Counting Moving People in Videos by Salient Points Detection

D. Conte, P. Foggia, G. Percannella, F. Tufano and M. Vento
Dipartimento di Ingegneria dell'Informazione ed Ingegneria Elettrica
Università di Salerno
Via Ponte don Melillo, I-84084 Fisciano (SA), Italy
dconte@unisa.it, pfoggia@unisa.it, pergen@unisa.it, ftufano@unisa.it, mvento@unisa.it

Abstract

This paper presents a novel method to count people for video surveillance applications. The problem is faced by establishing a mapping between some scene features and the number of people. Moreover, the proposed technique takes specifically into account problems due to perspective.

In the experimental evaluation, the method has been compared with respect to the algorithm by Albiol et al., which provided the highest performance at the PETS 2009 contest on people counting, using the same datasets. The results confirm that the proposed method improves the accuracy, while retaining the robustness of Albiol's algorithm.

1 Introduction

The estimation of the number of people present in an area can be an extremely useful information both for security/safety reasons (for instance, an anomalous change in number of persons could be the cause or the effect of a dangerous event) and for economic purposes (for instance, optimizing the schedule of a public transportation system on the basis of the number of passengers). Hence, several works in the fields of video analysis and intelligent video surveillance have addressed this task. The problem of people counting has been faced using two different approaches. In the *direct approach* (also called *detection-based*), people in the scene are first individually detected, using some form of segmentation and object detection, and then counted. In the *indirect approach* (also called *map-based* or *measurement-based*), instead, counting is performed using the measurement of some feature that does not require the separate detection of each person in the scene. The indirect approach is considered to be more robust, since the correct segmentation of people in the scene is by itself a complex problem that cannot be solved reliably, especially in crowded conditions.

Recent examples of the direct approach are [13], [4], [14], [15] and [17]. For the indirect approach, recent methods have proposed, among the others, the use of measurements such as the amount of moving pixels [5], blob size [8], fractal dimension [10] or other features [12], [11]. A recent method following the indirect approach has been proposed by Albiol et al. in [2]. This method has been submitted to the PETS 2009 contest on people counting, and has obtained the best performance among the contest participants. In Albiol's paper, the authors propose the use of corner points as features. Namely, corner points are found using a variant of the popular Harris corner detector [7]. The number of people is estimated from the number of moving corner points assuming a direct proportionality relation, with a constant factor determined using a frame of the video sequence. Although the assumptions underlying Albiol's paper may appear rather simplistic, the method has proven to be quite more robust than more sophisticated competitors. However, the accuracy it can attain is limited by the fact that it does not take into account problems like the instability of the Harris corner detector or the need of a perspective correction.

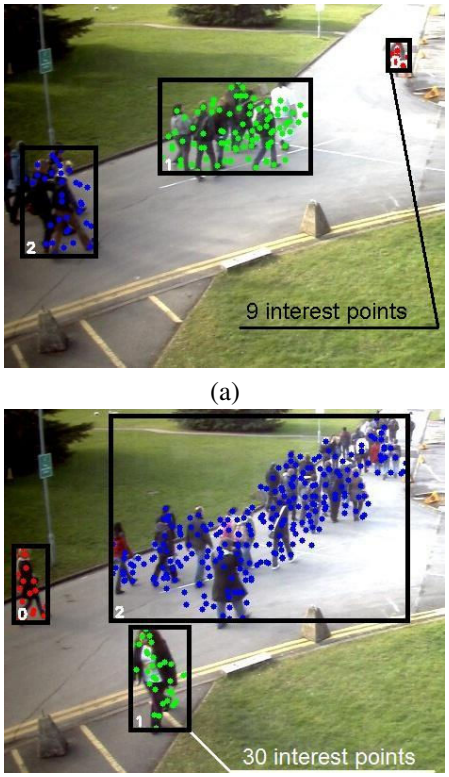
In this paper we propose a method that, while retaining the overall simplicity and the robustness of Albiol's approach, tries to provide a more accurate estimation of the count by considering also these factors.

2 The proposed method

The approach we propose in this paper is conceptually similar to the one by Albiol et al. [2], but introduces several changes to overcome some limitations of that method. The first problem that is addressed is the stability of the detected corner points. The points found by the Harris corner detector are strongly dependent on the perceived scale of the considered object: the same ob-

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.431

IEEE computer society



(a)

(b)

Figure 1. The effect of perspective on the number of detected interest points.

ject, even in the same pose, will have different detected corners if its image is acquired from different distances. This can cause problems in at least two different conditions. First, the observed scene contains groups of people whose distance from the camera is very different: in this case it is not possible to use a simple proportionality to estimate the number of people, since the average number of corner points per person is different between close people and far people. Second, the observed scene contains people walking on a direction that has a significant component orthogonal to the image plane, i.e. they are coming closer to the camera or getting farther from it: in this case the number of corner points for these people is changing even if the number of people remains constant.

To mitigate this problem we have chosen to adopt the SURF algorithm proposed by H. Bay et al. in 2006 [3]. SURF is inspired by the SIFT scale-invariant descriptor [9], but replaces the Gaussian-based filters of SIFT with filters that use the Haar wavelets, which are significantly faster to compute. The interest points found by SURF are much more independent of scale (and hence of distance from camera) than the ones provided by Harris detector. They are also independent of rotation, which is important for the stability of the points located on the arms and on the legs of the people in the scene.

As proposed in [8], the interest points associated to people are extracted in two steps. First, we determine all the SURF points within the frame under analysis. Then, we prune the points not associated to persons by taking into account their motion information. In particular, for each detected point we estimate the motion vector with respect to the previous frame by using a block-matching technique and prune those points with a null motion vector.

In addition to the use of SURF features, we also explicitly estimate the distance of people from the camera in order to account for the effects of perspective, which causes that the farther the person is from the camera, the fewer are the detected interest points. An example of the occurrence of this effect is shown in Figure 1 where a very different number of points is associated to the same person (9 versus 30 points in (a) and (b)) depending on the distance from the camera.

To perform this task, we first partition the detected points into groups corresponding to different groups of people. This can be treated as a clustering problem. However, the faced clustering problem is characterized by the fact that we do not have any a priori knowledge about the number and the shape of the clusters to be found. This depends on the fact that people can appear in different positions in the scene and can be aggregated in many different ways. In this situation more commonly used clustering methods (such as k-means) could not have been applied because they require the user to provide either the number of desired clusters or a threshold on cluster diameter or on inter-cluster distance. As observed in [16], the clustering algorithms based on graph theory are well suited to face clustering problems where no assumptions can be made about the clusters. In particular, we adopted the technique presented in [6], since (differently from other algorithms in the graph-based clustering family) it requires no parameters to be tuned or adapted to the particular application.

Once the detected points are divided into clusters, the distance of each cluster from the camera is derived from the position of the bottom points of the cluster applying an Inverse Perspective Mapping (IPM). The IPM is based on the assumption that the bottom points of the cluster lie on the ground plane. It has to be noted that the latter is a correct assumption when the clustering algorithm provides groups constituted by single persons or by persons close to each other and at the same distance from the camera (as an example, see the box 1 in Figure 1.a): in these cases, the error in the estimation of

the distance of the people from the camera is negligible. Conversely, when several persons at different distances from the camera are aggregated in a single cluster, the distance estimation error can be significant (see the box 1 in Figure 1.b). Nevertheless, even if the estimation for clusters formed by people at different distances may be inaccurate, it is still an improvement over the use of a global estimate based on all the detected points in the scene, as in Albiol et al.'s method. The inverse perspective matrix can be derived by calibration, using the images of several persons located at different distances from the camera and assuming that they have an average height.

Another limitation of the Albiol's approach addressed by our method is that the relation between the number of detected points and the number of people can have a form that is more complex than a simple direct proportionality, especially if we take into account the distance from the camera. So we have chosen to learn this relation by using a trainable function estimator. More precisely, we have used a variation of the Support Vector Machine known as *ϵ-Support Vector Regressor* (ϵ-SVR for short) as our function estimator. The ϵ-SVR receives as its inputs the number of points of a cluster and the distance, and is trained (using a set of training frames) to output the estimated number of people in the cluster. The ϵ-SVR is able to learn a non linear relation and shows a good generalization ability.

As with the Albiol's method, the output count is passed through a low-pass filter to smooth out oscillations due to image noise.

3 Experimental Results

The performance of the proposed method have been assessed using the PETS2009 dataset [1]. The dataset is organized in four sections, but we focused our attention only on the section named S1 that was used to benchmark algorithms for the "Person Count and Density Estimation" PETS2009 contest. In our experiments we used the same set of sequences adopted in PET2009, namely View 1 of the S1.L1.13-57, S1.L1.13-59, S1.L2.14-06 and S1.L3.14-17 sequences (hereinafter, the above sequences will be named V1, V2, V3 and V4, respectively). For all the sequences we calculated the number of people in each whole frame.

In order to use the proposed system for people counting we had to firstly train the ϵ-SVR regressor. The training set was built by collecting some samples of people groups from a subset (about 5%) of the test frames. For each selected box we calculated the feature vector and the associated ground truth, i.e. the true number of persons that are inside the box. Samples were carefully selected in order to guarantee that all the possible combinations in terms of number of persons in the group and distance from the camera were adequately represented in the training set. Testing has been carried out by comparing the actual number of people in the video sequences and the number of people calculated by the algorithm. The indices used to report the performance are the Mean Absolute Error (MAE) and the Mean Relative Error (MRE) defined as:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |G(i) - T(i)| \tag{1}$$

$$MRE = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{|G(i) - T(i)|}{T(i)} \tag{2}$$

where N is the number of frames of the test sequence and $G(i)$ and $T(i)$ are the guessed and the true number of persons in the i-th frame, respectively.

The MAE index is the same performance index used also to compare the performance of the algorithms that participated to the PETS2009 contest. This index is very useful to quantify the error in the estimation of the number of person which are in the focus of the camera, but it does not relate this error to the number of people; in fact, the same absolute error can be considered negligible if the number of persons in the scene is high while it becomes significant if the number of person is of the same order of magnitude. For this reason, we introduced also the MRE index which takes into account the estimation error related to the true people number.

In the Table 1 the performance of the proposed method on the adopted dataset is reported together with that of the Albiol's method. The motivation behind the choice of comparing our technique with respect to the Albiol's method is twofold: firstly, this method provided the best results on the PETS 2009 contest on people counting and, secondly, it constitutes the base from which we started for the definition of our method. It is worth noting that also the Albiol's method requires a training procedure for determining the optimal value of the interest points per person ratio. This value was determined by minimizing the MAE on the same set of frames already used for training our method.

From the results reported in Table 1 it is evident that the proposed method always outperforms Albiol's technique with respect to both MAE and MRE performance indices. However, in order to highlight the improvement introduced by the proposed method, we have also reported in Table 2 the relative improvements with respect to the performance of the algorithm by Albiol et al. The results reported in Table 2 highlight that the proposed method reduces in most cases the estimation

	Performance index MAE/MRE	
Sequence	Albiol	Our method
V1	2.80 / 12.6%	1.20 / 5.81%
V2	3.86 / 24.9%	1.39 / 11.0%
V3	5.14 / 26.1%	5.12 / 21.8%
V4	2.64 / 14.0%	1.92 / 9.6%

Table 1. Performance comparison.

Performance	Video sequence			
index	**V1**	**V2**	**V3**	**V4**
MAE	57.2%	64.0%	0.4%	27.2%
MRE	53.9%	55.9%	16.5%	31.4%

Table 2. Relative improvements.

error of more than 30%, both in terms of absolute and relative errors. The only exception is represented by the results obtained on the challenging S1.L2.14-06 sequence where the improvement obtained using the proposed method is lower than in the remaining cases.

The last remark regards the computational requirements. The Albiol's algorithm and the proposed one are both characterized by the fact that the computing time is mostly spent for detecting the interest points. This is due to the need of analyzing all the pixels of the image; the remaining steps employ only a small ratio of the total computing time, experimentally estimated below 1%. Repeated runs of both the algorithms confirmed comparable running times; 4CIF videos are processed at a rate of 25 frames per second on a Xeon 3GHz.

4 Conclusions

In this paper, we have proposed a novel method for counting moving people in a video surveillance scene. The method has been experimentally compared with the algorithm by Albiol et al. that was the winner of the PETS 2009 contest on people counting, highlighting the effectiveness of its enhancements. The experimentation on the PETS 2009 database has confirmed that the proposed method is in several cases more accurate than Albiol's but retains comparable robustness and computational requirements that are considered the greatest strengths of the latter. As a future work, a more extensive experimentation will be performed, adding other algorithms to the comparison and enlarging the video database to provide a better characterization of the advantages of the new algorithm.

References

[1] *http://www.cvg.rdg.ac.uk/PETS2009/.*

[2] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi. Video analysis using corner motion statistics. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 31–38, 2009.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[4] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 594–601, 2006.

[5] S.-Y. Cho, T. W. S. Chow, and C.-T. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 29(4):535–541, 1999.

[6] P. Foggia, G. Percannella, C. Sansone, and M. Vento. A graph-based algorithm for cluster detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(5):843–860, 2008.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[8] D. Kong, D. Gray, and H. Tao. A viewpoint invariant approach for crowd counting. In *International Conference on Pattern Recognition*, pages 1187–1190, 2006.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[10] A. N. Marana, L. da F. Costa, R. A. Lotufo, and S. A. Velastin. Estimating crowd density with mikowski fractal dimension. In *Int. Conf. on Acoustics, Speech and Signal Processing*, 1999.

[11] V. Rabaud and S. Belongie. Counting crowded moving objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 705–711, 2006.

[12] H. Rahmalan, M. S. Nixon, and J. N. Carter. On crowd density estimation for surveillance. In *The Institution of Engineering and Technology Conference on Crime and Security*, 2006.

[13] J. Rittscher, P. Tu, and N. Krahnstoever. Simultaneous estimation of segmentation and shape. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 486–493, 2005.

[14] M. D. Rodriguez and M. Shah. Detecting and segmenting humans in crowded scenes. In *ACM International Conference on Multimedia*, page 353356, 2007.

[15] O. Sidla, Y. Lypetskyy, N. Brandle, and S. Seer. Pedestrian detection and tracking for counting applications in crowded situations. In *AVSS '06: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, page 70, Washington, DC, USA, 2006. IEEE Computer Society.

[16] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Third Edition*. Academic Press, February 2006.

[17] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(7):1198–1211, 2008.