2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

A Method for Counting People in Crowded Scenes

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. Methods in the literature either follow a direct approach, by first detecting people and then counting them, or an indirect approach, by establishing a relation between some easily detectable scene features and the estimated number of people. The indirect approach is considerably more robust, but it is not easy to take into account such factors as perspective or people groups with different densities.

The proposed technique, while based on the indirect approach, specifically addresses these problems; furthermore it is based on a trainable estimator that does not require an explicit formulation of a priori knowledge about the perspective and density effects present in the scene at hand.

In the experimental evaluation, the method has been extensively compared with the algorithm by Albiol et al., which provided the highest performance at the PETS 2009 contest on people counting. The experimentation has used the public PETS 2009 datasets. The results confirm that the proposed method improves the accuracy, while retaining the robustness of the indirect approach.

1. Introduction

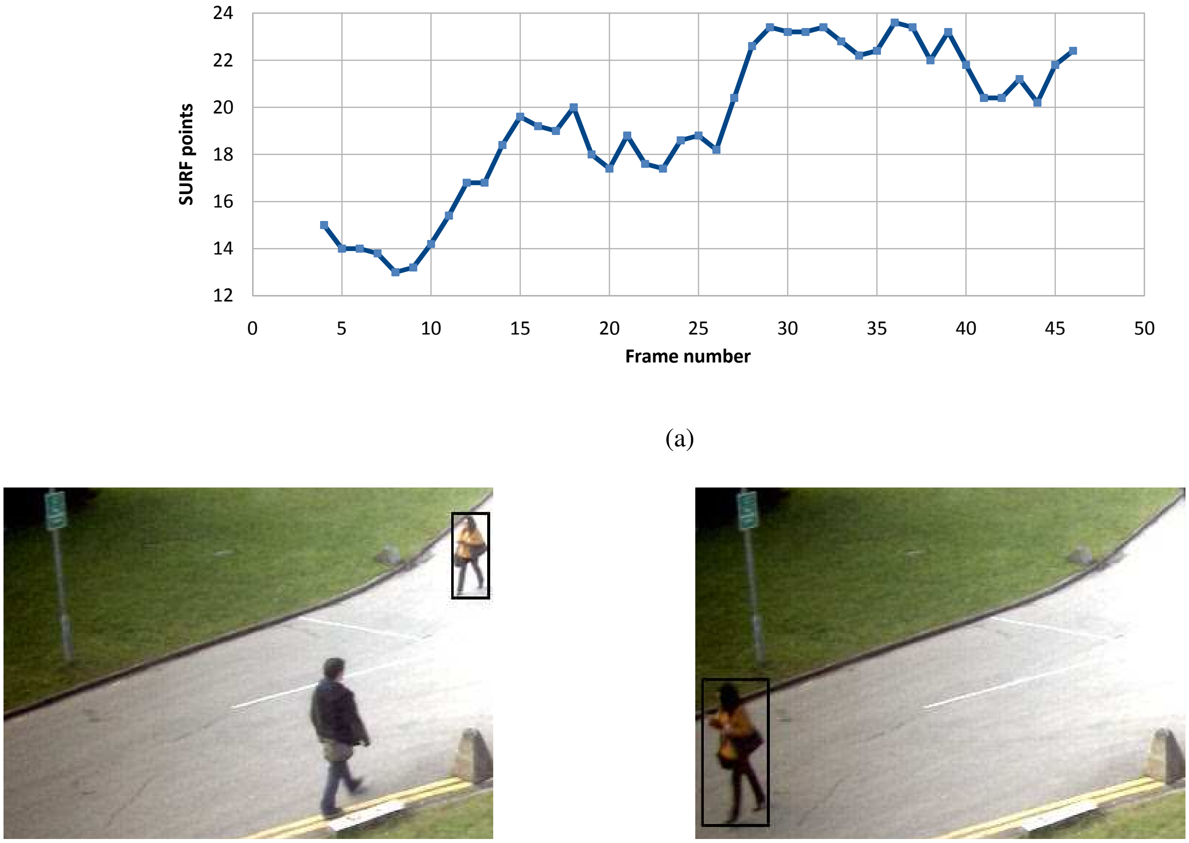
The estimation of the number of people present in an area can be an extremely useful information both for security/safety reasons (for instance, an anomalous change in number of persons could be the cause or the effect of a dangerous event) and for economic purposes (for instance, optimizing the schedule of 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 literature on people counting presents two conceptually different ways to face this task. In the *direct approach* (also called *detection-based*), each person in the scene is individually detected, using some form of segmentation and object detection; the number of people is then trivially obtainable. In the *indirect approach* (also called *map-based* or *measurement-based*), instead, counting is performed using the measurement of some features that do not require the separate detection of each person in the scene; these features then have to be put somehow in relation to the number of people.

The direct approach has the advantage that people detection is often already performed on a scene for other purposes (e.g. detecting events based on a person's position or trajectory), and as long as people are correctly segmented, the count is not affected by perspective, different people densities and, to some extent, partial occlusions. On the other hand, correct segmentation of people is a complex task by itself, and its output is often unreliable, especially in crowded conditions (which are of primary interest for people counting). The indirect approach instead is more robust, since it is based on features that are simpler to detect, but it is often not easy to find an accurate correspondance between these features and the number of people, especially if people may appear in the scene at different distances from the camera, and in groups with diverse densities.

Recent examples of the direct approach are [13], [4] and [15]. For the indirect approach, recent methods have proposed, among the others, the use of measurements such as the amount of moving pixels [6], blob size [9], fractal dimension [11] or other texture features [12]. Two recent methods following the indirect approach have been proposed by Albiol et al. in [2] and by Chan et al. in [5]. Both methods have been submitted to the PETS 2009 contest on people counting and have obtained very good performance among the contest participants. In Albiol's paper, the authors propose the use of corner points (detected using the Harris' algorithm [8]) as features. Static corner points (likely belonging to the background) are removed by computing motion vectors between adjacent frames. Finally, the number of people is estimated from the number of moving corner points assuming a direct proportionality relation.

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.78



(a)

(b)

(c)

Figure 1. The effect of perspective on the number of detected interest points. In a) it is reported the graph of the number of SURF points associated to the person denoted with the box in a video sequence whose first and last frames are shown in b) and c).

Altough Albiol's method has proved to be quite more robust than its competitors, the accuracy it can attain is limited by the fact that it does not take into account perspective effects, nor the influence of people density on the detection of corner points. Also, Harris' corner detector can be sometimes unstable for objects moving towards the camera or away from it.

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. Furthermore, the estimation is obtained through a trainable regressor (using the ϵ-SVR algorithm) that can be easily adapted to the characteristics of a new scene.

2. The proposed method

The approach we propose in this paper is based on the indirect approach. In particular, it uses as its features the moving interest points, where the interest points are first detected using a feature detector from the state of the art, and then the static ones are filtered out on the basis of a motion vector estimation. Under this respect the method is conceptually similar to the one by Albiol et al. [2], which has proved to be very successful at the PETS2009 people counting contest. However, while Albiol's algorithm assumes a very simple relation between the number of detected points and the number of persons (a direct proportionality), our method uses a more sophisticated estimation technique that takes into account several factors that could affect the relation between points and persons.

The first problem addressed is the the effect 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 problem is shown in Figure 1. Let us consider the woman denoted with the box that enters the scene (top right corner of the frame in Figure 1.b), goes progressively closer to the camera (up to the bottom left corner of the frame in Figure 1.c); from the graph in Figure 1.a it is evident how the closer is the box to the camera the higher is the moving SURF point associated to it.

In order to account for this effect, our algorithm computes the distance of each person or group of persons from the camera. To obtain this information, we first partition the detected points into groups corresponding to different groups of people. This can be treated as a clustering problem, but with the peculiarity that the shape of the clusters, their number and their densities are not known a priori. Because of this, commonly used clustering algorithms such as *k-means* and *DBSCAN* cannot be applied. So, to perform this task we have adopted the graph-based clustering algorithm presented in [7], which provides a good partitioning when the clusters are reasonably separated, without requiring any a priori information about the clusters.

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. 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 factor our algorithm takes into account is the effect of people density in a group. The more the persons in a group are close to each other, the more partial occlusions occur, reducing the visible part of the body, and thus the number of interest points per person. To consider this effect we compute a rough estimate of the people density by measuring how close are the interest points in the group. More precisely, we measure the ratio between the number of interest points in the group and the area covered by the group itself.

Given the need to consider not only the number of points, but also the distance from the camera and the density, the relation between these measurements and the number of people cannot be a simple direct proportionality as in Albiol's method. Actually, even if a single measurement were involved, the relation might have been non linear, at least in principle; with three measurements, there is the problem of understanding their relative weights and how they interact with each other to determine the count estimate.

Since this problem cannot be easily solved analytically, 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, the distance from the camera and the point density of the cluster, 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.

A further problem that is addressed in our method is the stability of the detected interest points. The points found by the Harris corner detector are somewhat dependent on the perceived scale and orientation of the considered object: the same object will have different detected corners if its image is acquired from a different distance or when it has a different pose.

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

Thus, Figure 2 shows the architecture of the proposed

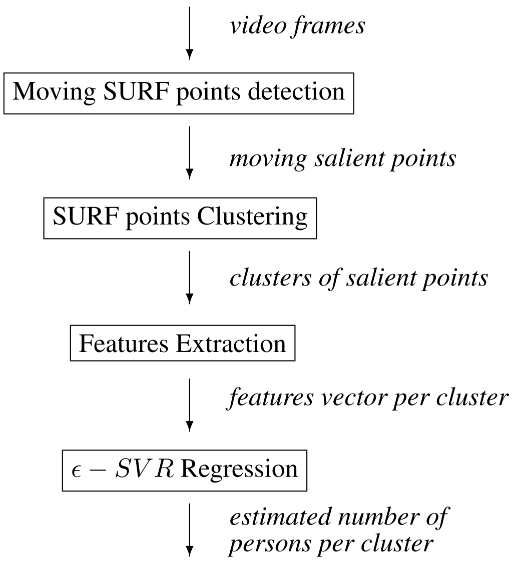

video frames
Moving SURF points detection
moving salient points
SURF points Clustering
clusters of salient points
Features Extraction
features vector per cluster
$\epsilon - SVR$ Regression
estimated number of persons per cluster

Figure 2. System architecture.

algorithm; an outline of the method is described by the following steps: Moving SURF points detection, SURF point Clustering and Feature extraction and regression.

2.1. Moving SURF points detection

In order to detect interest points associated to people we make two basic assumptions: persons within the scene are not static and there are no other moving elements in the scene. Thus, if we compute the interest points of the image and the associated motion information, the above assumptions guarantee that only the interest points with a non null motion vector must be associated to people.

As proposed in [2], the interest points associated to people are extracted in two steps. First, we determine all the interest points within the frame under analysis. Then, we prune the points not associated to persons by taking into account their motion information.

Interest points are determined by using the SURF algorithm [3] and not the Harris corner detector as in the paper by Albiol et al. [2].

In order to remove the static interest points (that are not associated to people), for each point detected by the SURF algorithm we estimate the motion vector with respect to the previous frame by using a block-matching technique. Then we distinguish between static and moving interest points on the basis of the following rule:

$$p(x,y) = \begin{cases} \text{moving point} & \text{if } |\vec{v}(x,y)| > \beta \\ \text{static point} & \text{if } |\vec{v}(x,y)| \leq \beta \end{cases} \quad (1)$$

where $p(x,y)$ is the interest point at the x, y coordinates, $|\vec{v}(x,y)|$ is the magnitude of the motion vector calculated in x, y with respect to the previous frame; β is a bias value (in our experiments we set $\beta = 0.0$).

2.2. SURF point Clustering

In order to compensate for changes in the number of points due to perspective and to partial occlusions, the algorithm needs to partition the detected points into clusters corresponding to separated groups of persons, so as to be able to compute for each group its distance from the camera and its density.

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. As observed in [14], 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 [7], since (differently from other algorithms in the graph-based clustering family) it requires no parameters to be tuned or adapted to the particular application.

This algorithm represents the set of points as a graph in which each point corresponds to a node and each edge is labeled with the distance between its endpoints. The minimum spanning tree (MST) of the graph is computed; this tree will contain some edges that are between nodes in the same cluster (intra-cluster edges) and other edges between nodes of different clusters (inter-cluster edges). Assuming that the clusters are well separated, it can be expected that the intra-cluster edges are shorter than the inter-cluster edges. So the algorithm uses a thresholding to divide the edges in two sets (the ones below the threshold, say it λ, and the ones above the threshold λ). The edges in the second set are deleted, and the remaining connected components are the clusters output by the algorithm.

The use of a fixed value for the threshold λ would be problematic, since the threshold would need to be adjusted depending on the resolution, the distance from the camera and so on. Instead, we have used a threshold proportional to the average edge length, computed as:

$$\lambda = \gamma \cdot \frac{1}{N} \sum_{i=1}^{N} x_i \quad (2)$$

where γ is the proportionality factor, N is the number of edges of the spanning tree, while x_i is the weight of the i-th edge of the tree. We have experimentally found that the choice $\gamma = 2.0$ works adequately for all the considered scenes.

2.3. Feature extraction and regression

In this stage of the algorithm, a feature vector is computed from each cluster detected in the previous step, and is fed into a regressor. The output of the regressor is the estimated number of persons in the group represented by the cluster.

The basic idea of the method in [2] is that the average number of interest points associated to each person is a global property of scene. Thus, once the scene has been defined, it is possible to assume a simple direct proportionality relation between the number of points and the number of persons.

We reasonably assume that when the density of people increase, the detected points get closer to each other. So we can consider the density of the points as related to people density, and we can indirectly take into account people density by establishing a relation between the average number of points per person and the point density.

the relation between the number of interest points and the number of people appears more complex than a direct proportionality, as we have to take into account also the distance of the people from the camera and the point density. We can formulate this relation as:

$$n_{people} = f(n_{points}, \rho, d) \quad (3)$$

where:

- n_{people} is the estimated number of people;
- n_{points} is the number of interest points within the cluster;
- ρ is the average density of the points in the cluster: the value is obtained as the ratio between the number of points into the cluster and the area of the bounding box. Note that the area of the bounding box is computed with respect to real world coordinates. This allows us to normalize the average density of the points to the value it would have if the cluster were moved to a predefined distance from the camera;
- d is the distance of the cluster from the camera: assuming that the bottom points of the bounding box lie on the ground plane, the calculation is done by applying an Inverse Perspective Mapping and is referred to the center of the bottom edge of the cluster's bounding box.

Since we do not know the analytical form of f, we have chosen to learn this function from a set of labeled examples by using an ϵ-SVR regressor. Once trained, the ϵ-SVR acts as a function estimator; for each detected cluster it receives as its input the above features and outputs the estimated number of people within the cluster. So the total number of persons in the frame (or in a predetermined region of interest) is obtained by summing the number of people calculated for each cluster.

Finally, in order to smooth out the oscillations in the number of the counted persons among consecutive frames, we employ a low-pass filter. Specifically, the final count of the persons within the scene is calculated as the average value of the people count on the last *k* frames of the video.

3. Experimental Results

The performance of the proposed method has been assessed using the PETS2009 dataset [1]. The dataset is organized in four sections, but we focused our attention primarily on the section named S1 that was used to benchmark algorithms for the "'Person Count and Density Estimation"' PETS2009 contest. The main characteristics of the subset of video sequences of the PETS 2009 dataset used for assessing the performance of the proposed method are summarized in the Table 1 in terms of their length, number of people in the scene (minimum, maximum and average number) and other elements as density of the crowd, illumination conditions, etc.

The videos reported in Table 1 refer to two different views obtained by using two cameras that contemporaneously framed the same scene from different points (see Figure 3 for an example frame of each view). For our experimentations, we used four videos of the view 1, which are also the same videos that were used in the people counting contest held in PETS2009. The videos in the second set refer to the view 2 which is characterized by a wide field depth that makes the counting problem more difficult to solve.

For all the sequences we calculated the number of people in the whole frame.

In order to use the proposed system for people counting, we had first to train the ϵ-SVR regressor. The minimum size of the training set needed to achieve an acceptable performance, as the statistical learning theory by Vapnik and Chervonenkis has demonstrated, depends on both the complexity of the problem and the complexity of the estimator to be trained. The method by Albiol et al. uses a very simple estimator, so that a single frame per sequence is sufficient for the training. Our estimator is more complex, so it needs more training frames. The training set was built by manually collecting some samples of people groups from a subset 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, points density and distance from the camera were adequately represented in the training set. It is worth pointing out that the required number of training frames has not to be very large to achieve a good performance level (in our tests we used about 30-40 training frames, which correspond to about 5% of the frames within the whole dataset), by taking into account also the fact that a single frame usually contains several people clusters at different distances, so it may cover several cases of the function to be learned.

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:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |G(i) - T(i)| \qquad (4)$$

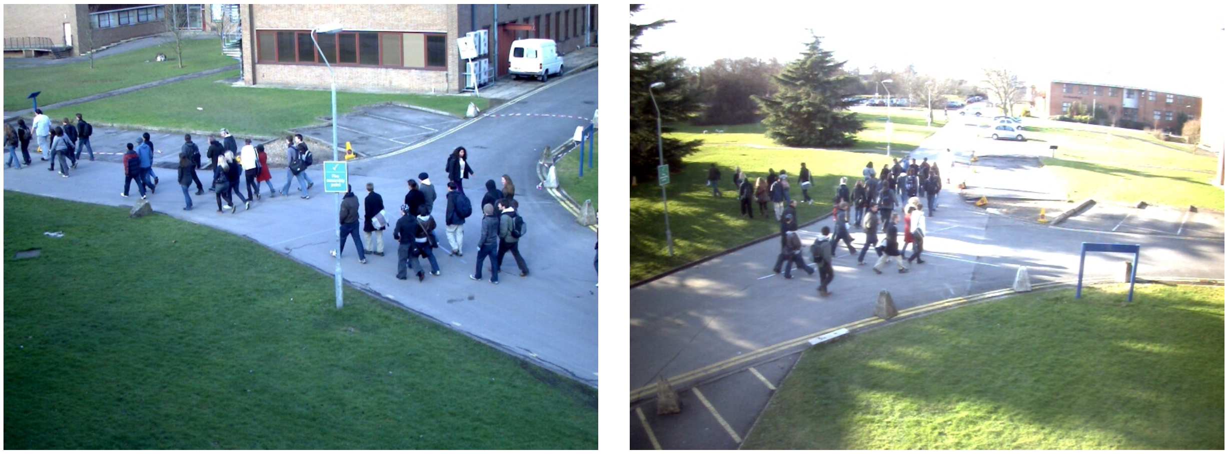
$$MRE = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{|G(i) - T(i)|}{T(i)} \qquad (5)$$

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 exactly quantify the error in the estimation of the number of person which are in the focus of the camera, but it does

				Number of people		
Video sequence	View	Length (frames)	Conditions	Min	AVG	Max
S1.L1.13-57	1	221	medium density crowd, overcast	5	22.61	34
S1.L1.13-59	1	241	medium density crowd, overcast	3	15.81	26
S1.L2.14-06	1	201	high density crowd, overcast	0	26.28	43
S1.L3.14-17	1	91	medium density crowd, bright sunshine and shadows	6	24.34	41
S1.L1.13-57	2	221	medium density crowd, overcast	8	34.19	46
S1.L2.14-06	2	201	high density crowd, overcast	3	37.10	46
S1.L2.14-31	2	131	high density crowd, overcast	10	35.19	43
S3.MF.12-43	2	108	very low density crowd, overcast	1	4.99	7

Table 1. Relevant characteristics of the four sequences of the PETS 2009 datasets used for assessing the performance of the proposed method.



(a) (b)

Figure 3. Examples of the frames of the video sequences used for the test: a) S1.L1.13-57 (view 1), b) S1.L2.14-31 (view 2)

not relate this error to 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.

The performance of the proposed method on the adopted dataset is reported together with that of the Albiol's method, for which we have provided our own implementation. The motivation behind the choice of comparing our technique with respect to the Albiol's method is twofold. First, it constitutes the base from which we started for the definition of our method; thus, the comparison allows us to quantify the improvement provided by the proposed modifications. Secondly, Albiol's method has already been compared to other algorithms based either on the direct or the indirect approach, in the PETS 2009 contest on people counting, and has consistently outperformed them. Since our test dataset contains also the video sequences used for the PETS 2009 contest on people counting, we can reasonably expect that, at least on that kind of scene, also our method should show an improvement over those other algorithms.

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.

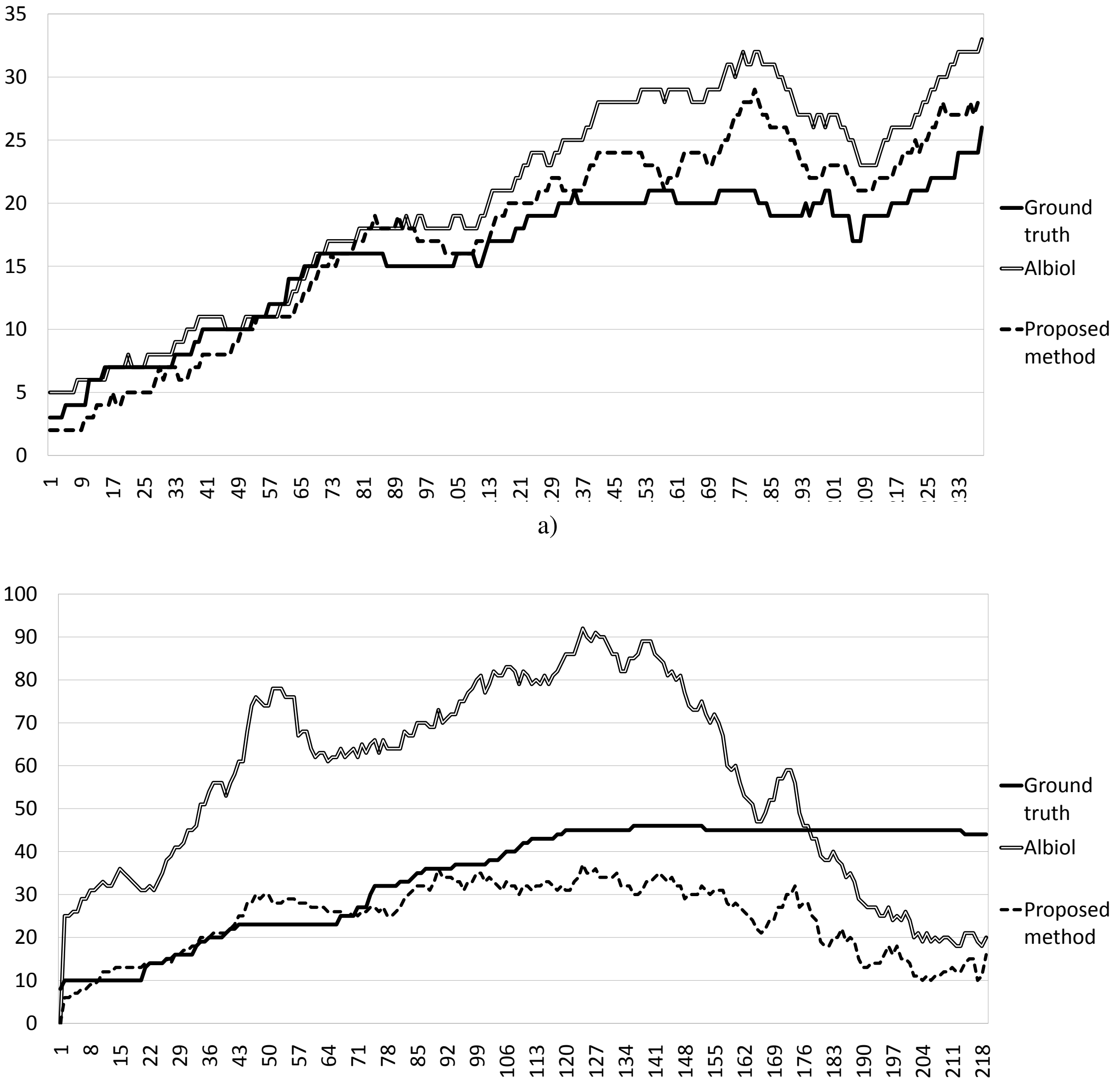
From the results reported in Table 2 it is evident that the proposed method always outperforms Albiol's technique with respect to both MAE and MRE performance indices.

In order to have a deeper insight into the behavior of the considered algorithms, Figure 4 shows the estimated number of people with respect to time for both our algorithm and Albiol's over two video sequences.

The different behavior of the considered algorithms can be explained by considering that the Albiol's method hypothesizes a linear relation between the number of detected interest points and the number of persons without taking into account the perspective effects and the people density. As a result this method provides better results when it is tested on videos characterized by conditions that are simi-

Video (view)	Albiol	Our	Rel. impr. %
S1.L1.13-57 (1)	2.80 (12.6%)	1.92 (8.7%)	31.4% (31.0%)
S1.L1.13-59 (1)	3.86 (24.9%)	2.24 (17.3%)	42.0% (30.6%)
S1.L2.14-06 (1)	5.14 (26.1%)	4.66 (20.5%)	9.3% (21.4%)
S1.L3.14-17 (1)	2.64 (14.0%)	1.75 (9.2%)	33.6% (34.3%)
S1.L1.13-57 (2)	29.45 (106.0%)	11.76 (30.0%)	60.1% (70.7%)
S1.L2.14-06 (2)	32.24 (122.5%)	18.03 (43.0%)	44.1% (64.9%)
S1.L2.14-31 (2)	34.09 (99.7%)	5.64 (18.8%)	83.4% (81.1%)
S3.MF.12-43 (2)	12.34 (311.9%)	0.63 (18.8%)	94.9% (94.0%)

Table 2. Performance of the Albiol's algorithm and of the proposed one. In each cell there are reported the values of the MAE and of the MRE (in parenthesis) performance indices for both Albiol's and our people counting method, while in the last column there are reported the relative improvements.



a)

b)

Figure 4. Curves of the number of people estimated by the Albiol's and our algorithms in each frame together with the ground truth on the video sequence S1.L1.13-59 view 1 (a) and S1.L1.13-57 view 2 (b). On the x-axis it is reported the frame number.

lar to those present in the training videos. Conversely, the proposed method is more robust with respect to the above problems.

In particular, the Figure 4.a refers to the view 1 of the video sequence S1.L1.13-59: this video is characterized by a group of persons that gradually enters and crosses the scene. In this view all the persons move in a direction that is ortogonal to the optical axis of the camera, so that their distance from the camera do not change significantly during their permanence in the scene. Consequently the main contribution to the performance improvement provided by our method can be abscribed to the fact that it takes into account the problem of the occlusions of the persons by means of points density. In fact, from the figure it is possible to note that the higher is the number of people, the higher is the estimation error of the method of Albiol.

In Figure 4.b, that refers to the view 2 of the sequence S1.L1.13-57, the persons move in a direction that is almost parallel to the optical axis of the camera; thus in this case the correction of the perspective effects plays a fundamental role in the performance improvements obtained by the proposed method. In fact, in this case the method of Albiol et al. tends to overestimate or underestimate the number of persons when they are close to or far from the camera while it provides a good estimate only when the persons are at an average distance from the camera (this is evident by considering the Albiol and the ground truth curves in the figure). On the contrary the proposed method is able to keep the estimation error low along almost all the sequence. The exception is represented by the last part of the sequence where the method tends to underestimate the number of the person: however, this can be explained by considering that in this part of the video the persons are very far from the camera and most of their interest points are considered static.

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] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2008.

[6] 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.

[7] 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.

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

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

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

[11] 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.

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

[15] 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.