Evaluation of People Counting Systems

T. Schlögl[1], B. Wachmann[2], W. Kropatsch[3], H. Bischof[3]

[1]Advanced Computer Vision, Wohllebengasse 6, A-1040 Vienna, Austria
[2]Siemens AG Österreich, Programm- und Systementwicklung, Graz, Austria
[3]Pattern Recognition and Image Processing Group, Vienna Technical University, Favoritenstr. 9/1832, A-1040 Vienna, Austria

This paper addresses the problem of the evaluation of people counting systems. The relation between the real people number and the output of the people counting system is used to motivate the computation of the maximum number of people as a scenario and geometry specific quantity which supports the evaluation of the system output. Based on the camera field-of-view we determine the maximum number of detectable people using a basic people model. The maximum number of people is computed for an overhead camera and for an arbitrary oblique camera orientation which is the typical geometry of standard surveillance applications.

1 Introduction

Vision-based real-time people counting comprises all techniques which are able to extract the number of people which are present in an observed area of the real world and which satisfy certain constraints of accuracy and performance. This paper emphasizes the idea that scenario- and geometry specific quantities can rise the accuracy of conventional people counting applications.

Recent research in people counting can be divided into techniques using neural-based crowd estimation ([1], [2], [3]) and methods which are based on blob detection and blob tracking ([4], [5], [6]). Both approaches are rather different from each other. The former employ simple image processing techniques for the extraction of significant features and feed those into a trained neural network to obtain an estimation value of the people in the scene. The accuracy of those systems depends strongly on the training set of the neural network and on the choice of the feature set. Typical elements of the feature set are the length of the edges and the background area ([1], [2]). People counting by blob tracking is currently an option of systems which are mainly aimed to classify objects and their activities by analyzing their shapes and their trajectories. Those systems are based on the capability to separate an object from the background. In scenarios where people are crowded, those systems have a poor accuracy since precise object extraction is hardly possible [5].

2 Output Accuracy

The output accuracy of a people counting application is evaluated by the relation between the true number of people N_{real} and the mean output of the people counting system N_{out}. Examples for those relations are illustrated in Figure 1. The straight line between the origin and the point (n_{max}, n_{max}) is the ideal case, of course. The diagram illustrates that an overall error in percent is no reasonable measure. Obviously, an error of 30 percent near n_{max} can cause a people number which even exceeds the maximum value. Curve *a* shows the optimum case of a people counting system which yields always the correct number. Curves *b* and *c* show possible variations.

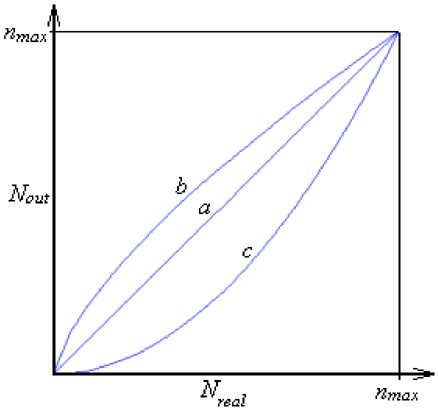

Figure 1: *Typical relations between mean output and real people number obtained from people counting systems: Every line represents a relation between the mean output of a people counting system and the real number of people in the viewing area.*

Figure 1 incorporates two constraints which can be used for the evaluation of current people counting techniques:

1. If there are no people in the viewed area, the people counting system must yield zero.
2. For every viewing area, there is a maximum number of people n_{max} who can gather there. The counting system must never exceed this value.

Constraint 1 is trivial and motivated naturally. Most current visual surveillance systems do provide information to human operators concerning the occurence of people within an area. Thereby, the decision whether no person or at least one is present, has a high priority and is of particular importance in security applications.

Constraint 2 includes the requirement that the maximum number of people n_{max} of a scenario must be known. Of course, the number is closely related to the chosen people model. The use of this essential number in people counting applications has not been published so far.

2.1 Neural-based crowd estimation

Constraint 1 is a critical point for people counting methods based on neural network estimators. They can hardly cope with this constraint since the image processing for feature extraction does not classify edges or blobs at all. Thus, they are sensible to other objects than people, shadows and ambient illumination changes. E. g., in [2], the evaluation of the proposed methods was not performed in respect to this condition. The maximum number of people in the scene is used implicitly in the training set. Good training data are spread over the whole range of possible numbers of people. Thus n_{max} is implicitly used during training, when images with the maximum density are fed into the training algorithm. But generally, there is no proof that n_{max} can not be exceeded by performing the estimation algorithm for images not in the training set.

2.2 Blob techniques

The capability of blob techniques to fulfill both constraints depends on the reliability and robustness of the realized blob detection and blob classification. Since blobs are analyzed based on their size and their shape [5], blob techniques have the general capability to separate humans from other objects. Thus, it is expectable that they fulfill constraint 1 more likely than neural network estimators. Similarly, constraint 2, the maximum number of people n_{max} must not be exceeded which can be evaluated by counting the blobs and comparing it to the theoretical value.

2.3 Parametric human body model

The computation of the maximum number assumes the definition of a person model. Figure 2 illustrates the people model which is used in this paper.

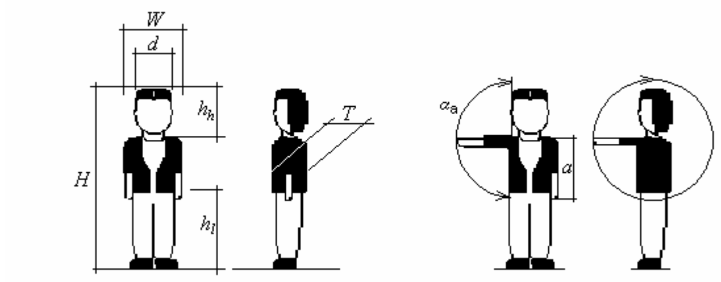

Figure 2: *People model: Dimensions used in the computation of n_{max}: H=tallness, W=shoulder width, T=chest depth, D=head diameter, h_h=head height, h_l=leg length, a=arm length, α_a=angle between body and head*

3 Computation of the maximum number of people

The size of a rectangular area of the video sensor is assumed with S_W x S_B pixels (the dimension of a pixel element is s_W *x* s_T). This area might be the whole sensor area, but can also cover only a smaller part like a bounding box of a blob. First, we do this for the special geometry of an overhead camera and generalize it later to an arbitrary oblique camera orientation.

3.1 Overhead geometry

This section demonstrates how the minimum number of pixels of the sensor and the maximum number of people can be estimated. All movements of the arms and the legs require more space for a person and reduce the overall number of people. Hence, the motion of arms and legs is neglected.

As Figure 3 illustrates, assuming that there is no sudden vertical movement (e. g. jumping) of a person the actual distance between the closest point of a person and the camera sensor is simply the difference between *h* and H_{max}. $\vartheta_{\{W,B\}}$ are the viewing angles of the lens of the camera parallel to the side lengths of the sensor. $S_{\{W,B\}}$ denotes the size of the sensor.

$$\vartheta_{\{W,B\}} = 2 \arctan \frac{S_{\{W,B\}}}{2f} \tag{1}$$

B_W and B_T are the projection lengths of the shoulder width *W* and the chest depth *T* respectively. Both lengths determine the bounding box of the projected upper body within the image plane. The minimum size of that person, approximated as a rectangle, is evaluated according to the basic projection equations [7]. *f* denotes the focal length.

$$B = \frac{object\ size. f}{object\ distance - f} \tag{2}$$

If the object distance is much greater than the focal length, the denominator is approximately equal to the object distance.

$$B_W = W \frac{f}{h - H} \tag{3}$$

$$B_T = T \frac{f}{h - H} \tag{4}$$

Those lengths are converted into pixel dimensions using equations (6) and (7). The use of the maximum value of both side lengths ensures a minimum number of pixels (Equation (8)). In image analysis the sampling interval should be less than or equal to half of the smallest interesting detail in the image in both dimensions [8]. This means that the required number of pixels representing an object must be doubled (in one dimension) due to the sampling theorem.

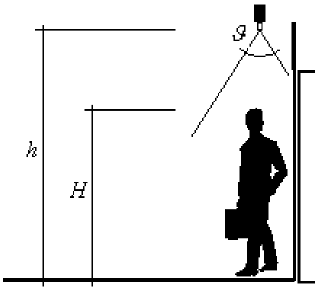

Figure 3: *Camera orientation of an overhead scenario: The geometry is determined by two geometrical parameters, h and ϑ.*

$$s_{\max} = \max(s_W, s_B) \tag{5}$$

$$n_W^{overhead} = 2.\left\lceil \frac{B_W}{s_{\max}} \right\rceil \tag{6}$$

$$n_T^{overhead} = 2.\left\lceil \frac{B_T}{s_{\max}} \right\rceil \tag{7}$$

The desired value of the maximum number of people in an overhead scenario follows as a ratio of the rectangular sensor area and the minimum area of a single person.

$$n_{\max}^{overhead} = \left\lceil \frac{S_W . S_B}{n_W^{overhead} . n_T^{overhead}} \right\rceil \tag{8}$$

3.2 Oblique geometry

The variety of the spatial conditions of actual realities in potential locations like waiting areas of supermarkets and banking halls makes it difficult to make any preconsiderations of the components. In contrast to a vertical camera orientation, the distance to the separate persons may vary.

Thus the required resolution is primarily determined by a person observed at the maximum distance *L* which is determined by the scene geometry. Substituting *L* instead of *h-H* into equations (3) and (4), the resulting value is the estimated maximum.

Equations (5), (6) and (7) allow the computation of the minimum size of a person onto the sensor.

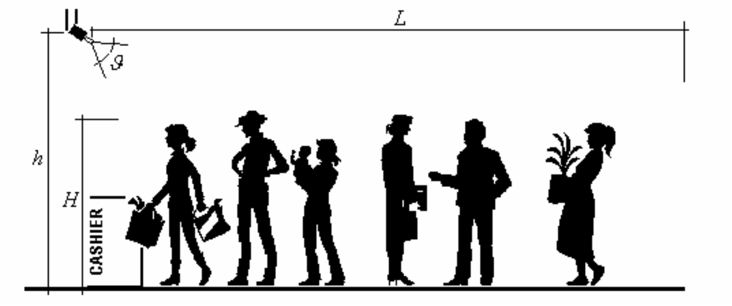

Figure 4: *Oblique camera orientation: Monitoring a queue*

$$n_W^{oblique} = 2.\left\lceil \frac{W.f}{s_{\max} L} \right\rceil \tag{9}$$

$$n_T^{oblique} = 2.\left\lceil \frac{T.f}{s_{\max} L} \right\rceil \tag{10}$$

T is the chest depth, *s* is the size of the video sensor and *L* is the maximum length of the queue shown in Figure 4. Replacing T_{min} by the corresponding minimum values for shoulder width and tallness equation (9) yields the minimum size of a projected person. Thus, the maximum number of people within the rectangular area of the sensor S_W x S_B is computed according to equation (11).

$$n_{\max}^{oblique} = \left\lceil \frac{S_W.S_B}{n_W^{oblique}.n_T^{oblique}} \right\rceil \tag{11}$$

4 Experimental Results

We implemented a motion detection algorithm based on an adaptive background and threshold model. The application of intensity profile analysis as proposed in [9] allows the detection of stationary and transient objects. The blob information was generated by morphological operations. The parameters of the people model (Figure 2) were used to classify the stationary and transient blobs into two classes: people and other objects. The relevant parameters of the people model were set to *H*=100...210 cm and *W*=30...80 cm. All blobs which have the appropriate size were counted for the overall people number. No blob analysis or correlation between successive frames was done.

This algorithm was applied to two test videos recorded at an oblique camera geometry in a railway station in Vienna. Figures 5a and 5b show two snapshots where the people density is very low and where constraint 1 is evaluated. The abscissa indicates the frame number of the video sequence while the ordinate gives the overall people number. Figure 5b illustrates that the person in the lower part of the scene does not fulfill the people model and is therefore not detected as human.

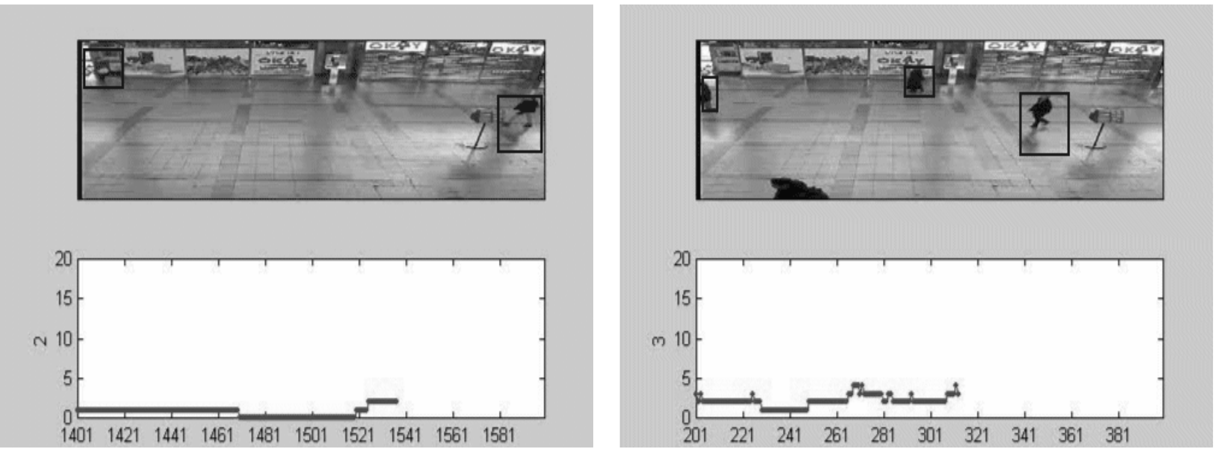

Figure 5: *Detection of people at low densities: (a) Proper detection of two people, (b) critical case where a person in the foreground is not detected properly*

Figure 6 shows the application of the algorithm for higher people densities. The bounding boxes illustrate that the algorithm is not able to distinguish people which occlude each other properly. This is not surprising since no blob analysis was performed. The maximum number of people which can be found within such a blob is computed using equation (11). This value is used as a reference value for an estimator of the people number in this blob.

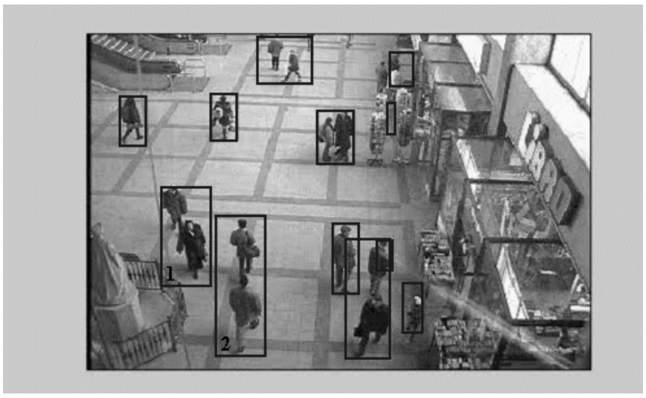

Figure 6: *Blob detection: The sizes of the bounding boxes of blobs 1 and 2 indicate that more than a single person is contained. Equation (11) allows the computation of the maximum number of people for every single blob.*

The camera was calibrated by measuring the bounding box in pixels of an object of known dimension placed at the nearest and farest point of the viewing area. The computation of the size of each detected blob S_W x S_B was interpolated linearly. Similarly, length L of equations (9) and (10) was linearly approximated from the total depth of the viewing area.

5 Conclusions

We described how the two current approaches of automatic people counting algorithm deal with our suggested constraints for the evaluation process. Based on the camera field-of-view and a basic people model, it was shown how the maximum number of detectable people is computed for overhead and oblique camera orientations.

6 References

[1] C. S. Regazzoni, A. Tesei, „Distributed data fusion for real-time crowding estimation“, Signal Processing 53, pp. 47-63, 1996.

[2] S.-Y. Cho, T. W. S. Chow, C.-T. Leung, „A Neural-Based Crowd Estimation by Hybrid Global Learning Algorithm“, IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 29, No. 4, August 1999, pp. 535-541.

[3] T. W. S. Chow, J. Y.-F. Yam, S.-Y. Cho, „Fast training algorithm for feedforward neural networks: application to crowd estimation at underground stations“, Artificial Intelligence in Engineering, Vol. 13, pp. 301-307, 1999.

[4] V. Kettnaker, R. Zabih, „Counting People from Multiple Cameras“, IEEE International Conference on Multimedia Computing and Systems, Vol. 2, pp. 267-271, 1999.

[5] I. Haritaoglu, D. Harwood, L. S. Davis, „W^4: Real-Time Surveillance of People and Their Activities“, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.8, August 2000, pp. 809-830.

[6] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, L. Wixson, „A System for Video Surveillance and Monitoring“, Carnegie Mellon University, CMU-RI-TR-00-12, 2000.

[7] F. Matossi, „Bergmann Schaefer: Lehrbuch der Experimentalphysik“, Bd. III, Optik, 4th Edition, Walter de Gruyter & Co., Berlin 1966, p. 75.

[8] M. Sonka, V. Hlavac, R. Boyle, „Image Processing, Analysis and Machine Vision“, 1st Edition, Chapman & Hall Computing, 1993.

[9] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D.Tolliver, N. Enomoto, O. Hasegawa, P. Burt, L. Wixson, „A System for Video Surveillance and Monitoring“, Carnegie Mellon University, CMU-RI-TR-00-12, 2000.