan 15588, Korea
* Correspondence: kibum@hanyang.ac.kr

Featured Application: The proposed technique is an application for people detection and counting which is evaluated over several challenging benchmark datasets. The technique can be applied in heavy crowd assistance systems that help to find targeted persons, to track functional movements and to maximize the performance of surveillance security.

Abstract: Automatic head tracking and counting using depth imagery has various practical applications in security, logistics, queue management, space utilization and visitor counting. However, no currently available system can clearly distinguish between a human head and other objects in order to track and count people accurately. For this reason, we propose a novel system that can track people by monitoring their heads and shoulders in complex environments and also count the number of people entering and exiting the scene. Our system is split into six phases; at first, preprocessing is done by converting videos of a scene into frames and removing the background from the video frames. Second, heads are detected using Hough Circular Gradient Transform, and shoulders are detected by HOG based symmetry methods. Third, three robust features, namely, fused joint HOG-LBP, Energy based Point clouds and Fused intra-inter trajectories are extracted. Fourth, the Apriori-Association is implemented to select the best features. Fifth, deep learning is used for accurate people tracking. Finally, heads are counted using Cross-line judgment. The system was tested on three benchmark datasets: the PCDS dataset, the MICC people counting dataset and the GOTPD dataset and counting accuracy of 98.40%, 98%, and 99% respectively was achieved. Our system obtained remarkable results.

Keywords: Apriori-Association; Cross-line judgment; deep learning; head tracking; Hough Circular Gradient Transform; Fused intra-inter trajectories

check for updates

Citation: Gochoo, M.; Rizwan, S.A.; Ghadi, Y.Y.; Jalal, A.; Kim, K. A Systematic Deep Learning Based Overhead Tracking and Counting System Using RGB-D Remote Cameras. *Appl. Sci.* **2021**, *11*, 5503. https://doi.org/10.3390/app11125503

Academic Editor: Emilio Soria-Olivas

Received: 27 April 2021
Accepted: 5 June 2021
Published: 14 June 2021

1. Introduction

Head and shoulders detection has become a research hotspot which plays a significant role in people counting [1] and crowd analysis which can be used for several practical applications such as surveillance, logistics and resource management coding and public transportation systems [2,3]. Many studies have been carried out on RGB image based head and shoulders counting but, due to the development of depth cameras and sensors, researchers are now studying RGB-Depth images for crowd counting using head and shoulders tracking. Compared with RGB images, RGB-D images provide additional and more general depth map information for the detection of heads and shoulders.

Computer vision techniques provide remarkable performance improvements to the problem of automatic head and shoulders detection and tracking in complex indoor/outdoor environments [4,5]. However, research and development is usually carried out on whole body detection and counting using RGB videos which are challenged by multiple issues such as variations in occlusion, illumination, clutter, shadows etc. Thus, different RGB-D cameras (e.g., Kinect V1 [6–10], Vzence DCAM and many more) are used to solve these issues by providing depth information. However, head and shoulder counting using depth datasets is still a challenging task for many researchers due to various unsolved problems related to occlusions and noise.

Vision-based head counting is a challenging task that involves different techniques such as object detection, human detection, object and human tracking and recognition [11]. Moreover, techniques used in this area are categorized into three main streams; (1) clustering-based methods, (2) regression-based methods, and (3) detection-based methods. Clustering-based methods target certain objects, track their features and cluster the object trajectories to count them [12–17]. Regression based methods use regression function learning which uses human object and non-human object features and utilizes them to count people [18,19]. Detection based methods have a common architecture which is divided into image/video pre-processing, object or body detection, feature extraction and classification [20–27]. These three main streams are further divided based on the data types they use, e.g., color/depth or hybrid videos etc. These approaches face some common issues during real time people counting under practical conditions, e.g., restricted camera angles, computational time and complexity, failure to handle cluttered scenes and excessively occluded images etc. [28–34]. In this paper, we describe a novel method of head and shoulders tracking and counting using depth datasets that address such issues.

Our proposed work flow is divided into six main phases. First, video preprocessing is done in which videos are converted into frames and complex backgrounds are removed from the image frames. Second, heads are detected in the frames using Hough Circular Gradient Transform and shoulders are detected by a joint HOG based symmetry method. Third, Robust features such as Fused joint HOG-LBP, Energy based Point clouds and Fused intra-inter trajectories are extracted. Fourth, Apriori-Association is implemented to select the best features. Fifth, Convolution Neural Network (CNN) is used for accurate people tracking. Finally, heads are counted using the Cross line judgment technique.

The main contributions of our system are listed below:

- Complex backgrounds with excessive occlusions in videos cause mis-detection of individual sets of heads and shoulders. We use novel techniques to mitigate the problems associated with occlusions and to more precisely detect individual sets of heads and shoulders.
- Our salient feature vectors provide far better accuracy than other state-of-the art techniques.
- The Apriori-Association rule is used for the selection of the ideal sets of features along with the CNN classifier for head tracking.
- Our head and shoulders tracking and counting (HASTAC) system performance is evaluated using three benchmark datasets; (1) the PCDS dataset, (2) the MICC people counting dataset and (3) the GOTPD dataset. Our proposed model was fully validated for its efficacy, outperforming other state-of-the-art methods.

This article is structured as follows: Section 2 describes related work. Section 3 gives a the detailed overview of the proposed HASTAC model. In Section 4, the proposed model's performance is assessed on three publicly available benchmark datasets through various experiments. Lastly, in Section 5 we sum up the paper and outline future directions.

2. Related Work

Over the past few years, various studies on head tracking and counting using RGB and RGB-D datasets have been reported. In this section we are going to give a broader view of the latest techniques and methodologies used in these systems. They are divided into two main streams; (1) Head tracking and counting using RGB datasets (See Section 2.1) and (2) Head Tracking and counting using RGB-D datasets (See Section 2.2).

2.1. Head Tracking and Counting using RGB Datasets

Many head tracking and counting systems that work on RGB datasets have been developed in recent years. Table 1 gives a detailed account of these systems.

Table 1. Detailed report of head tracking and counting systems that used RGB datasets.

Paper Name	RGB Datasets	Methodology	Classification/Regression Results
A people counting method based on head detection and tracking [35]	Pedestrian dataset and Internet Images	The foreground was extracted by some morphological operations. For head detection, an LBP based Adaboost classifier was used. MeanShift algorithm was used for head tracking. Finally, the heads were counted based on the Cross-Line Judgement technique.	People counting accuracy was 96%.
A people counting system based on head-shoulder detection and tracking in surveillance video [36]	Monocular surveillance videos	The system detects people from video sequences using HOG and Dalal HOG-SVM Bootstrapping framework. Each pedestrian is tracked using a Kalman filter and each pedestrians motion trajectory is calculated to track each individual.	The system uses three video sequences. Each video sequence is divided into two parts (in and out). Results show the detection accuracy achieved lies between 85–95%.
Vision-based People Counting for Attendance Monitoring System [37]	Custom dataset	The system was developed for attendance monitoring. The system was divided into three parts. First, the people were detected using the trained deep learning model, i.e., Mobile-Net SSD. Secondly, the Tracking process was implemented using the Correlation Filter and SORT. Finally, people were counted based on reference lines and tracked path.	Accuracy can be measured using the COCO evaluation metric. They acheieved performance score of 83% was achieved.
People Counting Based on Head Detection and Reidentification in Overlapping Cameras System [38]	Gallery-Probe database	This system was developed to count the people who occur simultaneously on two cameras located at different positions in overlapping areas. A trained head detector was used to detect the heads from the images taken from the two cameras and these pictures of the two cameras are passed to a Siamese network. Then the re-identification of humans step was carried out to count the people in the repeating area. Finally, the median of the total number of individuals that were counted from the video intervals at a certain time was the attendace of the class.	The highest accuracy rate achieved was 71.40%.
Counting People in the Crowd Using a Generic Head Detector [39]	PETS 2012 and Turin metro station	The system was developed for counting people in a crowd. Initially, the head of each individual is detected using cascade-of boosted integral features. A gradient orientation feature was used to detect the interest points. Background subtraction was done using Vibes and Idiap techniques. At last, heads were counted based on Adaboost classifier.	A regression based method is used for the counting of people in crowded environemnts. The MAE rate lies between 68–95 for 6 different views from the Turin metro station dataset.

2.2. Head Tracking and Counting using RGB-D Datasets

From the past few years, due to the arrival of depth cameras and sensors, many researchers are carrying out studies on head tracking and counting using depth imagery. Table 2 gives a detailed account of these recent systems.

Table 2. Detailed explanation of head tracking and counting systems that used RGB-D datasets.

Paper Name	RGB-D Datasets	Methodologies	Classification/Regression Results
People counting base on head and shoulder information [40]	Depth Surveillance Videos	The Nearest Neighbor interpolation technique was used for background subtraction which results in greater computational effeciency. Candidate heads and shoulders were detected using edge detection and circle detection techniques. The tracking of heads was done by tracking the circle using the nearest neighbor interpolation method. Finally, a virtual line is used to count the people crossing that line.	The precision results obtained were based on two scenerios, (1) Two to five people were passing the line without holding any object and, (2) People with bags, luggage and crossing the virtual line. The precision results in both scenerios were 100% and 95%, respectively.
Benchmark Data and Method for Real-Time People Counting in Cluttered Scenes Using Depth Sensors [41]	PCDS dataset	The system was developed to calculate the number of people entering or exiting a bus. Intially, the background was removed from each frame of a video by the farthest background model. A 3D human model and seed fill method can reliably detect the human heads in frames. After that, the human heads were tracked to identify their trajectories which further helps in human counting.	The dataset is divided into four parts that are identified by $\boldsymbol{N^-C^-}, \boldsymbol{N^-C^+}, \boldsymbol{N^+C^-}, \boldsymbol{N^+C^+}$. People counting accuracy over these four parts on entering were 85.40%, 83.25%, 77.54% and 75.32%. People counting accuracy rate over these four parts upon exiting were 93.04%, 92.66%, 93.71% and 91.30%, respectively.
Real-time people counting from depth imagery of crowded environments [42]	MICC dataset	This system was developed to count people in crowded environments. In order to detect the head of an individual, background subtraction was implemented first using selective running average background subtraction. For head tracking, connected components are used. To eliminate the problem of overlapping connected pixels, edges were detection. A multitarget tracker using greedy data association was used to track the entrance and exit of people from a designated area.	The precision rate achieved over the MICC dataset was 97.9%
Depth driven people counting using deep region proposal network [43]	CBSR dataset	The main goal of this system was to count people in crowded environments. The authors used CNN for head detetcion. The authors also explored the impact of the number and quality of RPN anchors over the faster RCNN model and improved the performance by proposing a new solution.	The precision rate was 97.54%
3D Head Pose Estimation through Facial Features and Deep Convolutional Neural Networks [44]	Pointing'04, BU, AFLW, and ICT-3DHPE	The authors introduced an end to end face parsing algorithm which tries to address a challenging problem of face pose estimation. The face parsing model through DCNNs by extracting useful information from different face parts was trained. The face parsing model provides a class label for each pixel in a face image. We use a probabilistic classification technique and create PMAPS in the form of grey scale images for each face class	The accuracy achieved on the Pointing'04 dataset is 96.5% and MAE of BU, AFLW, and ICT-3DHPE was 3.6, 2.1 and 3.0 respectively.

3. The Proposed System Methodology

This section discusses the overall proposed methodology that is used in our HASTAC system. The system framework is divided into 6 major steps. First, video preprocessing is done in which the video is converted into frames and complex backgrounds are removed from the image frames using the Kernel Density Estimation (KDE) technique. Second, heads are detected in the frames using Hough circular gradient transform and shoulders are detected by joint HOG based symmetry methods. Third, Robust features such as Fused joint HOG-LBP, Energy based point clouds and Fused intra-inter trajectories are extracted. Fourth, Apriori-Association is implemented to select the best features. Fifth, Convolution Neural Network (CNN) is used for accurate people tracking. Finally, heads are counted using the Cross-line judgment technique. The overall architecture of our HASTAC system is shown in Figure 1.

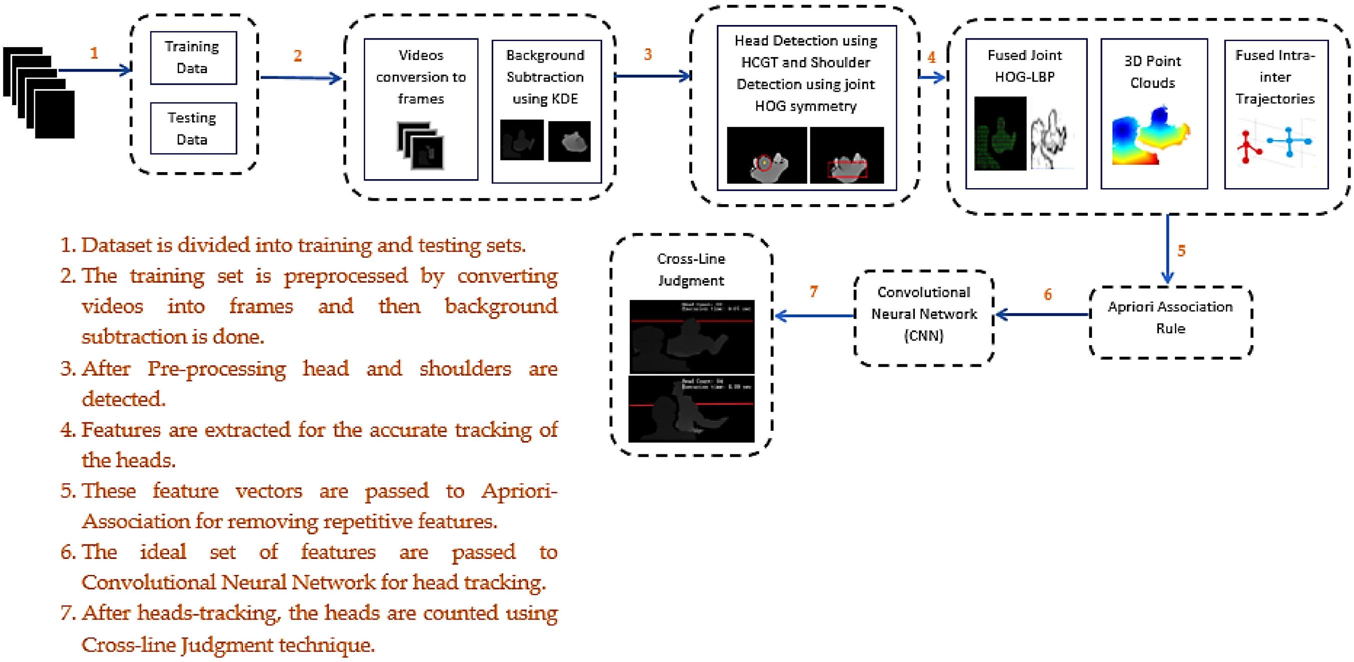

1
Training Data
Testing Data
2
Videos conversion to frames
Background Subtraction using KDE
3
Head Detection using HCGT and Shoulder Detection using joint HOG symmetry
4
Fused Joint HOG-LBP
3D Point Clouds
Fused Intra-inter Trajectories
5
Apriori Association Rule
6
Convolutional Neural Network (CNN)
7
Cross-Line Judgment
1. Dataset is divided into training and testing sets.
2. The training set is preprocessed by converting videos into frames and then background subtraction is done.
3. After Pre-processing head and shoulders are detected.
4. Features are extracted for the accurate tracking of the heads.
5. These feature vectors are passed to Apriori-Association for removing repetitive features.
6. The ideal set of features are passed to Convolutional Neural Network for head tracking.
7. After heads-tracking, the heads are counted using Cross-line Judgment technique.

Figure 1. System architecture of the proposed HASTAC system.

3.1. Preprocessing Stage

During video pre-processing, the first step is the conversion of videos into image frames. Thus, the depth frames are extracted from static videos $V = [F_1, F_2, \ldots, F_n]$ where n is the number of frames. The next step is to remove complex backgrounds from these depth frames using the Kernel Density Function (KDE) technique. KDE is a non-parametric technique for estimating the densities of the pixels. The main idea of KDE is that each frame background is identified by the histogram of the most recent pixels N, where each pixel is smoothed by a kernel (generally a Gaussian kernel). Perhaps, the main objective of using KDE is to detect changes occurring frequently in the background and to accurately detect and identify the target objects in the forground with high sensitivity.This technique works by selecting the most recent frames of a video and updating these frames continuously in order to update changes in the background. These frames are collected as samples and their pixels are further processed to obtain the intensity values of each pixel.

By using the intensity value samples of the pixels the probability density function of each pixel has the intensity value i_T at time T, and kernel estimator E can be used to estimate it as [45];

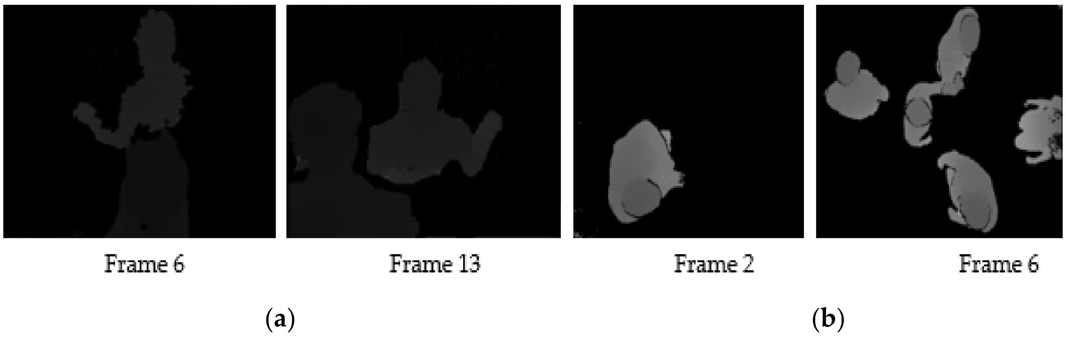
$$P(i_T) = \frac{1}{n} \sum\nolimits_{x=1}^{n} E(i_T - i_x) \tag{1}$$

where $i_1, i_2, \ldots, i_n$ are the recent intensity value samples of a pixel.

E can be calculated using Gaussian distribution for the depth frames as in [45];

$$P(i_T) = \frac{1}{n} \sum_{x=1}^{n} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{-1}{2} \frac{(i_{T_k} - i_x)^2}{\sigma_k^2}} \tag{2}$$

By calculating the probability estimate, the set of pixels are considered as foreground pixels if $P(i_T) < Th_1$ where Th_1 is the global threshold that can be adjusted for every image frame in order to get a desired false positive rate. Figure 2a,b show the results obtained after background subtraction using the KDE technique over PCDS and GOTPD datasets, respectively.



(**a**) (**b**)

Figure 2. Results obtained after background subtraction using the KDE technique over (**a**) PCDS dataset and (**b**) GOTPD dataset.

3.2. Head and Shoulders Detection

Hough Circular Gradient Transform (HCGT) is a technique that detects the circles by first passing an imperfect image to the edge detection phase. In our proposed work the HCGT technique is used for detecting heads in an image by first finding the edges of an image. These edges will further help to find the circular heads in an image by considering the local gradient. The circle in an image is defined using Equation (3);

$$(x - p)^2 + (y - q)^2 = r^2 \tag{3}$$

where p and q are the circle center coordinate points, r is the radius and x and y are the arbitrary edge points. Because the dataset images are taken from the top view cameras the heads are more likely to appear circular. Circular candidates are detected in the Hough parameter space by voting and selecting the local maxima in an accumulator matrix. The location of each non-zero pixel is identified to get the centers of the heads from the points in an accumulator that are above the threshold and larger than their candidate neighbors. These candidate centers are then sorted in descending order by their accumulator values, causing the most supported pixels to appear first. All the non-zero pixels are considered for every center. These pixels are then sorted according to the minimum distance from the centers of the circles. Considering the smallest distance to the largest distance, the radius of the circle that is less than the given threshold *thresh_R* is selected. The center that provides sufficient support for the non-zero pixels on the edge of the image is kept and provides sufficient distance from a previously selected center. Thus, the HCGT technique can divide the problem of finding the circular head into two sub-stages. First, the candidate centers are found and then the appropriate circle radius is found. A 2D array is required to store the votes of each edge point whereas the distance between each point is accumulated to find the radius of the circle. Figure 3 shows the final results of head detection using the HCGT technique over the GOTPD dataset.

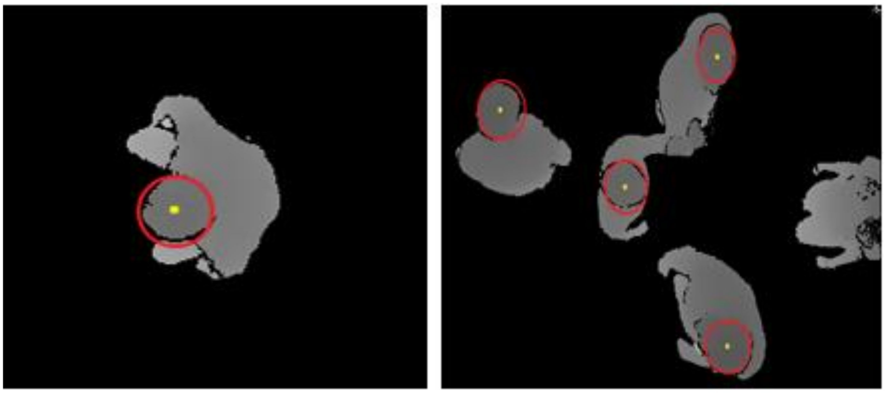


Figure 3. Final results of head detection using the HCGT technique over the GOTPD dataset.

For the detection of shoulders, the joint HOG symmetry-based detection method is used. This type of HOG detection method uses two descriptors of the same size which use the symmetry and continuity of shoulder contours to get higher discrimination and greater detection accuracy. d_i and d_j are the two descriptors, the sample p joint HOG features can be calculated as [46];

$$HOG_{d_i,d_j}(p) \tag{4}$$

The pixel position on x and y has a depth value l, gradient magnitude g_m and the direction of the gradient is θ. We extracted a 1D center gradient operator [−1, 0]. The horizontal gradient and the vertical gradient are calculated as [46];

$$hg_x(x,y) = l(x+1,y) - l(x-1,y) \tag{5}$$

$$hg_y(x,y) = l(x,y+1) - l(x,y-1) \tag{6}$$

whereas, the pixel gradients located at position (*x*,*y*) are calculated as [46];

$$hg(x,y) = \sqrt{hg_x(x,y)^2 + hg_y(x,y)^2} \tag{7}$$

$$\theta(x,y) = tan^{-1}\frac{hg_x(x,y)}{hg_y(x,y)} \tag{8}$$

Thus, each cell of 4 × 4 pixels their gradient magnitude according to their gradient direction are calculated and accumulated in bin direction. A histogram of the gradient is obtained. The directional range is from 0–180° for the joint HOG features of the shoulders, with 5 bin directions. The joint HOG features for the shoulders are then expressed as [46];

$$H_{d_i,d_j} = \left\{H_{d_i}, H_{d_j}\right\} = v_1, v_2, \ldots, v_{L+1} \tag{9}$$

Finally, these HOG features are normalized to get the joint HOG features of different blocks. Figure 4 shows the results obtained for shoulder detection over the GOTPD dataset.

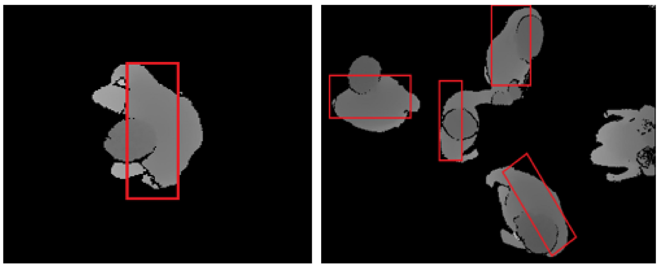


Figure 4. Results obtained for shoulder detection using joint the HOG symmetry technique over the GOTPD dataset.

3.3. Feature Extraction

To track the number of heads in depth frames, local and global feature extraction plays a vital role. Therefore, Fused joint HOG-LBP, Energy based Point clouds and Fused intra-inter trajectories are passed to an Apriori Association algorithm to remove the unnecessary redundant features (See Algorithm 1).

Algorithm 1: Feature Extraction.

```
Input: Y: Detected head region;
Output: Extracted Features;
/* Fused joint HOG-LBP*/
/* Energy based point clouds*/
/* Fused intra-inter trajectories*/
Fused joint HOG-LBP
/* For a sample $p$ HOG feature can be calculated as;
$HOG_{d_i,d_j}(p)$;
/* The Horizontal and vertical gradient */
$hg_x(x,y) = l(x+1,y) - l(x-1,y)$;
$hg_y(x,y) = l(x,y+1) - l(x,y-1)$;
/* The magnitude and direction of the pixel gradient */
$hg(x,y) = \sqrt{hg_x(x,y)^2 + hg_y(x,y)^2}$;
$\theta(x,y) = tan^{-1}\frac{hg_x(x,y)}{hg_y(x,y)}$;
/* HOG features description of heads and shoulders */
$H_{d_i,d_j} = \left\{H_{d_i}, H_{d_j}\right\} = v_1, v_2, \ldots, v_{L+1}$;
Energy based point clouds
/* Energy based point clouds w..r.t central point of head and other 5 fiducial points marked */
$D = \begin{cases} A=\{A=D_x+D_y+\sqrt{\Delta} & where\ \Delta\geq 0 \\ B=\min(D_x,D_y)+a & otherwise \end{cases}$
$\Delta = 2a^2 - \left(D_x - D_y\right)^2$;
Fused intra-inter point trajectories
/* Five fiducial points are marked on head, neck, two shoulders and chest. Then, find the change
in displacement of the trajectories */
$\Delta p_t = (x_{t+1}, -x_t, y_{t+1} - y_t)$;
$d_{x,y} = \frac{(\Delta p_1,\ \Delta p_2,\ldots,\Delta p_{T-1}}{\sum_{i-1}^{T-1}||\Delta p_i||}$;
Augment all features extracted;
$A = \left|H_{d_i,d_j}\right||D||\Delta||\Delta p_t|d_{x,y}\Big|$;
Pass to Apriori-Association
end;
```

3.3.1. Fused Joint HOG-LBP

After the normalization of all the HOG features to obtain joint HOG features of different blocks (See Section 3.2), histograms for all the overlapping blocks are collected over the detection window [46]. Then, detection window values are fused with Local Binary Pattern (LBP) to improve the performance of feature extraction. Using a specific threshold, image pixels are labeled by comparing it with every neighboring pixel and then converting it into binary. Again, the window of an image is divided into cells of 16 × 16 pixels and each pixel is compared to its neighboring pixel, i.e., a central pixel's value is compared with the 8 neighboring pixels. If the value of the central pixel is greater than a neighboring pixel, it is replaced with the value 0, otherwise it is replaced with value 1. Figure 5 shows the results of fused joint HOG-LBP results over the PCDS dataset.

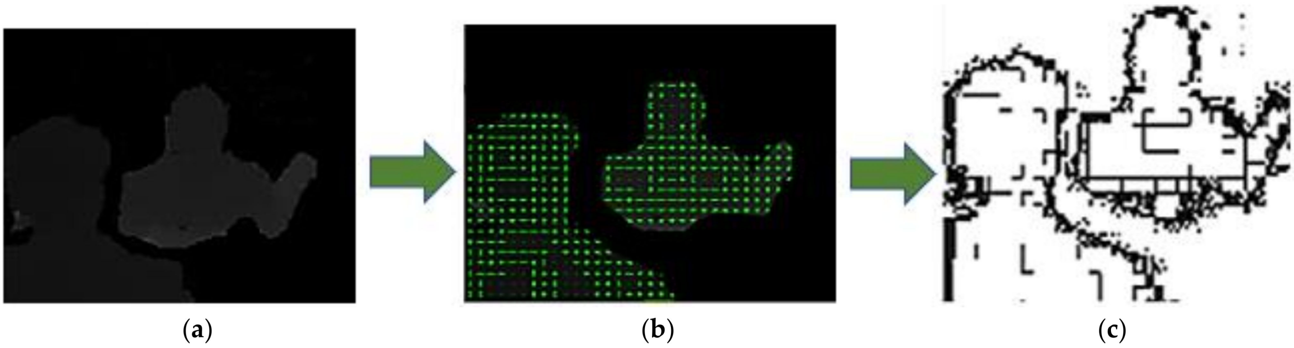


Figure 5. Results of fused joint HOG-LBP over the PCDS dataset where (**a**) shows the original depth image, (**b**) shows the HOG results and (**c**) shows the LBP results.

3.3.2. Energy Based Point Clouds

The Energy based point cloud method is quite similar to the geodesic distance algorithm. According to our knowledge, it is used for automatic head counting for the first time. It is robust, efficient and simple to implement. In this method, a central point v ϵ V is selected as the anchor point on the head and is given the fixed distance of zero $d(v) = 0$. It is inserted as priority queue S, priority being based on the smallest distance. All the other points $s \nexists V$ are labeled with distance $d(s) = \infty$. One point v from the priority queue is selected then, based on the geodesic distance algorithm which works on the principle of the Dijkstra algorithm, the shortest distance between the central point to the other varying fiducial points is found. Then, energy based point clouds can be displayed. These point clouds are changed according to the varying positions of the head, i.e., as the positions of the fiducial points are changed. The distances between the central points to the other points are examined as optimal features. The change in distance between fiducial points is calculated as [33];

$$D = \{ {}^{A=\{A=D_x+D_y+\sqrt{\Delta}\ \mathit{where}\ \Delta\geq 0}_{B=\min(D_x,D_y)+a\quad \mathit{otherwise}} \tag{10}$$

$$\Delta = 2a^2 - \left(D_x - D_y\right)^2 \tag{11}$$

where $D_x = min\left(d_{p+1,q},\ d_{p-1,\,q}\right)$ and $D_y = min\left(d_{p,\,q+1},\ d_{p,\,q-1}\right)$. Figure 6 shows the results for the energy based point clouds obtained over the PCDS dataset.

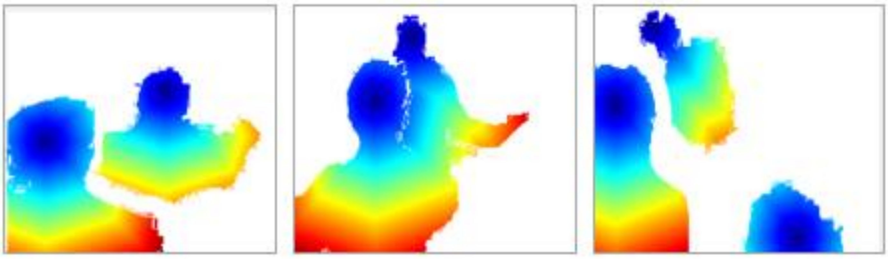


Figure 6. Energy based point cloud results obtained over frames of the PCDS dataset.

3.3.3. Fused Intra-Inter Trajectories

We propose "fused intra-inter depth silhouette localized point trajectories" for the first time for human head tracking and counting. This type of trajectory method contains a subset X, containing a set of human joints which gives localized points to form trajectories. First, a subset X containing five localized points is plotted on human the depth silhouette s, i.e., X = {*head*, *neck*, *left_shoulder*, *right _shoulder* and *chest*}. The number of localized points varies according to the number of human depth silhouettes N in each frame i.e., $N = \{n_1,\ n_2,\ \ldots,\ n_\infty\}$. These localized points are then joined in the form of trajectories. This results in four trajectories for each human silhouette. These four trajectories are represented as $T = \{HN, NRS,\ NLS,\ NC\}$ where *HN* is the trajectory between head and neck, *NRS* is the trajectory between neck and right_shoulder, *NLS* represents the trajectory between neck and left_shoulder and at last *NC* represents the trajectory between neck and chest. Figure 7 shows the fused inta-inter silhouettes trajectories over the MICC dataset.

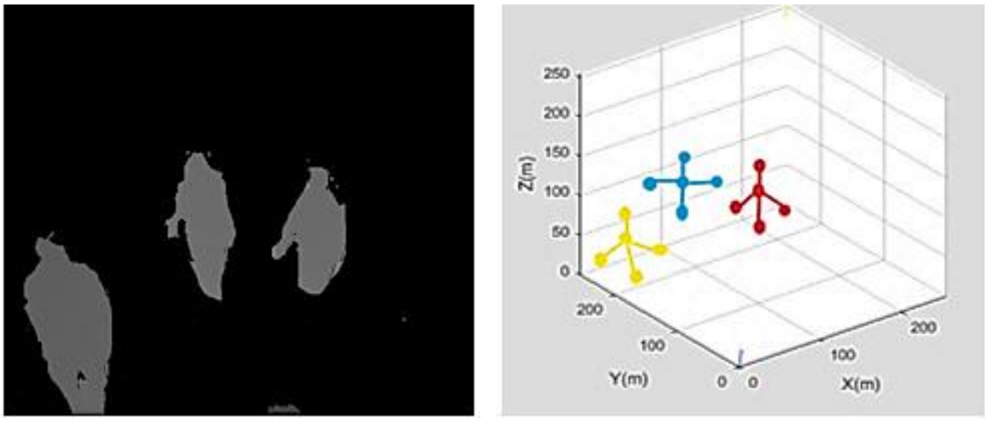


Figure 7. Fused intra-inter silhouette trajectories over the MICC dataset.

After the formation of all the trajectories, shape descriptors are extracted which can calculate the displacement of the changes in length L in each silhouette, frame by frame, during the time t along x and y coordinates. This change during time t can be measured using Equation (12) and the normalized displacement vector can be measured using Equation (13) [33];

$$\Delta p_t = (x_{t+1}, -x_t, y_{t+1} - y_t) \tag{12}$$

$$d_{x,y} = \frac{(\Delta p_1,\ \Delta p_2, \ldots, \Delta p_{T-1})}{\sum_{i-1}^{T-1} ||\Delta p_i||} \tag{13}$$

3.4. Apriori Association

The Apriori-Association technique is used to detect and extract meaningful association relationships between the quantities in a dataset. It is used in several practical applications such as the study of disease, improvements in the production process, correlations in alarm analysis etc., and it is used, for the first time, in our proposed head tracking and counting system. An Apriori-association technique identifies the data itemset which frequently occurs in a dataset. At first, the minimum support of the individual itemset is calculated to identify frequent items in the first pass. In the second pass, the same procedure is repeated by taking each seed set of the previous transaction and finding the frequently occurring itemset. This procedure is repeated several times until no frequent itemset is found. This algorithm generates candidate item sets on each transaction in order to improve the computational efficiency [47]. Figure 8 illustrates the Apriori-Association graph of the accuracy of all three datasets across three feature extraction techniques. Algorithm 2 defines the step by step procedure of the working of the Apriori-Association.

3.5. Head Tracking Using Convolution Neural Network

All the local and global features extracted from the above mentioned feature extraction methods (See Section 3.3) are then passed through CNN, resulting in the accurate tracking of heads over the three benchmark datasets. CNN always promised to give better results than other deep learning techniques in both RGB and RGB-D image and video based features [48–53]. Figure 9 illustrates the overall architecture of our proposed 1D CNN over the PCDS dataset.

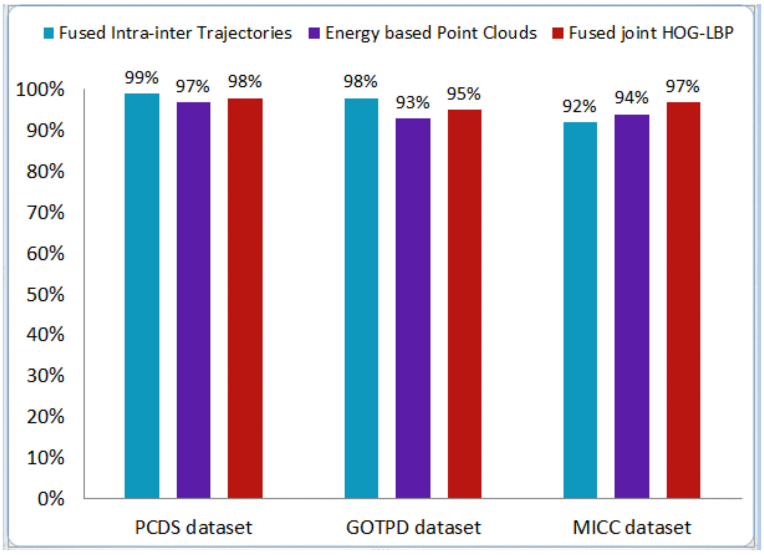


Figure 8. Illustration of the Apriori-Association graph of the accuracy of all three datasets across three feature extraction techniques.

Algorithm 2: Head Tracking (Training).

```
Apriori-Association
Begin
Project A on Apriori-Association;
For $p$ = 1, $W_1 = all\ 1 - A$;
For $p > 1,\ generate\ W_p\ from\ Q_{p-1}\ as\ follows:$
/* The join step */
$W_p = p - 2$ way join $Q_{p-1}$ with itself;
If both $\{c_1, \ldots, c_{p-2}, c_{p-1}\}$ and $\{c_1, \ldots, c_{p-2}, c_p\}$ are in $Q_{p-1}$;
Then add $\{c_1, \ldots, c_{p-2}, c_{p-1}, c_p\}$ to $W_p$;
/* All features are sorted. Now the prune step */
Remove $\{c_1, \ldots, c_{p-2}, c_{p-1}, c_p\}$ if it contains a non-frequent $(p-1)$ subset;
Obtain P;
End for
Assign all depth features P to train CNN;
End;
```

1D CNN is proposed for the first time for the accurate tracking of heads in both indoor and outdoor complex environments. The PCDS dataset contains 10,500 feature sets in 4500 videos. The proposed CNN consists of three convolution layers, three max_poolinglayers and one fully connected layer. Each layer has its own specific purpose. The first layer is the convolution layer C_1 and contains the input matrix. This layer is convolved with 32 kernels each having a size of 1 × 13. As a result a matrix of 4500 × 10,488 × 32 is produced. The convolution matrix can be calculated as [54];

$$C_n^{(m-1)}(a,b) = ReLU(x) \tag{14}$$

$$ReLU(x) = \sum_{u=1}^{y} \Omega\left(a, \left(b - u + \frac{y+1}{2}\right)\right) W_n^m(u) + \alpha_n^m \tag{15}$$

where $C_n^{(m-1)}\ (a,b)$ generates the results of the convolution layer for the coordinates (a, b) of the $(m-1)$ layer with the nth convolution map. Ω represents the previous layer map and the size of the kernel is represented as x. W_n^m is the mth convolution kernel for the layer nth. α_n^m is the mth bias of the nth layer.

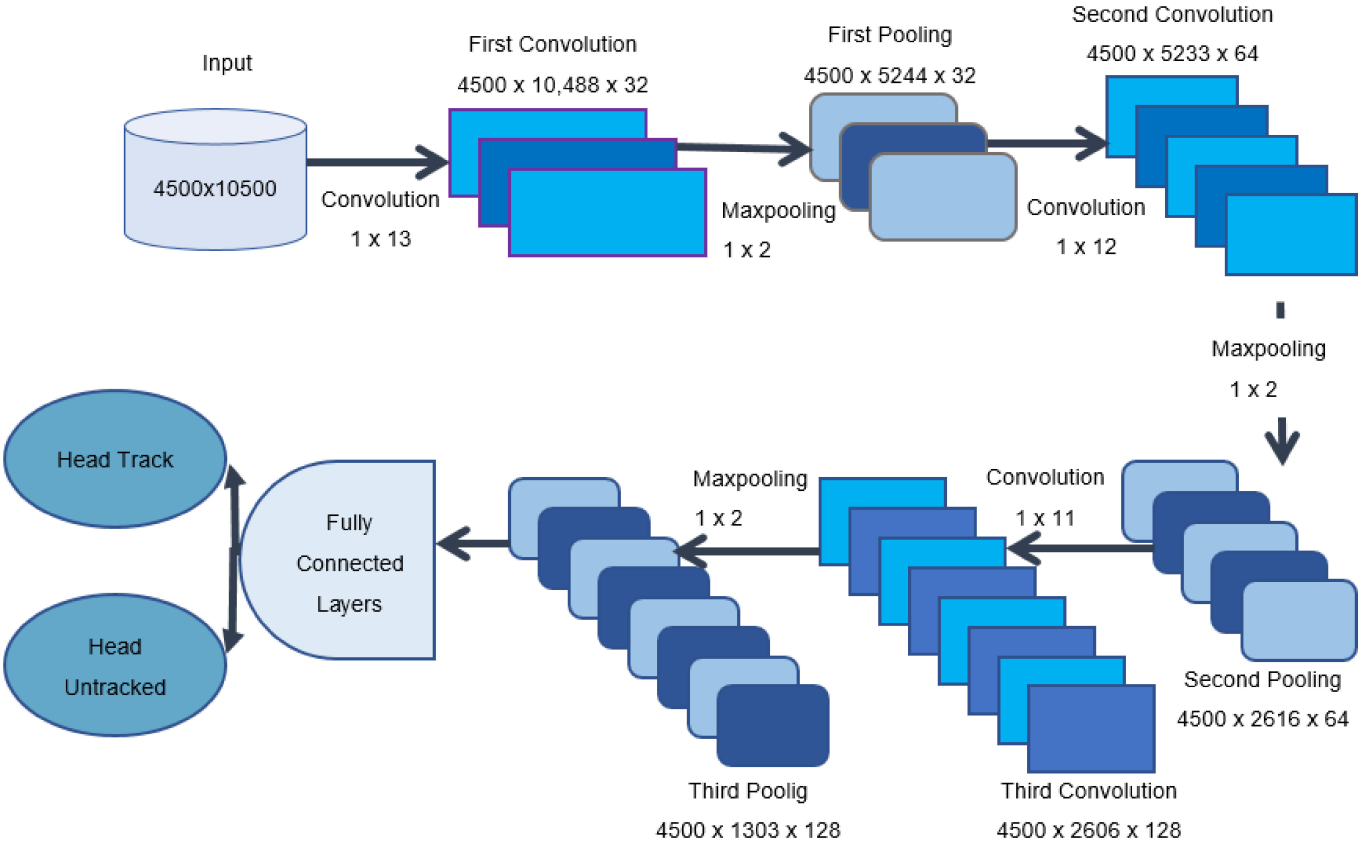


Figure 9. Illustration of the overall architecture of our proposed 1D CNN over PCDS dataset.

The result of the first convolution layer is passed through the first max_pooling layer P_1. A ReLU activation function is used between each convolution and max_pooling layer containing the sum of weights and bias of the previous layer passed to the next layer. The max_pooling layer downsamples the result obtained by the convolution layer by using a sliding window of size 1 × 2. Thus the pooling results of the $(p-1)$th layer, q kernel and x row and y column can be calculated as [55–58];

$$P_n^{(m-1)}(x,y) = max(C_n^m(x,((y-1)\times(p+q))) \qquad (16)$$

where $1 \leq p \leq q$ and n is the pooling window size. The result obtained by the first pooling layer is passed through the second convolution layer C2, convolved with 64 kernels and then the result is passed through the second pooling layer P2. The same procedure is repeated for the third convolution layer convolved with 128 kernels. Finally, a fully connected layer FC result is obtained as;

$$FC_v^{(m+1)} = ReLU(\textstyle\sum_i x_i^m W_{mv}^m + \alpha_v^m) \qquad (17)$$

where $W_{iv}^m w_{iv}^m$ represents the matrix having weights from the ith node of the mth layer to the vth node of the $(m+1)$th layer. x_i^m represents the content of the mth node at ith layer. Figure 10 represents the convergence plot of the CNN using 250 epochs of all three benchmark datasets.

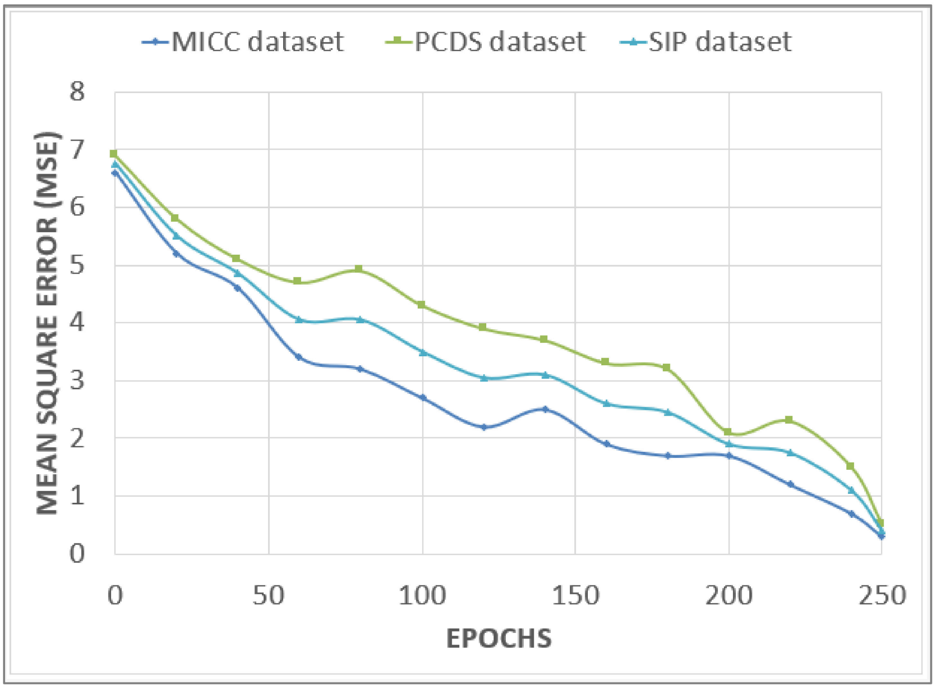


Figure 10. Convergence plot of the CNN using 250 epochs of all three benchmark datasets.

3.6. Head Counting Using Cross-Line Judgement

First, the counting line and region are specified in each frame of a video. Secondly, the movements of the heads are detected, i.e., either the heads are moving from left to right, right to left, upward to downward or downward to upward in direction. Hence, the relative location between heads and crossing lines can be determined and this information is further used to detect the position of the initial movement of the heads using the horizontal and vertical crossing lines. So if the crossing line is aligned horizontally, then the directions of the moving heads are calculated as in Equation (18), and, if the crossing line is aligned vertically, then the directions of the moving heads are calculated using Equation (19);

$$Direc_{head} = \begin{cases} up\_to\_down,\ y_{head_{init}} < y_u \\ down\_to\_up,\ y_{head_{init}} > y_d \end{cases} \tag{18}$$

$$Direc_{head} = \begin{cases} left\_to\_right,\ x_{head_{init}} < x_l \\ right\_to\_left,\ x_{head_{init}} > x_r \end{cases} \tag{19}$$

where $Direc_{head}$ denotes the initial movement of the head. $x_{head_{init}}$ and $y_{head_{init}}$ denotes the initial x and y coordinates of the center of the heads.

After detecting the direction of the initial head movements, the heads are counted based on the cross-line judgment as in [35];

$$Head_{count} = \begin{cases} left\_to\_right,\ Direc_{head} = left\_to\_right \text{ } and \text{ } x_{head} > x_r \\ right\_to\_left,\ Direc_{head} = right\_to\_left \text{ } and \text{ } x_{head} < x_l \\ up\_to\_down,\ Direc_{head} = up\_to\_down \text{ } and \text{ } y_{head} > y_d \\ down\_to\_up,\ Direc_{head} = down\_to\_up \text{ } and \text{ } y_{head} < y_u \end{cases} \tag{20}$$

where $Head_{count}$ denotes the direction of the heads being counted, x_{head} and y_{head} are the x and y coordinates of the center of the heads. If one of the above conditions is true, a counter of 1 is added in $Head_{count}$, otherwise the head will be discarded. Figure 11 shows the results obtained for head counting over the PCDS dataset.

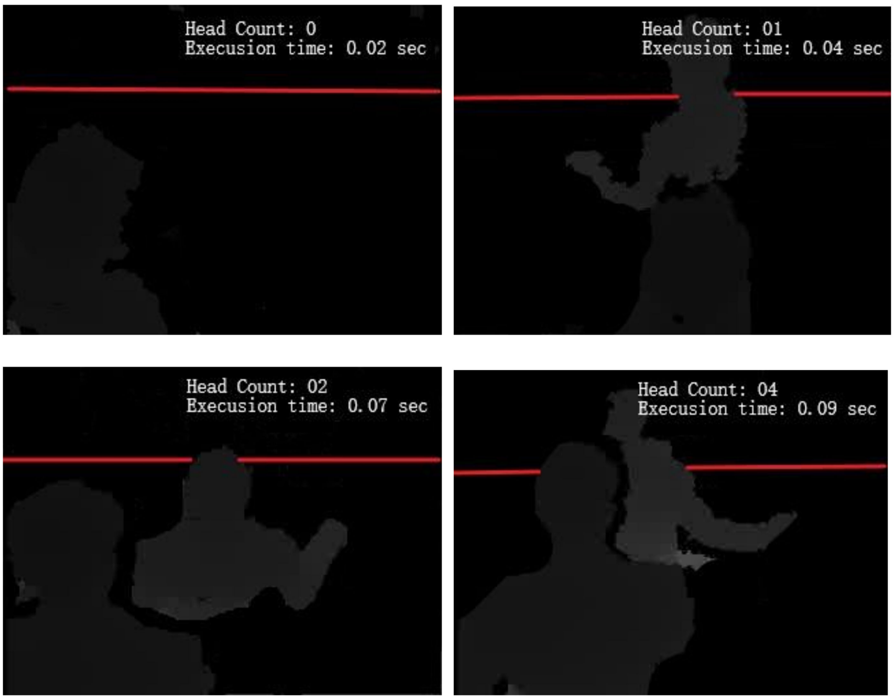


Figure 11. Results obtained for head counting using cross-line judgment over the PCDS dataset.

4. Experimental Setup and Performance

In this section, three benchmark datasets are described in detail (See Section 4.1). Different experiments are conducted over three benchmark datasets and their performances are evaluated. In experiment 1, head tracking performance is evaluated on all three datasets. In experiment 2, a comparison of the computational efficiency of our proposed model with the other state of the art methods is conducted. In experiment 3, the mean accuracy for head counting on all three benchmark datasets is illustrated in the form of a column chart.

4.1. Dataset Descriptions

This section describes the three benchmark datasets in detail.

4.1.1. The People Counting Dataset (PCDS)

The People Counting dataset (PCDS) is a first RGB-D dataset which contains 4500 videos that were taken at the entrance and exit of the bus in both normal and cluttered scenes. The videos are captured using a Kinect V1 camera. The dataset videos present large variations in illumination, occlusion, clutter and noise. This dataset is publicly available at [41]. Figure 12 shows video frame examples of the PCDS dataset.

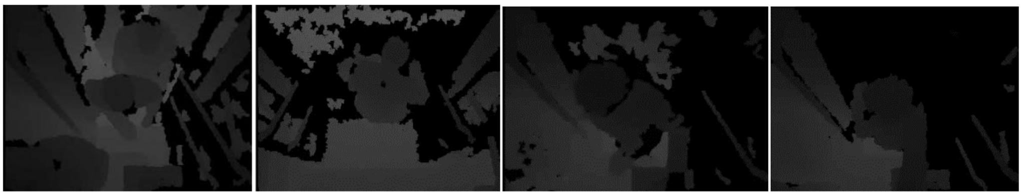


Figure 12. Video frame examples of the PCDS dataset.

4.1.2. The MICC People Counting dataset

The MICC people counting dataset is another publicly available RGB-D dataset [42]. This dataset is divided into three sequences; Flow Sequence, Queue Sequence and Groups Sequence. In the Flow sequence the participants are moving straight from one point to another point in the room. This sequence contains 1260 frames with 3542 persons. In the Queue sequence, the participants are waiting in a line in a room. This sequence contains

918 frames with 5031 persons. In the Groups sequence, the participants are split into two groups which talk to each other without exiting the room. The number of frames in this sequence is 1180 and the number of persons is 9057 [42]. Figure 13 shows examples of video frames form the MICC people counting dataset.

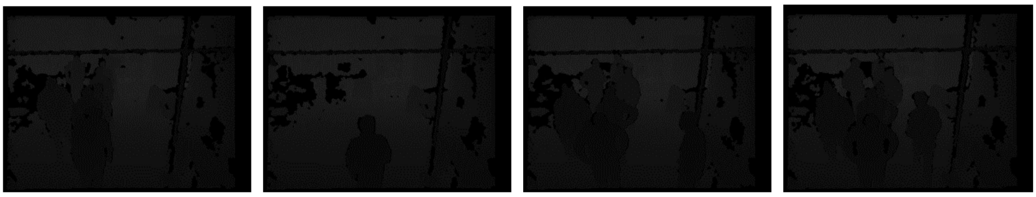


Figure 13. Some examples of the MICC people counting dataset.

4.1.3. The Geintra Overhead ToF People Detection dataset (GOTPD)

The GOTPD is a multimodel dataset containing both depth and infra-red video recordings which were captured using Kinect 2 camera. The camera was positioned above the heads for people detection. The dataset is composed of 48 video sequences each having different illumination conditions, various complex environments, various objects like hats, chairs, caps etc. The dataset frames consist of both single and multiple persons. Figure 14 shows some examples of the GOTPD dataset [59,60].

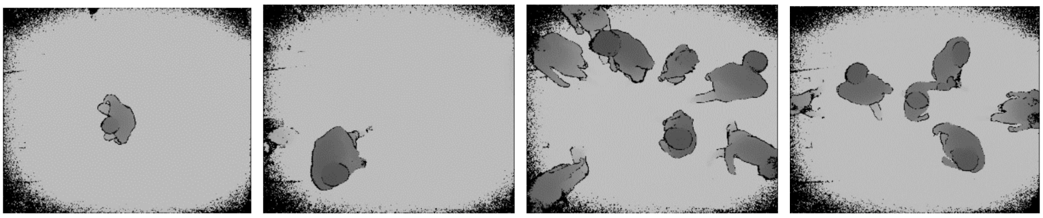


Figure 14. Some examples of the GOTPD dataset.

4.2. Experiment 1: Evaluation of Tracking Performance

Tables 3–5 shows the overall tracking accuracy, sensitivity and specificity performance metrices of the PCDS dataset, the MICC people counting dataset and the GOTPD dataset under various conditions. The accuracy, sensitivity and specificity measures are calculated as;

$$Accuracy_{head\_track} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

$$Sensitivity_{head\_track} = \frac{TP}{TP + FN} \tag{22}$$

$$Specificity_{head_{track}} = \frac{TN}{TN + FP} \tag{23}$$

where *TP* is the True Positive, *TN* is the True Negative, *FP* is the False Positive and FN is the False Negative.

Table 3. Tracking Accuracy achieved over the PCDS dataset.

Condition	Confusion Matrix	Head Track	Non-Head Track	Accuracy	Sensitivity	Specificity
Entering	Head track	0.98	0.02	0.986	0.99	0.98
	Non-head track	0.008	0.992			
Exiting	Head track	0.993	0.007	0.994	0.995	0.993
	Non-head track	0.005	0.995			

Table 4. Tracking Accuracy achieved over the MICC People Counting dataset.

Condition	Confusion Matrix	Head Track	Non-Head Track	Accuracy	Sensitivity	Specificity
Flow	Head track	0.973	0.027	0.971	0.969	0.972
	Non-head track	0.031	0.969			
Queue	Head track	0.992	0.008	0.993	0.995	0.992
	Non-head track	0.005	0.995			
Group	Head track	0.971	0.029	0.972	0.973	0.971
	Non-head track	0.026	0.974			

Table 5. Tracking Accuracy achieved over the GOTPD dataset.

Confusion Matrix	Head Track	Non-Head Track	Accuracy	Sensitivity	Specificity
Head track	0.965	0.035	0.951	0.939	0.964
Non-head track	0.062	0.938			

4.3. Experiment 2: Comparison of the Computational Efficiency of the Proposed Model with Other State of the Art Techniques

In this section, Figures 15–17 illustrate the comparison of computational time in milliseconds of head tracking between our proposed model and other state of the art techniques. The number of frames is 25 per sequence for all three bechmark datasets. Results show that our model is more efficient than other state of the art techniques. For the PCDS dataset, the computational time is calculated using the first 7000 video frames. For the MICC datset and the GOTPD dataset, the number of frames taken for testing computational efficiency is 3499 and 2950 respectively.

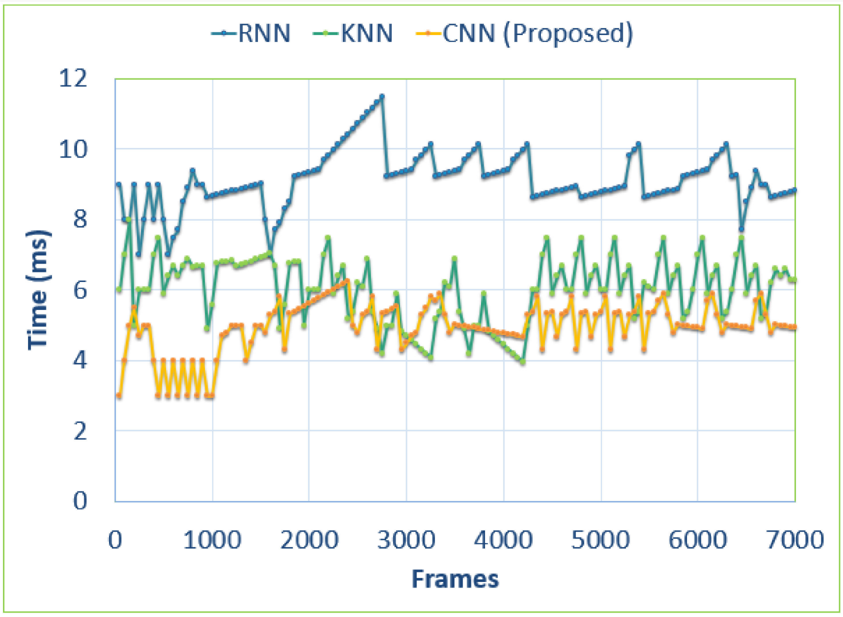


Figure 15. Comparison of the computational efficiency of our proposed model with other state of the art models over the PCDS dataset.

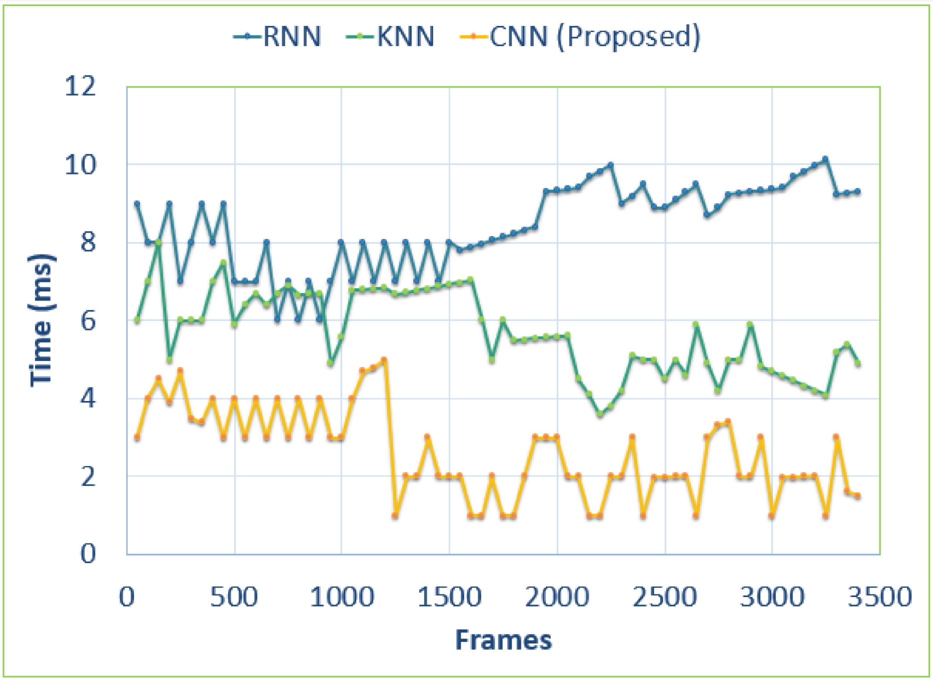


Figure 16. Comparison of the computational efficiency of our proposed model with other state of the art models over the MICC dataset.

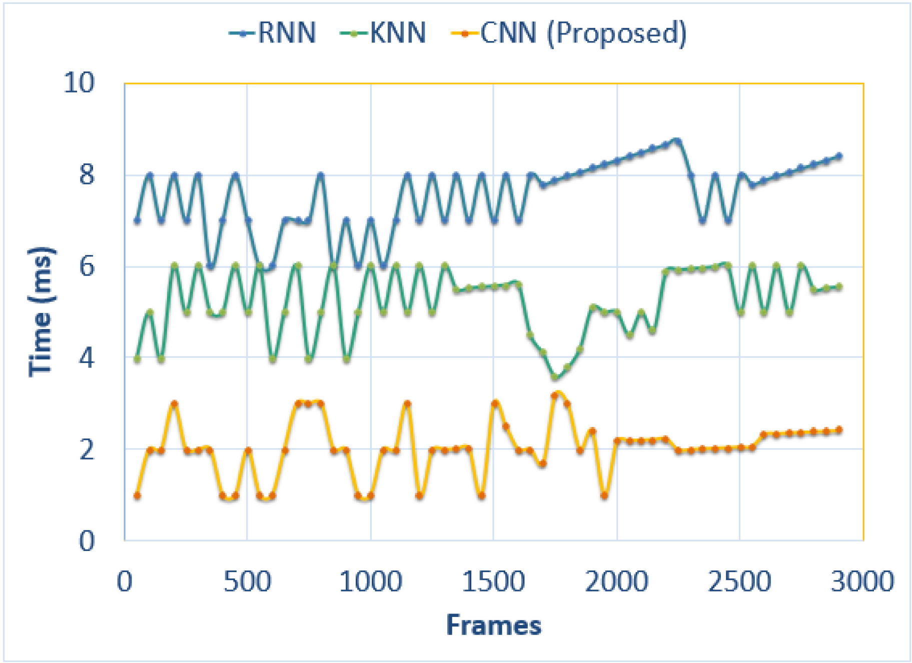


Figure 17. Comparison of the computational efficiency of our proposed model with other state of the art models over the MICC dataset.

4.4. Experiment 3: Performance Evaluation of Heads Counting Over all Three Benchmark Datasets

In this section, the mean head counting accuracy over all three benchmark datasets is shown in Figure 18. The counting accuracy can be calculated as;

$$Count_{heads} = \frac{Predicted\ Number\ of\ Heads}{Actual\ Number\ of\ Heads} \tag{24}$$

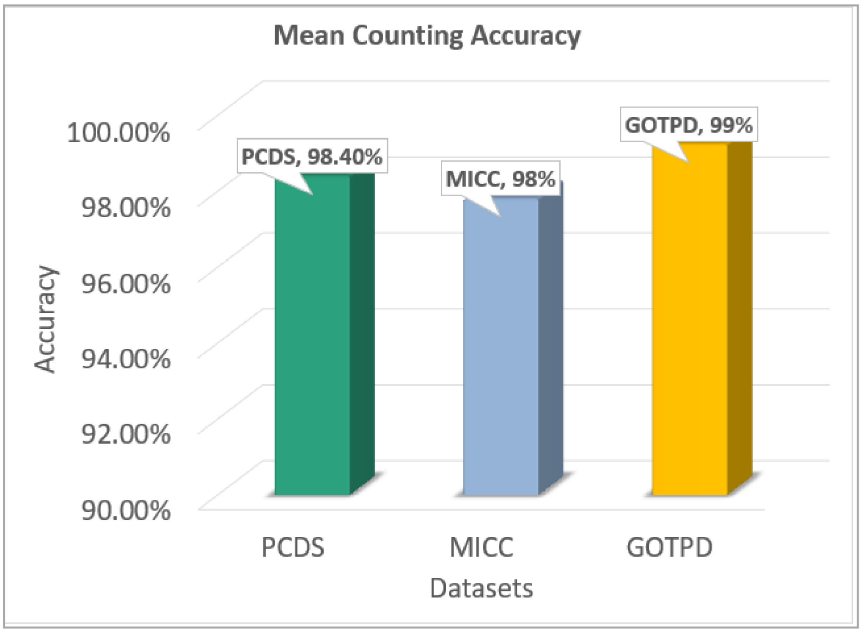


Figure 18. Accumulative head counting accuracies achieved on the three benchmark datasets.

5. Discussion

The automatic HASTAC system is primarily built to track human heads and count the number of heads in depth imagery. We have tested our system on three benchmark datasets i.e., the PCDS dataset, the MICC dataset and the GOTPD dataset. The results acquired on the PCDS dataset and GOTPD dataset exhibit greater accuracy i.e., 98.40% and 99% respectively for head tracking and counting compared to the MICC dataset which produced an acuracy rate of 98%. The reason for getting lower head tracking and counting accuracy on the MICC dataset is the overlapping persons in the Queue condition when the persons are walking one after the other; our system causes some misdetection of heads which results in low tracking and counting accuracy. Secondly, the challenges we face during the processing of the MICC dataset is that it provides noisy depth images with greater illumination and complex backgrounds which hinders the detection of heads in some images. The heads in some of the images of this dataset are overlapped with the noisy background which results in mis-tracking and mis-counting of heads.

6. Conclusions

We have proposed an efficient method for head tracking and counting which works well despite variations in occlusion, illuminaton and background complexity. The proposed HASTAC model is sub-divided into six domains. First, using Kernel Density Function (KDE), preprocessing is done to remove complex backgrounds under varying illumination conditions. Second, head detection is achieved using Hough Circular Gradient Transform and shoulders detection is achieved using the HOG based symmetry method. Third, robust features, like Fused joint HOG-LBP, Energy based point clouds and Fused intra-inter trajectories are extracted which further help the tracking of the heads and shoulders of each individual detected in the video frames. These features are then projected to the Apriori-Association technique which can find frequent itemsets and remove them to get ideal results for head tracking. Fifth, these features are then passed to our proposed 1D CNN model in order to distinguish between heads and non-heads in the video frames. In the last step, the cross line judgement technique is used to count the number of heads frame by frame. We evaluated the proposed model on three publicly available datasets and found that the results are very promising with respect to accuracy and computational time.

6.1. Theoretical Implications

Our proposed HASTAC model has various practical applications in security, visitor counting, queue management etc. Our system is efficient and applicable for people counting and crowd analysis etc.

6.2. Research Limitations

Due to the complex backgrounds and varrying illumination conditions of the MICC dataset, we faced minor challenges in the processing and we achieved less accurate results for this dataset in the Queue condition compared to other conditions, namely, the Flow and Group conditions. Similarly, due to overlaps between persons in the video frames in the Queue condition, our system produced some mis-detection of heads in this dataset. In future, we will work on this problem using different head detection, feature extraction techniques and deep learning models to get better results.

Author Contributions: Conceptualization, S.A.R. and K.K.; Methodology, S.A.R. and A.J.; Software, S.A.R.; Validation, Y.Y.G. and M.G.; Formal analysis, M.G., Y.Y.G. and K.K.; Resources, M.G., A.J. and K.K.; Writing—review and editing, S.A.R., Y.Y.G. and K.K.; Funding acquisition, M.G., Y.Y.G. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2018R1D1A1A02085645). Also, this work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 202012D05-02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mahwish, P.; Jalal, A.; Kim, K. Hybrid algorithm for multi people counting and tracking for smart surveillance. In Proceedings of the IEEE IBCAST, Islamabad, Pakistan, 12–16 January 2021.
2. Sun, Y.; Wang, Y.; He, Y.; Hua, Y. Head-and-Shoulder Detection in Varying Pose. In *Advances in Natural Computation. ICNC*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 12–20.
3. Barabino, B.; Di Francesco, M.; Mozzoni, S. An Offline Framework for Handling Automatic Passenger Counting Raw Data. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2443–2456. [CrossRef]
4. Jalal, A.; Mahmood, M. Students' behavior mining in e-learning environment using cognitive processes with information technologies. *Educ. Inf. Technol.* **2019**. [CrossRef]
5. Ahmed, A.; Jalal, A.; Kim, K. A novel statistical method for scene classification based on multi-object categorization and logistic regression. *Sensors* **2020**, *20*, 3871. [CrossRef]
6. Jalal, A.; Kim, Y.; Kim, D. Ridge body parts features for human pose estimation and recognition from RGB-D video data. In Proceedings of the Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Hefei, China, 11–13 July 2014; pp. 1–6.
7. Akhter, I.; Jalal, A.; Kim, K. Pose Estimation and Detection for Event Recognition using Sense-Aware Features and Adaboost Classifier. In Proceedings of the 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), Islamabad, Pakistan, 12–16 January 2021.
8. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
9. Tahir, S.B.; Jalal, A.; Kim, K. Wearable inertial sensors for daily activity analysis based on adam optimization and the maximum entropy Markov model. *Entropy* **2020**, *22*, 579. [CrossRef]
10. Tahir, S.; Jalal, A.; Batool, M. Wearable Sensors for Activity Analysis using SMO-based Random Forest over Smart home and Sports Datasets. In Proceedings of the 3rd International Conference on Advancements in Computational Sciences, ICACS, Lahore, Pakistan, 17–19 February 2020.
11. Gochoo, M.; Tan, T.-H.; Huang, S.-C.; Batjargal, T.; Hsieh, J.-W.; Alnajjar, F.S.; Chen, Y.-F. Novel IoT-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning. *IEEE Internet Things J.* **2019**, *6*, 7192–7200. [CrossRef]

12. Rafique, A.; Jalal, A.; Kim, K. Automated Sustainable Multi-Object Segmentation and Recognition via Modified Sampling Consensus and Kernel Sliding Perceptron. *Symmetry* **2020**, *12*, 1928. [CrossRef]
13. Ahmed, A.; Jalal, A.; Kim, K. Region and decision tree-based segmentations for Multi- objects detection and classification in Outdoor Scenes. In Proceedings of the IEEE Conference on Frontiers of Information Technology, Islamabad, Pakistan, 16–18 December 2019.
14. Lee, M.W.; Nevatia, R. Body part detection for human pose estimation and tracking. In Proceedings of the 2007 IEEE Workshop on Motion and Video Computing, WMVC, Austin, TX, USA, 23–24 February 2007.
15. Antonini, G.; Thiran, J. Counting Pedestrians in Video Sequences Using Trajectory Clustering. *IEEE Trans. Circuits Syst. Video Technol.* **2006**, *16*, 1008–1020. [CrossRef]
16. Topkaya, I.; Erdogan, H.; Porikli, F. Counting people by clustering person detector outputs. In Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014; pp. 313–318.
17. Kong, Y.; Liang, W.; Dong, Z.; Jia, Y. Recognising human interaction from videos by a discriminative model. *IET Comput. Vis.* **2014**, *8*, 277–286. [CrossRef]
18. Nadeem, A.; Jalal, A.; Kim, K. Human Actions Tracking and Recognition Based on Body Parts Detection via Artificial Neural Network. In Proceedings of the 3rd International Conference on Advancements in Computational Sciences, ICACS, Lahore, Pakistan, 17–19 February 2020.
19. Jalal, A.; Kamal, S.; Kim, D. Depth silhouettes context: A new robust feature for human tracking and activity recognition based on embedded HMMs. In Proceedings of the 2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI, Goyang City, Korea, 28–30 October 2015.
20. Moschini, D.; Fusiello, A. Tracking human motion with multiple cameras using an articulated model. In *Computer Vision/Computer Graphics Collaboration Techniques. MIRAGE 2009*; Gagalowicz, A., Philips, W., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009.
21. Barandiaran, J.; Murguia, B.; Boto, F. Real-Time People Counting Using Multiple Lines. In Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria, 7–9 May 2008; pp. 159–162.
22. Fradi, H.; Dugelay, J. Low level crowd analysis using frame-wise normalized feature for people counting. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), Costa Adeje, Spain, 2–5 December 2012; pp. 246–251.
23. Zeng, C.; Ma, H. Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting. In Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2069–2072.
24. Madiha, J.; Jalal, A.; Kim, K. Wearable sensors based exertion recognition using statistical features and random forest for physical healthcare monitoring. In Proceedings of the IEEE International Conference on Applied Sciences and Technology, Bangkok, Thailand, 1–3 April 2021.
25. Quaid, M.A.K.; Jalal, A. Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm. *Multimed. Tools Appl.* **2020**. [CrossRef]
26. Shokri, M.; Tavakoli, K. A review on the artificial neural network approach to analysis and prediction of seismic damage in infrastructure. *Int. J. Hydromechatronics* **2019**. [CrossRef]
27. Pizzo, L.; Foggia, P.; Greco, A.; Percannella, G.; Vento, M. Counting people by RGB or depth overhead cameras. In *Pattern Recognition Letters*; ACM: New York, NY, USA, 2016; pp. 41–50.
28. Jalal, A.; Sarif, N.; Kim, J.T.; Kim, T.S. Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. *Indoor Built Environ.* **2013**. [CrossRef]
29. Chen, C.; Chen, T.; Wang, D.; Chen, T. A Cost-Effective People-Counter for a Crowd of Moving People Based on Two-Stage Segmentation. *J. Inf. Hiding Multimed. Signal Process.* **2009**, *3*, 12–23.
30. Li, G.; Ren, P.; Lyu, X.; Zhang, H. Real-time top-view people counting based on a Kinect and NVIDIA jets on TK1 integrated platform. In Proceedings of the 6th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 468–473.
31. Gao, C.; Liu, J.; Feng, Q.; Lv, J. People-flow counting in complex environments by combining depth and color information. *Multimed. Tools Appl.* **2016**, *75*, 9315–9331.
32. Tingting, Y.; Junqian, W.; Lintai, W.; Yong, X. Three-stage network for age estimation. *CAAI Trans. Intell. Technol.* **2019**. [CrossRef]
33. Rizwan, S.; Jalal, A.; Gochoo, M.; Kim, K. Robust Active Shape Model via Hierarchical Feature Extraction with SFS-Optimized Convolution Neural Network for Invariant Human Age Classification. *Electronics* **2021**, *10*, 465.
34. Khalid, N.; Gochoo, M.; Jalal, A.; Kim, K. Modeling Two-Person Segmentation and Locomotion for Stereoscopic Action Identification: A Sustainable Video Surveillance System. *Sustainability* **2021**, *13*, 970.
35. Jalal, A.; Kim, Y. Dense Depth Maps-based Human Pose Tracking and Recognition in Dynamic Scenes Using Ridge Data. In Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014; pp. 119–124.
36. Li, B.; Zhang, J.; Zhang, Z.; Xu, Y. A people counting method based on head detection and tracking. In Proceedings of the International Conference on Smart Computing, Hong Kong, China, 3–5 November 2014; pp. 136–141.

37. Xu, H.; Lv, P.; Meng, L. A people counting system based on head-shoulder detection and tracking in surveillance video. In Proceedings of the International Conference On Computer Design and Applications, Qinhuangdao, China, 25–27 June 2010; pp. 394–398.
38. Le, M.; Le, M.; Duong, M. Vision-based People Counting for Attendance Monitoring System. In Proceedings of the 5th International Conference on Green Technology and Sustainable Development (GTSD), Ho Chi Minh City, Vietnam, 27–28 November 2020; pp. 349–352.
39. Wang, S.; Li, R.; Lv, X.; Zhang, X.; Zhu, J.; Dong, J. People Counting Based on Head Detection and Reidentification in Overlapping Cameras System. In Proceedings of the International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Jinan, China, 14–17 December 2018; pp. 47–51.
40. Subburaman, V.; Descamps, A.; Carincotte, C. Counting People in the Crowd Using a Generic Head Detector. In Proceedings of the Ninth International Conference on Advanced Video and Signal-Based Surveillance, Beijing, China, 18–21 September 2012; pp. 470–475.
41. Kuo, J.; Fan, G.; Lai, T. People counting base on head and shoulder information. In Proceedings of the IEEE International Conference on Knowledge Engineering and Applications (ICKEA), Singapore, 4–7 June 2021; pp. 52–55.
42. Sun, S.; Akhtar, N.; Song, H.; Zhang, C.; Li, J.; Mian, A. Benchmark Data and Method for Real-Time People Counting in Cluttered Scenes Using Depth Sensors. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3599–3612.
43. Bondi, E.; Seidenari, L.; Bagdanov, A.; Bimbo, A. Real-time people counting from depth imagery of crowded environments. In Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014; pp. 337–342.
44. Song, D.; Qiao, Y.; Corbetta, A. Depth driven people counting using deep region proposal network. In Proceedings of the IEEE International Conference on Information and Automation (ICIA), Macao, China, 18–20 July 2017; pp. 416–421.
45. Khan, K.; Ali, J.; Ahmad, K.; Gul, A.; Sarwar, G.; Khan, S.; Thanh Hoai Ta, Q.; Chung, T.; Attique, M. 3D Head Pose Estimation through Facial Features and Deep Convolutional Neural Networks. *Comput. Mater. Contin.* **2021**, *66*, 1757–1770.
46. Jianzhao, C.; Victor, O.; Gilbert, O.; Changtao, W. A fast background subtraction method using kernel density estimation for people counting. In Proceedings of the 9th International Conference on Modelling, Identification and Control (ICMIC), Kunming, China, 10–12 July 2017; pp. 133–138.
47. Chen, L.; Wu, H.; Zhao, S.; Gu, J. Head-shoulder detection using joint HOG features for people counting and video surveillance in library. In Proceedings of the IEEE Workshop on Electronics, Computer and Applications, Ottawa, ON, Canada, 8–9 May 2014; pp. 429–432.
48. Adebayo, O.; Abdul Aziz, N. Improved Malware Detection Model with Apriori Association Rule and Particle Swarm Optimization. *Secur. Commun. Netw.* **2019**, *2019*, 1–13. [CrossRef]
49. Park, E.; Han, X.; Berg, T.L.; Berg, A.C. Combining multiple sources of knowledge in deep CNNs for action recognition. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, WACV, Lake Placid, NY, USA, 7–10 March 2016.
50. Li, Y.; Liu, S.G. Temporal-coherency-aware human pose estimation in video via pre-trained res-net and flow-CNN. In Proceedings of the International Conference on Computer Animation and Social Agents (CASA), Seoul, Korea, 22–24 May 2017; pp. 150–159.
51. Shehzad, A.; Jalal, A.; Kim, K. Multi-Person Tracking in Smart Surveillance System for Crowd Counting and Normal/Abnormal Events Detection. In Proceedings of the International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 27–29 August 2019; pp. 163–168.
52. Jalal, A.; Khalid, N.; Kim, K. Automatic Recognition of Human Interaction via Hybrid Descriptors and Maximum Entropy Markov Model Using Depth Sensors. *Entropy* **2020**, *22*, 817. [CrossRef] [PubMed]
53. Gochoo, M.; Akhter, I.; Jalal, A.; Kim, K. Stochastic Remote Sensing Event Classification over Adaptive Posture Estimation via Multifused Data and Deep Belief Network. *Remote Sens.* **2021**, *13*, 912. [CrossRef]
54. Uddin, M.; Khaksar, W.; Torresen, J. Facial Expression Recognition Using Salient Features and Convolutional Neural Network. *IEEE Access* **2017**, *5*, 26146–26161. [CrossRef]
55. Basavegowda, H.; Dagnew, G. Deep learning approach for microarray cancer data classification. *CAAI Trans. Intell. Technol.* **2020**, *5*, 22–33. [CrossRef]
56. Jiang, R.; Mou, X.; Shi, S.; Zhou, Y.; Wang, Q.; Dong, M.; Chen, S. Object tracking on event cameras with offline–online learning. *CAAI Trans. Intell. Technol.* **2020**, *5*, 165–171. [CrossRef]
57. Murlidhar, B.; Sinha, R.; Mohamad, E.; Sonkar, R.; Khorami, M. The effects of particle swarm optimisation and genetic algorithm on ANN results in predicting pile bearing capacity. *Int. J. Hydromechatronics* **2020**, *3*, 69. [CrossRef]
58. Shahgoli, A.; Zandi, Y.; Heirati, A.; Khorami, M.; Mehrabi, P.; Petkovic, D. Optimisation of propylene conversion response by neuro-fuzzy approach. *Int. J. Hydromechatronics* **2020**, *3*, 228. [CrossRef]
59. Luna, C.; Losada-Gutierrez, C.; Fuentes-Jimenez, D.; Fernandez-Rincon, A.; Mazo, M.; Macias-Guarasa, J. Robust people detection using depth information from an overhead Time-of-Flight camera. *Expert Syst. Appl.* **2017**, *71*, 240–256.
60. Luna, C.; Macias-Guarasa, J.; Losada-Gutierrez, C.; Marron-Romera, M.; Mazo, M.; Luengo-Sanchez, S.; Macho-Pedroso, R. Headgear Accessories Classification Using an Overhead Depth Sensor. *Sensors* **2017**, *17*, 1845. [CrossRef]