DTV: Detection, Tracking and Validation Framework for Unique People Count

A. Satarupa Mukherjee , B. Nilanjan Ray

satarupa@ualberta.ca
nray1@ualberta.ca

University of Alberta, Edmonton, Alberta, Canada

Abstract

Counting the unique number of people in a video (i.e., counting a person only once while the person is within the field of view), is required in many significant video analytical applications, such as transit passenger and pedestrian volume count in railway stations, shopping malls and road intersections and many others. In this paper, a novel framework is proposed for counting passengers, mainly in a railway station. The framework has three components: people detection, tracking and validation. In the detection step, a person is detected when he or she enters the field of view. Then, the person is tracked by optical flow based tracking until the person leaves the field of view. Finally, the trajectory generated by the tracker is validated through a spatio-temporal validation technique. The number of valid trajectories denotes the number of people. The novelty of the framework is the inclusion of the validation step, which is overlooked by the existing methods. Extensive experiments have been conducted on the datasets having both top views and whole body views of the passengers. Experimental results demonstrate that the proposed framework generates more than 90% accuracy on both types of views. It also detects and tracks persons having different hair colors and wearing hoodies, caps, long winter jackets, carrying bags and more. The proposed algorithm shows promising results also for people moving in different directions.

Keywords: Unique people count, detection, tracking.

I. Introduction

Accurate people count is an important but challenging task in the field of computer vision. It has impact in solving real life applications like traffic management, detection of overcrowded situations in public buildings, tourist flow estimation, surveillance and many others. It is also a significant component in video analytics. By unique people count, we mean the computation of the total number of people in a specific time interval by counting a person only once while the person is present within a field of view (FOV) or a region of interest (ROI) within the FOV.

People count was manually done in the past, which was both labor and cost intensive. So, developing an automatic method for counting passengers has become a crucial issue. Development of automatic people counting systems based on digital image processing techniques has become a popular research topic in the past few years (Chan et al., 2008, Conte et al., 2010, Ge and Collins, 2009, Harasse et al., 2005, Kim et al., 2003, Lempitsky and Zisserman, 2010, Tan et al., 2011, Zeng and Ma, 2010).

There are two types of methods prevalent in literature for people counting - frame based count and unique count. The frame based count methods use extraction of features from individual video frames and count the number of people in each video frame with the help of some machine learning techniques (Chan et al., 2008, Conte et al., 2010, Lempitsky and Zisserman, 2010, Tan et al., 2011). But these methods fail to produce a unique count in a video over an interval of time, as
they estimate the number of people on a frame-by-frame basis and they fail to consider the correspondence among the same person in multiple frames.

The second type of method is unique people count. One of the prevalent approach of different unique count methods, is based on detection and tracking (Harasse et al., 2005, Kim et al., 2003, Zeng and Ma, 2010). In these types of methods, the individual persons are first detected and then they are tracked. The number of tracked trajectories accounts for the estimate of the number of people. These types of methods detect the people first and then track them. Thus, the count is not dependent on individual frames, but on a sequence on frames. The drawback of these methods is that, often they avoid a validation step, where the detected objects or tracked trajectories should be classified as a human or non-human. Consequently, these methods are imprecise. Therefore, this paper aims to develop a framework, which has three major steps: people detection, tracking and validation.

In order to develop the framework, initially, top views of passengers were considered for a Light Railway Transit (LRT) where a Hough circle (Gonzales and Woods, 2008) based detection algorithm was used to detect people, then optical flow (Horn and Schunck, 1981) based tracking algorithm was evoked to track each person detected in a frame, and finally all trajectories resulted from the tracking algorithm were sieved through a spatio-temporal validation technique for classifying the trajectories into human or non-human (Mukherjee et al., 2011). The total number of valid trajectories indicated the total number of people within a specific time interval.

Apart from top views, the second focus of this paper is to work with whole body views of people. In this case, background subtraction method (McFarlane and Schofield, 1995) is proposed for the initial detection of human being, optical flow method (Horn and Schunck, 1981) for tracking and a motion histogram based technique (Davis, 1999) for classifying the trajectories into human or non-human.

The novelty of this framework is the introduction of the validation step where the trajectories are classified into human or non-human after the detection and tracking works are completed. This state of art algorithms [13,20] skip the validation step. Here it is argued that the validation step is crucial as most of the existing detection algorithms produce a significant number of false alarms, which will lead to a wrong count.

Extensive experiments have been carried out on 7000 frames of overhead views and 10000 frames of whole body views of passengers passing through railway stations. Results demonstrate that the performance of the proposed framework is acceptable as well as promising for both the cases of overhead and full body views. Apart from LRT video data, we also run our algorithm on LHI dataset (Chan et al., 2008). This is a publicly available dataset where overhead views of people are captured using 90° camera angles. The entire methodology has been described in details in section 2.

II. Detection-Tracking-Validation (DTV) Framework

In this section we describe our proposed people counting framework: DTV (detection, tracking, validation). The DTV framework has been tested on two major types of datasets - top views and whole body views of people. The input of the proposed algorithm is a sequence of views of passengers and the output is a set of valid trajectories. The number of valid trajectories is the number of people. The framework can be categorized into three consecutive steps-
(a) The object detection step where the trajectories are initiated.
(b) The object tracking step where the trajectories are generated.
(c) The object validation step where the trajectories are classified into a human or non-human.

After execution of these three steps, the number of valid trajectories denotes the total number of people in the given sequence of frames.

First, the approach for top views is discussed in the following section. Next, the experimental setup on the whole body views are provided.

A. Top Views

The first type of dataset on which the proposed framework is tested, consists of top views of passengers (Mukherjee et al., 2011) as shown in Figure 1a. The advantage of working with top views is that, there is no occlusion and a person can be tracked failsafe. But the disadvantage is that, there is less number of features that make the automatic detection process challenging. The different stages of working with top views are described below.

i. Object Detection

Taking into account that there is circularity in the top view of a human body, the whole body of the person have been captured as a circular object. Initially, the frames that do not have any people are removed with an approximate median (AM) based background subtraction method (McFarlane and Schofield, 1995) to speed up the process. Canny's edge detector (Gonzales and Woods, 2008) is used to detect edges on the frames which remain after background subtraction. Circles having radii within 60 to 80 are detected on the edge image using Hough circle method (Gonzales and Woods, 2008). A square template is constructed around the center of the detected circle to denote the object and track it in the following frames Figure 1b. As detection and tracking are performed at each frame, color information is used for distinguishing newly arrived persons and the persons already detected in the previous frame. For each frame, the value of the Bhattacharya coefficient (Fukanaga, 1990) between the color distribution of a newly detected and previously detected person is calculated. If this value becomes very high, then the newly detected person is ignored as he or she is already considered as detected in the previous frame.

Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) is also used for the detection process for comparing with the Hough circle method. In the experiments, as the top views of passengers have a parametric circular shape, the performance of Hough circle method is better than HOG method. Moreover, passengers are having different hair colors, wearing different types of hoodies, caps, long winter jackets, carrying bags etc. These make an irregular shape of the outer body that is very difficult to learn with supervised shape based object detection technique like HOG.

ii. Object Tracking

After an object is detected in a frame, Horn-Schunck optical flow method (Horn and Schunck, 1981) is used to track the center of the object along with its template in the consecutive frames. In the optical flow based tracking method, the average velocity of pixels within a template of the previous frame is calculated. As the template is in the previous frame, its center position is already known. So, the center of the template in the current frame is obtained by adding the
average velocity with the center of the template of the previous frame. Thus, the tracking is done by obtaining the position of the center of the templates in each frame. As a person is detected in a frame, tracking is initiated and it continues as the person moves through the frame. When the person leaves the field of view, tracking is stopped and the trajectory gets generated as shown in Figure 1c. Apart from Horn-Schunck method, other two well-known methods proposed by Lucas-Kanade method (Lucas and Kanade, 1981) and Brox et al. (2004) are also used for tracking. All the three methods show good performance on the top views of the passengers. Since Horn-Schunck technique has fewer parameters, it is used in our proposed DTV framework.

iii. Object Validation

Two types of false alarms are generated from Hough or HOG based object detection technique: – (a) clutter detected as people; (b) duplicates: detecting different body parts of the same person Figure 2. For clutter removal, an approximate median (AM) based background subtraction method (McFarlane and Schofield, 1995) is used in the spatio temporal domain and a measure of overlap of two trajectories is calculated for duplicate removal. The proposed validation framework is named as spatio-temporal validation (STV). Let \( T = \{T_1, T_2, ..., T_i, ..., T_p\} \) be a trajectory generated by the tracking algorithm on \( p \) consecutive frames, where \( T_i \) is the set of pixels contained by the trajectory on \( i \)th frame. Trajectory \( T_i \) belongs to a person correctly if

\[
\sum_{i=1}^{p} \sum_{(x,y) \in T_i} F_i(x,y) / \sum_{i=1}^{p} N_i > LF_i \text{ is the output of AM for frame } i. \]

The value of each pixel \((x,y)\) in \( F_i \) is either 1 or 0 if it belongs to foreground or background respectively. \( N_i \) is the number of pixels contained by the trajectory \( T_i \) on \( i \)th frame. So, the proportion of foreground pixels to the total number of pixels in a whole trajectory is calculated to conclude whether the trajectory actually belongs to a person or not as a greater proportion of foreground pixels indicates that the trajectory belongs to a human being. For duplicate removal, one trajectory \( A \) is considered as duplicate of another trajectory \( B \) if

\[
\frac{\sum_{i=1}^{m} A_i \cap B_i / A_i \cup B_i}{m} > O \text{ where, } m \text{ is the number of common frames for both trajectories } A \text{ and } B. \]

The values of \( L \) and \( O \) are determined empirically.

![Figure 1. Top Views of Passengers](image-url)
B. Whole Body Views

The second type of dataset on which the framework is tested, consists of whole body views of people where the passengers are going down or climbing up the staircase of a LRT station. As the view consists of whole body of the passengers, lots of features are available. But the two major challenges in this situation are occlusion and scaling effect. The scaling effect is due to the fact that the size of the people is decreasing or increasing as they are going down or coming up the stairs. The second challenge, occlusion, is not severe in the dataset, as in most of the cases, people are descending or coming up the stairs one by one. In case of two or more people, there is only partial occlusion. In this paper, occlusion is tried to be avoided by creating a region of interest (ROI) at the bottom of the frame as shown in Figure 3a. When one person enters the region of interest, he/she is detected. If a second person also comes in along with the first person within the region of interest, the second person is not detected initially due to partial occlusion and also because the entire body of the second person does not fit simultaneously with the whole body of the first person within the small ROI. Once the initially detected person is tracked for a few frames and the whole body of the second person fully enters the ROI, the second person is captured Figure 3b. In this way, partial occlusion is avoided. Moreover scaling effect can also be handled in this way as the size of a person does not change much within a small region. The various steps of this methodology are described below.

![Figure 2. Different false alarms on top views of passengers](image)

(a) Clutter detected as person  
(b) Duplicate: A person detected twice

![Figure 3. Whole body views of passengers](image)

(a) Region of interest on current frame  
(b) Partial occlusion handling

![Figure 4. Detection algorithm on whole body views](image)

(a) Input frame  
(b) Output of background subtraction method

Figure 2. Different false alarms on top views of passengers

Figure 3. Whole body views of passengers

Figure 4. Detection algorithm on whole body views
i. Object Detection

In the case of whole body views, background subtraction method is adopted for the object detection process. Initially, foreground pixels are obtained from the background detection process. Different connected components are found out from this foreground image. The connected component having area greater than or equal to an average human size corresponds to a single person Figure 4. A rectangular template is constructed around the center of the detected blob to denote the object and track it in the following frames.

Histogram of Oriented Gradients (HOG) method is also used for the detection process for comparing with background subtraction method. But in this case, the performance of background subtraction method is better as HOG produces a lot of false alarms. Moreover, the background subtraction method is much faster than the HOG method as established in the result section.

ii. Object Tracking

After an object has been detected in a frame, Horn-Schunck optical flow method is used to track the center of the object along with its template in the consecutive frames. In the optical flow based tracking method, the average velocity of pixels within a template of the previous frame is calculated. As the template is in the previous frame, its center position is already known. So, the center of the template in the current frame is obtained by adding the average velocity with the center of the template of the previous frame. Thus, tracking is done by obtaining the position of the center of the templates in each frame. As a person is detected in a frame, tracking is initiated and it continued as the person moved through the rectangular region of interest. When the person left the region of interest, tracking is stopped and the trajectory gets generated.

iii. Object Validation

The false alarms are again of two types for both the background subtraction and HOG detection techniques– (a) clutter detected as people; (b) duplicates: detecting different body parts of the same person Figure 5. For clutter removal, a motion histogram based technique (Davis, 1999) is adopted to classify the trajectories into human and non-human. At first, half of the number of frames is taken for the training process. In the training phase, the motion histograms of all the trajectories are constructed and a support vector machine is trained with these histograms. In the testing phase, the motion histograms of the generated trajectories are constructed and using these histograms and the trained support vector machine, the trajectories are analysed to classify them as human or non-human. For duplicate removal, one trajectory A is considered as duplicate of another trajectory B if they have some overlap between them as described in section 2.1.3.

![Figure 5. Different false alarms on whole body view of passengers](image_url)
III. Results and Discussion

The proposed DTV framework is tested on 7000 frames of top views and 10000 frames of whole body views of the LRT dataset. For top views, comparisons among different methods for each of the detection, tracking and validation steps are demonstrated. In the case of top views, the video has different crowd densities whereas for the whole body views, the crowd density is almost constant.

In addition to the LRT dataset, the framework is also tested on 3300 frames of an university campus (LHI) dataset which consists of top views of students walking along the campus.

A. Results for Top Views

Some visual results of the algorithm on top view sequences for both sparse and dense crowds of the railway dataset are shown in Figure 6. Figure 7 demonstrates the visual results of HOG and Hough based detection algorithms. The quantitative comparison in terms of Accuracy, Recall, Precision and F-measure between HOG and Hough are shown in Figure 8. In case of detection with HOG, first 50% of the total frames has been used for training and the remaining 50% has been used for testing. F-measure combines recall and precision into a single quantity by computing harmonic mean of Recall and Precision. A better performance is indicated by a higher F-measure. It can be concluded from both Figures 7 and 8 that Hough circle method performs better than the HOG method. While performing the performance evaluations, Horn-Schunck method is used for tracking and STV method is used for validation for both HOG and Hough based detection procedures.

In the Hough circle method, there is a noise sensitive parameter. This parameter is the ratio of the number of detected edge pixels to the calculated perimeter of the circle. Recall and precision are computed on 100 randomly selected images for different values of this parameter ranging from 0.2 to 0.9 at an interval of 0.1 and it is found that the recall is 100% but precision is 70% at a threshold value of 0.7. This experiment is shown in Table 1. The threshold value 0.7 is chosen while performing detection with Hough circle method as the recall value is maximum in this case and the precision value increases significantly by introducing the validation step into spatio-temporal domain.

Horn-Schunck method is chosen for tracking in the proposed framework as it has only one tuning parameter. Average time taken in seconds to track a person between two consecutive frames of resolution 480-by-640 by Horn-Schunck method is 1.32 seconds.

The validation step, which is the noteworthy novelty of the algorithm, enhances the performance of the proposed method. Before validation, the recall and precision of the system was 100% and 70% respectively, which was analyzed on a frame by frame basis. After validating the entire trajectories through the spatio-temporal validation process, the recall and precision of the system is 97% and 92% respectively. The values of $L$ and $O$, for clutter and duplicate trajectory removal are chosen to be 0.6 and 0.1 experimentally.

The proposed validation technique has been compared with two other validation techniques viz., motion histogram (MH) (Davis, 1999) and spatio-temporal gradient histogram (STG) (Hua et al., 2010) and their comparisons are illustrated in Figure 9, where it shows that the proposed validation technique outperforms both the MH and STG techniques. Both MH and STG are supervised validation algorithms and first 50% of the total frames are used as training and the remaining 50% are used as testing. Here, it is to be kept in mind that Hough circle method is
used as detection and Horn-Schunck algorithm is used as tracking while evaluating the performances of STV, MH and STG.

People entering from any of the four borders of the frame and moving in any direction can be detected and tracked successfully using the proposed framework as shown in Figure 10.

The algorithm also exhibits excellent performance on LHI dataset. Visual results on LHI dataset are shown in Figure 11. The video is of duration 5 minutes 30 seconds consisting of 3300 frames where overhead views of passengers are captured. Similarly as the other overhead views of the railway station, Hough circle method is used as detection, Horn-Schunck method is used for tracking and STV is used for validation for the LHI dataset.

Accuracy, Recall, Precision and F-measure of the proposed algorithm for top views on both the LRT and the LHI dataset, are demonstrated in Table 2.

![Visual results on top views](image)

**Figure 6.** Visual results on top views

**TABLE 1**

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.89</td>
</tr>
<tr>
<td>0.8</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>0.6</td>
<td>0.91</td>
<td>0.69</td>
</tr>
<tr>
<td>0.5</td>
<td>0.91</td>
<td>0.63</td>
</tr>
<tr>
<td>0.4</td>
<td>0.91</td>
<td>0.56</td>
</tr>
<tr>
<td>0.3</td>
<td>0.91</td>
<td>0.56</td>
</tr>
</tbody>
</table>
TABLE II

PERFORMANCE OF PROPOSED FRAMEWORK

<table>
<thead>
<tr>
<th>Views</th>
<th>Frames</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>Accuracy (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top (LRT)</td>
<td>7000</td>
<td>97</td>
<td>92</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>Whole body (LRT)</td>
<td>10000</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>92</td>
</tr>
<tr>
<td>Top (LHI)</td>
<td>3300</td>
<td>99</td>
<td>96</td>
<td>96</td>
<td>97</td>
</tr>
</tbody>
</table>

Figure 7. Results of HOG and Hough circle based detection

B. Results for whole body views

The framework is tested on 10000 frames of a video where people were walking down or coming up the stairs of a LRT station. The people were moving mainly in two directions and they wore different types of colored dresses. Some of them were even carrying bags with them. Some visual results are illustrated in Figure 12.
Figure 8. Accuracy, Recall, Precision and F-measure for two detection methods: HOG and Hough

Figure 9. Accuracy, Recall, Precision and F-measure for three spatio-temporal based validation methods: motion histogram (MH), spatio-temporal gradient (STG) and proposed spatio-temporal validation (STV)

Figure 10. People moving in different directions. Top row shows people moving in opposite directions and bottom row shows people moving in perpendicular directions
Background subtraction method is chosen for the initial segmentation of human beings. Quantitative comparison between Background Subtraction and HOG based detection methods in terms of Recall, Precision, Accuracy and F-measure are shown in Figure 13. Receiver Operating System (ROC) curves for both the detection methods for different ROI's have been plotted in Figure 14. The height of the rectangular ROI is varied several times and different values of True Positive Rates (TPR) and False Positive Rates (FPR) are observed which generate the points on the ROC curve. It is noticed that the area under the ROC curve for the proposed background subtraction based detection method is 0.83, whereas the area under the curve for the HOG method is 0.74. The greater area under the ROC curve of the proposed detection method demonstrates its superiority over the HOG method. It is also observed that the background subtraction method is faster than HOG method. The time taken by these two detection methods on each frame implemented in Matlab on a desktop (Intel duo 2 core processor, 2GHz and 4 GB RAM) is illustrated in Table 3. The proposed validation step described in section 2.2.3 enhances the performance of proposed method. After passing the entire trajectories through validation process, the recall and precision of the system is 95% and 90% respectively. The recall, precision, accuracy and F-measure of the framework on all the 10000 frames are illustrated in Table 2 which proves the competency of the framework.

**Table III**

<table>
<thead>
<tr>
<th>Detection Algorithm</th>
<th>Background subtraction</th>
<th>HOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)/frame</td>
<td>0.66</td>
<td>2.76</td>
</tr>
</tbody>
</table>
Figure 12. Visual results on whole body views

Figure 13. Quantitative comparison between Background Subtraction and HOG methods

Figure 14. ROC curve for comparison between Background Subtraction and HOG methods
C. Comparison with an existing method

The proposed method is compared with the method proposed by Zeng et al. (2010). In Zeng et al.'s paper, the approach is almost similar that includes only detection and tracking, but not the validation step. The detection is a supervised method where they use both HOG and Local Binary Pattern (LBP) (Ojala et al., 2002) features to detect the head and shoulders of people to avoid occlusion. In tracking, they use a particle filter tracker. So, basically their approach develops a detection-tracking framework while the proposed approach in this paper designs a detection-tracking-validation framework.

Zeng et al.'s method is applied on the same 10000 frames where the proposed framework is implemented. As it is a supervised method, 50% of the total number of frames is used for training and the remaining 50% for testing. As there is no validation step, the false positives cannot be removed; the accuracy of Zeng et al.'s method is 75% whereas the accuracy of the proposed method is 90%. The quantitative comparison of the two methods is illustrated in Figure 15 which proves the superiority of the discussed method. ROC curves for both the methods for different ROI's have been plotted in Figure 16. The height of the rectangular ROI is varied several times and different values of True Positive Rates (TPR) and False Positive Rates (FPR) are observed which generate the points on the ROC curve. It is noticed that the area under the ROC curve for the proposed DTV framework is 0.83 whereas the area under the ROC curve for Zeng et al.'s method is 0.72. The greater area under the ROC curve of the proposed detection method illustrates its superiority over Zeng et al.'s method. It is also observed that the DTV framework is faster than Zeng et al.'s framework. The time taken by these two methods on each frame implemented in Matlab on a desktop (Intel duo 2 core processor, 2GHz and 4 GB RAM) is illustrated in Table 4.

![Figure 15. Quantitative comparison between Proposed method and Zeng et al.'s [20] method](image)

<table>
<thead>
<tr>
<th>People Counting Algorithm</th>
<th>Proposed method</th>
<th>Zeng et al.'s method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)/frame</td>
<td>0.66</td>
<td>2.52</td>
</tr>
</tbody>
</table>

Table IV. RUNNING TIME OF TWO PEOPLE COUNTING ALGORITHMS
Figure 16. ROC curve comparing Proposed method and Zeng et al.’s [20] method

IV. Conclusion and Future work

In this paper, a robust framework is proposed for computing unique people count in video sequences. The framework is consisted of three components: object detection, object tracking and object validation. A person is detected in the object detection step as he enters the image frame, the object tracking step tracks the person as he moves through the frame and as a result, a trajectory is generated. The validation step analyses the trajectories and the total number of trajectories is counted. The number of valid trajectories represents the number of people. The algorithm is experimented on both top views and whole body views of people. The algorithm succeeds in detecting various types of appearances like persons having different hair colors, wearing hoodies, caps, long winter jackets, carrying bags etc. The proposed framework is also smart enough for handling of people entering in the frame from any direction. Although some specific methods are proposed for the different stages of the framework, the framework is not constrained by these methods. Thus this framework is flexible enough for future work.

Our future work includes implementing the algorithm by creating multi-resolution pyramid of the original image frames for better accuracy and also working with various types of pedestrian scenarios. The framework is also intended to be extended for car counting scenarios.

Acknowledgement

The authors would like to thank Dr. Yang Cong for the LHI dataset. The authors also acknowledge the following sources of funding for this work: NSERC, AQL Management Consulting Inc., and Computing Science, University of Alberta.

References


