A Methodology for Evaluation of People Counting Methods based on Video Analysis

V. H. C. Melo, S. S. Almeida, J. C. Mendes, D. Menotti
UFOP - Federal University of Ouro Preto
Computing Department
Ouro Preto, MG - Brazil
Email: {victorhcmelo,susilvaalmeida,jean.mendes,menottid}@gmail.com

***Abstract*—People counting based on video analysis may be very useful for many commercial applications, such as in the monitoring of public spaces or in sporting events. However, the methods in the literature tend to only check if the total counting is correct, independent of where each count happens. In this paper, we propose a methodology for the assessment of people counting methods (PCMs) based on video from cameras in a zenith position. Initially, it is required to manually indicate, in a given video, when each person passes into and out of the counting zone, generating the ground-truth data. From this reference data and the output of a PCM for a given video, we propose a greedy algorithm to solve the problem of matching the best tracked people from the reference to the output counting, as modeled in a bipartite graph. Once the matching is performed, the non-saturated vertices, *i.e.*, people, indicate false positive and negative countings of the method, while the saturated vertices indicate the true positive counting. From these figures, standard measures such as $precision$, $recall$ and $F{-}score$ can be automatically computed and may also identify where errors occur. In addition, the use of this methodology on a PCM brings benefits for comparison purposes and adjusting parameters.**

***Index Terms*—people counting, evaluation methodology, measures, ground-truth data**

I. INTRODUCTION

People detection, tracking and counting may be very useful for many commercial applications, such as in the monitoring of public spaces and in sporting events. Subway stations in large urban centers have heavy traffic of people and can use these systems to measure flow of people and users. The information collected from the counting process helps to identify patterns in vehicle traffic and to monitor the public present at events. In surveillance systems, it can be used to assign accurate numbers of security personnel at key places and designate efficient evacuation plans.

The methods presented in the literature can be divided into three broad categories [1]. In the first category, systems using mechanical counters, such as turnstiles count only one person at a time and may obstruct the paths causing congestion if there is heavy traffic of people. Due to their design, systems based on turnstiles are subject to undercounting because people are able to jump over or pass under. The second category which is composed of systems using sensors, such as infrared beam or heat sensors, while not obstructing the path, suffer from the same undercounting problem due to the overlapping of people. In addition, these systems may have a tendency to overcount since the distinction among people and objects in the counting zone is hard to identify. The third category consists of vision-based systems using cameras, which surmounts the drawbacks of the first category and aids in solving the problems of the second [1], [2], [3]. Computer Vision, Image Processing and Pattern Recognition techniques are used for segmenting, tracking, and, then, counting people from a monitoring camera video within an interest region.

One of the main problems of people counting methods - from thusforth referred to as PCM – consists of the evaluation methodology employed in works published. Many authors use a single measure, such as accuracy, to evaluate their results. Employing a single measure does not allow to distinguish between false alarms for missing persons. In turn, the use of a set of measures, such as precision, recall and F-score, are more informative. However the process of manually counting the number of false positives, false negatives and true positives for computing these measures is exhaustive and impractical for larger videos with many people.

The study of these PCMs face other difficulties common to several works in this digital age, as detailed by [4]. These include: complications in ensuring the validity of research data; restrictions on data sharing that reduce the ability of researchers to verify results and build on previous research; huge increases in the amount of data being generated, creating severe challenges in preserving data for long-term use.

Many authors of vision-based methods partially follow, or do not follow the recommendations of the report, complicating the analysis as regards the reproduction and verification of experiment results. The key information missing is mainly configuration parameters of the PCMs, characteristics of the recordings and the sharing of videos used in the experiments.

In this paper, in order to overcome such difficulties, we introduce an automatic methodology for evaluation of PCMs, aiming at those which employ an overhead camera – also known as zenith positioning [5]. Our proposed methodology can be briefly described in three steps: 1) For a given video, the ground-truth data is manually, just once, produced indicating where (frame) each person passes to and out of the counting zone; 2) From this reference data and the output indicated by a typical PCM having as input a given video, the matching of the best tracked people from the reference to the output counting is established. This matching problem is instantiated as a bipartite graph, and a simple greedy strategy is employed to solve it; 3) Then standard measures such as $precision$, $recall$ and $F{-}score$ can be computed. In addition, we are able to analyze several specific situations of counting from the matching data. As a result, the main advantage of this methodology is to automatically quantify the false positive and negative counts of the method and also to identify where these counting errors occur. Furthermore, several experiments and parameters in order to set up people counting methods can be tested once the reference data is built. In addition, the source code of people counting methods is not required to be evaluated since the evaluation methodology only needs input and output indication data.

In order to validate our evaluation methodology, we employ it on the results of a people counting method [5] using three 10-minute videos. We also analyze the results based on the proposed measures and go on to illustrate some benefits. The original videos used and ground-truth data generated are available in [6].

The remainder of this work is organized as follows. Related works of evaluation methodologies, measures and public video database are briefly discussed in Section II. Section III describes the proposed evaluation methodology. Experiments are presented in Section IV. Finally, in Section V, conclusions and future works are pointed out.

II. Related Works

There are many approaches and hardware setups [7] in people counting systems based on video processing. Camera position is an important issue. A single common camera mounted in a zenithal position (over the head) is the most common and can be found in [8], [1], [9], [10], [11], [12], [13], [14], [2], [15], [5], [16], [17]. In [18], the authors used a stereovision camera, also in a zenithal position. Another option is a multiple camera approach as used in [19], [20]. In this work, we focus on a single common camera mounted overhead.

This kind of system needs to be very accurate, so another important aspect in people counting systems is the result evaluation methods. The simplest way to report results is the accuracy (number, in percentage form, of people counted related to a real number of people) [3], [1], [9], [11], [12], [10], [14], [2], [18], [16], [21], [22], [23]. Few authors, such as in [8] reported results counting a number of false positives, false negatives, and true positives as well as others the use of $precision$, $recall$, and $F{-}Score$ [13], which are defined based on the former figures.

III. The Evaluation Methodology

As mentioned before, our proposed methodology can be divided into three main steps: 1) Ground-truth generation 2) Maximum matching between tracked people of reference and PCM output 3) The measures computed from matching (Figure 1).

In the following subsections, we will explain in detail each of these steps.

TABLE I
EXAMPLE OF REFERENCE GENERATION

$Frame$	**Up**	**Down**
1004	0	1
1019	2	0
1083	-1	0
1113	0	-2
2058	3	0
2067	4	0
2114	-3	0
2150	0	-4

A. Reference Generation

The ground-truth data is generated by analyzing the video as a whole and references are made only for the frames where events are detected. An event is considered as a person passing into or out of the counting zone or region of interest (ROI). Once an event happens, we insert it in a reference table as follows. If the event is someone passing into the ROI, we add the unique ID with positive signal in the corresponding direction column, *i.e.*, **Up** or **Down**. Conversely, if it is a passing out of the ROI, we add the same ID used when that person was passing into the ROI with a negative signal in the corresponding direction column. Note that this convention of *Up* and *Down* can be adapted to videos where the people come (and go) from left to right, and vice-versa, as well. Also observe that two or more people can pass into or out of the ROI in the same frame without loss of generality of our convention. An example of reference generation for 4 people counting is shown in Table I.

B. Matching Problem

The maximum matching of tracked people between the reference and the PCM output works as follows. From the reference and output counting data, we extract a simpler representation. Each tracked person is represented as a triple composed of its ID and the number of the frame when they pass into and out of the ROI. Together with the number of frames in or out we put the direction information. The i-th tracked people from the reference data can be represented as $(RID^i, RF^i_{in}, RF^i_{out})$, whilst the j-th tracked people for the method as $(MID^j, MF^j_{in}, MF^j_{out})$. An example of such a transformation of the reference data representation in Table I is shown in Table II. Note that the output of the PCM evaluated respects the same rules and conventions as those imposed to the reference generation, but the cardinality of these two sets may be different.

Each set of tracked people for R reference and M method can be viewed as disjointed sets where the connection weight

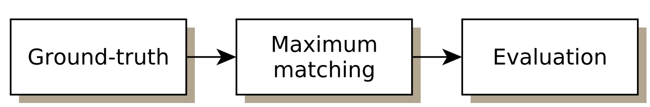

Fig. 1. The three main steps of our proposed methodology

TABLE II
EXAMPLE OF REFERENCE VERTICES (PEOPLE COUNTING) FROM TABLE I USED FOR MATCHING

ID	$Frame_{in}$	$Frame_{out}$
1	-1004	+1083
2	+1019	-1113
3	+2058	+2114
4	+2067	-2150

between the elements of these sets are proportional to their overlapping in time domain and the problem of matching of tracked people from the reference to the method can be modeled as the following graph. For computing the edge weight, W_{ij}, between RID^i and MID^i we propose to take into account their intersection and union time intervals, *i.e.*,

$$W^{ij} = \frac{|RInt^i \cap MInt^j|}{|RInt^i \cup MInt^j|} \tag{1}$$

where $RInt^i$ and $MInt^j$ stand for the time interval of RID^i and MID^j tracked people, respectively, and $0 \leq W^{ij} \leq 1$. When the direction of the movement of the people tracked for RID^i and MID^j are different, zero is assigned to its edge weight. From this definition, we see that the more the intersection time interval between RID^i and MID^j, the more the edge weight is; and, in contrast, the more the union time interval between RID^i and MID^j, the less the edge weight is. Figure 2 illustrates a graph resulting from this procedure, where darker edges stand for greater edge weight, while lighter edges stand for smaller values.

Our matching problem may be directly related to the maximum matching in bipartite graphs or the maximum correspondence in a net flow [24]. In this specific case, we prefer a solution that obtains the largest number of matches from the reference vertices to the output method. Despite the fact that this problem has combinatorial solution space, there are algorithms available in the literature for solving these problems in polynomial time, *i.e.*, $O(n^3)$ time complexity [25]. Nevertheless, we propose a greedy algorithm for selecting the best tracked people matching from the reference to output counting. After computing the edge weight for all possible pairs of RID^i and MID^j for $0 < i < |RID|$ and $0 < j < |MID|$, where $|RID|$ and $|MID|$ stand for the cardinality of RID and MID, respectively, we sort the edges in descending order by their weight. Once sorted, we have removed from this pool the edges W^{ij} with the greatest weight and if both of its vertices do not correspond to a match, the match between them are established. This process is repeated until there is no more freedom of matched vertices in R or in M or until there are no more edges to analyze (see Algorithm 1).

C. Measures

At the end of the process, all possible matching from R to M were made. The non-matched vertices from R and M sets are directly considered as false negatives (FN) and positives (FP), respectively, in counting evaluation.

From the number of matched vertices from R to M (true positives), the number of non-indicated vertices in R (FN) and in M (FP), we can use three very useful measures, *i.e.*, $precision$, $recall$, and $F{-}score$.

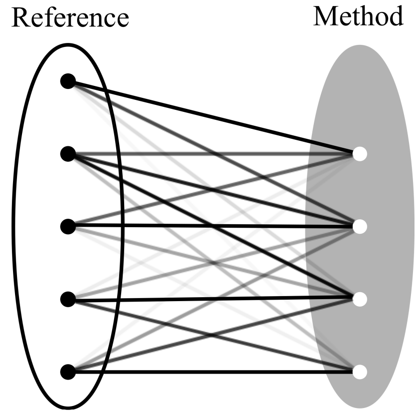

Fig. 2. Bipartite graph representing the matching problem. Each edge has a weight based on the Equation 1. Edges with less opacity have lower weight; more opacity means greater weight

$$precision = \frac{TP}{TP + FP}, \tag{2}$$

$$recall = \frac{TP}{TP + FN}, \tag{3}$$

and

$$F{-}score = \frac{2 \times precision \times recall}{precision + recall}. \tag{4}$$

As stated, this is not the first time in people counting methods literature that have been used to measure and to evaluate counting results. However, according to our evaluation methodology, we are able to automatically determine in which situations these errors occur (number of people present in the counting zone).

In order to automatically compute these situations from the reference data, we first have to compute the expected number of people in the counting zone at each frame. This information can easily be estimated by accumulating the interval time of each tracked person in a total time vector. Once this total time vector is computed, we are able to obtain the expected number of people for each tracked person in the reference data and in output methods by taking the maximum frequency in the total time vector for its own interval time. As we are aware of FN and FP sets, for building the histogram of situations we may

Algorithm 1 GREEDY-MATCHING

Require: A bipartite graph $G = (V, E)$.
Ensure: A matching M between the vertices.

SORT-EDGES-DESCENDING(G)
for all $(u, v) \in E$ not matched **do**
  Insert (u, v) in M
  Mark u and v as matched
end for
return M

TABLE III

VARYING PARAMETER OF THE PEOPLE COUNTING METHOD AND ITS IMPACT IN THE EVALUATION OF THE RESULTS FOR DIFFERENT VIDEOS

		Number of People			Measures (%)		
Video	D_{min}	TP	FP	FN	$prec.$	$recall$	$F{-}score$
stm1	**50**	29	6	4	82.9	87.9	85.3
	65	28	4	5	87.5	84.8	86.2
	80	26	0	7	100.0	78.8	88.1
stm2	**50**	16	25	0	39.0	100.0	56.1
	65	15	17	1	46.9	93.8	62.5
	80	15	15	1	50.0	93.8	65.2
stm3	**50**	21	25	1	45.7	95.5	61.8
	65	21	18	1	53.8	95.5	68.9
	80	21	10	1	67.7	95.5	79.2

TABLE IV

DIFFERENT NUMBER OF PEOPLES IN ROI AND NUMBER OF TRUE POSITIVES (TP), FALSE NEGATIVES (FN), AND FALSE POSITIVES (FP) USING $D_{min} = 50$ IN THE IMPLEMENTED PCM

Video		Number of People				
		0	1	2	3	Tot
stm1	**TP**	x	15	12	2	29
	FP	0	4	2	0	6
	FN	x	1	2	1	4
stm2	**TP**	x	14	2	0	16
	FP	5	20	0	0	25
	FN	x	0	0	0	0
stm3	**TP**	x	14	7	0	21
	FP	0	21	4	0	25
	FN	x	0	1	0	1

simply accumulate each error in its corresponding position in the histogram. A similar process can be executed to obtain the situations where TP tracked people occur.

IV. EXPERIMENTS

In order to validate our evaluation methodology, we implemented in MATLAB a PCM [5] and ran it on three videos recorded using different parameters of the method. We therefore employed our methodology on the output of the PCM and analyzed the results obtained. We also illustrate how the methodology may identify where errors occur for the output by the PCM. Section IV-A briefly explains the implemented method and the discussion of the results is presented in Section IV-B.

A. The People Counting Method

The implemented PCM used here is designed to work in a system with a zenith camera [5], which can be described as follows. From the initial block-wise background subtraction, k-means clustering is used to provide the segmentation of isolated people in the scene. The number of people in the scene is estimated as the maximum number of clusters with acceptable inter-cluster separation. Tracking of segmented people is addressed as a problem of dynamic cluster assignment between two consecutive frames and is solved in a greedy fashion.

B. Experiments

The videos used in the experiments have a resolution of 640×480 pixels, 30 fps, 10 minutes each and are available at [6]. The videos are named `stm1`, `stm2`, and `stm3` and the number of people passing through their ROI are 33, 16, and 22, respectively. Observe that all the values presented in this section were automatically generated by our methodology.

Apart from automatically evaluating the result of a PCM applied on a given video, the proposed methodology can be used to help the choice of the parameter setting of a PCM. To illustrate this task, we chose to evaluate the output generated by the implemented PCM by varying its main parameters, *i.e.*, the minimum allowable intercluster distance, D_{min}, which is related to the average size in pixels of a person, using the three videos. We used three values for parameter $D_{min} = \{50, 65, 80\}$ in the experiment. Table III presents, in detail, the number of true positive (TP), false negative (FN) and false positive (FP) countings and the measures computed for these parameters of the PCM, automatically obtained by the proposed evaluation methodology.

Observe that the sum of TP and FN values is constant for each video, since it represents the real/expected number of people passing through the video, *i.e.*, the ground-truth, while the sum of TP and FP values varies for each video depending on the D_{min}, since this sum represents the number of counted people for the PCM. By observing the figures of Table III, it is noteworthy that the values of $precision$, which depends directly on the FP value, and consequently $F{-}score$ measures are smaller for the `stm2` and `stm3` videos than the ones of the `stm1` video, while the values of $recall$ measures, which depends directly on the FN value, are not so sensitive for the three videos as $precision$. This result can be explained by sudden illumination changes which happen more often in the `stm2` and `stm3` videos and cause the appearance of more false alarms in segmentations steps, and it can be noted that our implementation of PCM is quite sensitive to background changes. Despite this fact, we can observe that in the three videos TP and FN decreases as D_{min} increases, whilst FP increases as D_{min} decreases. This knowledge about the PCM can be used to its setup according to the user system requirements that can prefer less FP to more FN, or to the contrary. And this information been obtained by analyzing the values in Table III that are automatically taken from our proposed evaluation methodology, without the need of manually recounting each output of the implemented PCM.

Moreover, the proposed evaluation methodology can be used to identify where the counting errors occur. That is, where a FP or FN happens how many people we expected to have in the scene. As already claimed at the end of Section III, it is straightforward to compute these data.

In order to illustrate this benefit, Table IV presents in details the number of true positive (TP), false negatives (FN) and false positives (FP) counting for different numbers of people in the videos using $D_{min} = 50$. In the interval time (set of frames) where we have no people (0 people), it is not possible to have any TP and FN, then we fulfill this entry in table with 'x'. By analyzing the figures of FP and FN is possible to know where the counting errors occur and try to change the PCM implementation for overcome such difficulties.

V. Conclusions

In this paper, we have proposed a methodology for automatic evaluation of people counting methods. Our evaluation methodology was made possible by automatically quantifying the true positive, false positive, and negative counting, in addition to identifying when the errors occur (*i.e.*, the number of people presented in the counting zone).

The proposed methodology can automatically evaluate situations of varying people speed, such as normal, fast and abrupt by means of thresholds. Moreover, it is possible to establish the situations where uni-directional and bi-directional people movements occur. However, our methodology may not evaluate situations where the people passing are separated or merge-splitting takes place, since we do not take into account the real positions of people during his/her passing through. We propose as possible future directions of work to overcome this drawback by perfectly segmenting each person during his/her passing through, or more approximately, by indicating his/her mass center during tracking. In this last case, the separating or splitting-merge situation can be estimated using average values for people sizes. It is noticeable that in both cases, the working time required to label the videos increases considerably.

We are in the finishing stages of implementing two more people counting method [13], [14] and planning to implement five others [1], [17], [16], [26], [20], [7], [27] in order to perform an extensive evaluation of the state-of-the-art methods, in terms of results by using the proposed methodology here. In addition to these methods, we plan to disclose about two dozen videos: longer (1 hour, for example), as well as collected in different places, using different cameras and illumination conditions. Moreover, the number of people in the counting zone is required to collect varying videos. In other words, from rush hour, where we can easily have five or more people in the counting zone to the very passive moments, *e.g.*, one people per minute. From such videos, we believe that a more realistic evaluation of the methods proposed in the literature may be performed.

Acknowledgment

The authors would like to thank FAPEMIG/PROPP/UFOP for financing this study.

References

[1] S. Velipasalar, Y.-L. Tian, and A. Hampapur, "Automatic counting of interacting people by using a single uncalibrated camera," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2006, pp. 1265–1268.

[2] C.-H. Chen, Y.-C. Chang, T.-Y. Chen, and D.-J. Wang, "People counting system for getting in/out of a bus based on video processing," in *International Conference on Intelligent Systems Design and Applications (ISDA)*, 2008, pp. 565–569.

[3] D. Huang and T. W. S. Chow, "A people-counting system using a hybrid rbf neural network," *Neural Processing Letters*, vol. 18, pp. 97–113, 2003.

[4] D. A. Committee, *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. National Academy Press, 2009. [Online]. Available: http://www.nap.edu/catalog.php?record_id=12615

[5] B. Antic, D. Letic, D. Culibrk, and V. Crnojevic, "K-means based segmentation for real-time zenithal people counting," in *IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2565–2568.

[6] "Codes, databases and ground-truth data," available at http://www.iceb.ufop.br/decom/lapdi/peoplecounting/index.html.

[7] L. Gasparini, M. Gottardi, N. Massari, D. Petri, and R. Manduchi, "FPGA implementation of a people counter for an ultra-low-power wireless camera network node," in *Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, 2011, pp. 169–172.

[8] J. Bescos, J. M. Menendez, and N. Garcia, "DCT based segmentation applied to a scalable zenithal people counter," in *IEEE International Conference on Image Processing (ICIP)*, vol. 3, 2003, pp. 1005–1008.

[9] H. Septian, J. Tao, and Y.-P. Tan, "People counting by video segmentation and tracking," in *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2006, pp. 1–4.

[10] M. Bozzoli and L. Cinque, "A statistical method for people counting in crowded environments," in *IEEE International Conference on Image Analysis and Processing (ICIAP)*, 2007, pp. 506–511.

[11] J.-W. Hsieh, C.-S. Peng, and K.-C. Fan, "Grid-based template matching for people counting," in *IEEE Workshop on Multimedia Signal Processing (MMSP)*, 2007, pp. 316–319.

[12] H. Yu, Z. He, and J. Liu, "A vision-based method to estimate passenger flow in bus," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2007, pp. 654–657.

[13] J. Barandiaran, B. Murguia, and F. Boto, "Real-time people counting using multiple lines," in *International Workshop on Image Analysis for Multimedia Interactive Services (IAMIS)*, 2008, pp. 159–162.

[14] S. Yu, X. Chen, W. Sun, and D. Xie, "A robust method for detecting and counting people," in *IEEE International Conference on Audio, Language and Image Processing (ICALIP)*, 2008, pp. 1545–1549.

[15] Y. Xin, G.Sun, and Q. Wu, "A preprocessing method for tracking and counting pedestrians in bus video monitor," in *IEEE International Conference on Industrial Informatics (INDIN)*, 2008, pp. 1689–1693.

[16] T.-Y. Chen, T.-H. Chen, and D.-J. Wang, "A cost-effective people-counter for passing through a gate based on image processing," *International Journal of Innovative Computing, Information and Control (ICIC International)*, vol. 5, no. 3, pp. 785–800, 2009.

[17] K. Jaijing, P. Kaewtrakulpong, and S. Siddhichai, "Object detection and modeling algorithm for automatic visual people counting system," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, vol. 2, 2009, pp. 1062–1065.

[18] T. Yahiaoui, C. Meurie, L. Khoudour, and F. Cabestaing, "A people counting system based on dense and close stereovision," in *International Conference on Image and Signal Processing (ICISP)*, vol. 5099 (LNCS), 2008, pp. 59–66.

[19] R. Muoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato, "Multi-camera people tracking using evidential filters," *International Journal of Approximate Reasoning (IJAR)*, vol. 50, no. 5, pp. 732–749, 2009.

[20] A. Belbachir, S. Schraml, and N. Brandle, "Real-time classification of pedestrians and cyclists for intelligent counting of non-motorized traffic," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 45–50.

[21] D. Lamovsky and R. Sadykhov, "Method of pedestrians traffic assessment based on analysis of video data in surveillance systems," in *Proceedings of the 33rd International Convention MIPRO*, 2010, pp. 704–706.

[22] J. Li, L. Huang, and C. Liu, "Robust people counting in video surveillance: Dataset and system," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2011, pp. 54–59.

[23] X. Wang, C. Wang, and J. Yao, "A heuristic information based system for people counting," in *International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 1, 2011, pp. 22–26.

[24] J. A. Bondy and U. S. R. Murty, *Graph Theory With Applications*. Elsevier Science Ltd, 1976, iSBN-13: 978-0444194510.

[25] D. Jungnickel, *Graphs, Networks and Algorithms*, 3rd ed. Springer Publishing Company, Incorporated, 2007.

[26] S. Mukherjee, B. Saha, I. Jamal, R. Leclerc, and N. Ray, "Anovel framework for automatic passenger counting," in *IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2969–2972.

[27] Y. Benabbas, N. Ihaddadene, T. Yahiaoui, T. Urruty, and C. Djeraba, "Spatio-temporal optical flow analysis for people counting," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 212–217.