AN EFFECTIVE METHOD FOR COUNTING PEOPLE IN VIDEO-SURVEILLANCE APPLICATIONS

D. Conte, P. Foggia, G. Percannella, F. Tufano and M. Vento
Dipartimento di Ingegneria Elettronica e Ingegneria Informatica, University of Salerno
Via Ponte don Melillo, 1, 84084, Fisciano (SA), Italy

Keywords: People counting, Crowd density estimation, Video surveillance.

Abstract: This paper presents a method to count people for video surveillance applications. The proposed method adopts the indirect approach, according to which the number of persons in the scene is inferred from the value of some easily detectable scene features.

In particular, the proposed method first detects the SURF interest points associated to moving people, then determines the number of persons in the scene by a weigthed sum of the SURF points. In order to take into account the fact that, due to the perspective, the number of points per person tends to decrease the farther the person is from the camera, the weight attributed to each point depends on its coordinates in the image plane. In the design of the method, particular attention has been paid in order to obtain a system that can be easily deployed and configured.

In the experimental evaluation, the method has been extensively compared with the algorithms by Albiol et al. and by Conte et al., which both adopt a similar approach. The experimentations have been carried out on the PETS 2009 dataset and the results show that the proposed method obtains a high value of the accuracy.

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, the counting is performed using the measurement of suitable 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 (Rittscher et al., 2005), (Brostow and Cipolla, 2006) and (Zhao et al., 2008). For the indirect approach, recent methods have proposed, among the others, the use of measurements such as the amount of moving pixels (Cho et al., 1999), blob size (Kong et al., 2006), fractal dimension (Marana et al., 1999) or other texture features (Rahmalan et al., 2006). Some recent methods following the indirect approach have been

Conte D., Foggia P., Percannella G., Tufano F. and Vento M..
AN EFFECTIVE METHOD FOR COUNTING PEOPLE IN VIDEO-SURVEILLANCE APPLICATIONS.
DOI: 10.5220/0003370400670074
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2011), pages 67-74
ISBN: 978-989-8425-47-8

proposed in (Albiol et al., 2009), in (Chan et al., 2008) and in (Conte et al., 2010). All these methods have been submitted to the PETS 2009 and 2010 contests on people counting and have obtained very good performance among the contests participants. In particular, in Albiol's paper, the authors propose the use of corner points detected using the Harris' algorithm (Harris and Stephens, 1988). 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.

Although Albiol's method has proved to be quite more robust than its competitors, its accuracy 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. Moreover, the Harris' corner detector is sometimes unstable for objects moving towards the camera or away from it.

In the paper (Conte et al., 2010), the authors propose a method that provides a more accurate estimation of the people number by considering also the issues related to perspective effects and occlusions. In particular, the authors propose to carry out the estimation of the count through a trainable regressor (using the ε-SVR algorithm) suitably trained on the used scene. Tests performed on very crowded scenes characterized by a large field depth demonstrated high performance improvements with respect to the method by Albiol et al. However, this is obtained at the cost of complex set up procedures for training the ε-SVR regressor.

In this paper we describe a method that is able to obtain performance comparable to those obtained by the method of Conte et al., but at the same time retains the overall simplicity of Albiol's approach.

2 SYSTEM ARCHITECTURE

The approach we propose in this paper is conceptually similar to the one in (Albiol et al., 2009), but introduces several changes to overcome some limitations of that method and draws some ideas from the approach in (Conte et al., 2010).

The first problem addressed is the stability of the detected corner points. The latter are strongly dependent on the perceived scale of the considered object: the same object, 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. Firstly, the observed scene contains groups of people whose distance from the camera is very different: in this case it is not effective to use a simple proportionality law to estimate the number of people, since the average number of corner points per person is different passing from close people to far ones. Secondarily, 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, as in (Conte et al., 2010) we adopt the SURF algorithm proposed in (Bay et al., 2008). SURF is inspired by the SIFT scale-invariant descriptor (Lowe, 2004), 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 rotation invariant, which is an important issue for the stability of the points located on the arms and on the legs of the people in the scene. The interest points associated to people are obtained 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 pruning those one with a null motion vector.

The second issue we address in this paper is the perspective effect, which causes that the farther the person is from the camera, the fewer are the detected interest points. As a consequence, a simple proportionality relation between the number of detected interest points and the number of persons in the scene provides acceptable results only when the average distance of the persons is close to a reference distance used to determine the proportionality factor, otherwise this approach tends to overestimate the number of people that are close to the camera and to underestimate it when people are far from the camera.

The authors in (Conte et al., 2010) propose to segment each single person or small group of persons at similar distances from the camera by clustering the detected interest points. 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), assuming that the bottom points of the cluster lie on the ground plane. Then, the number of persons in each cluster is determined using an ε-*Support Vector Regressor* that receives the number of points of a cluster, the distance and the point density. However, the main limitation of this approach is the costly and annoying procedure for training the regressor, which requires that the training samples are manually and carefully selected, so as to guarantee an adequate coverage of the possible situations in terms of the number and density of persons in the group and distance from the camera. Furthermore, this procedure has to be repeated for each camera and requires also that the calibration parameters of the camera are available.

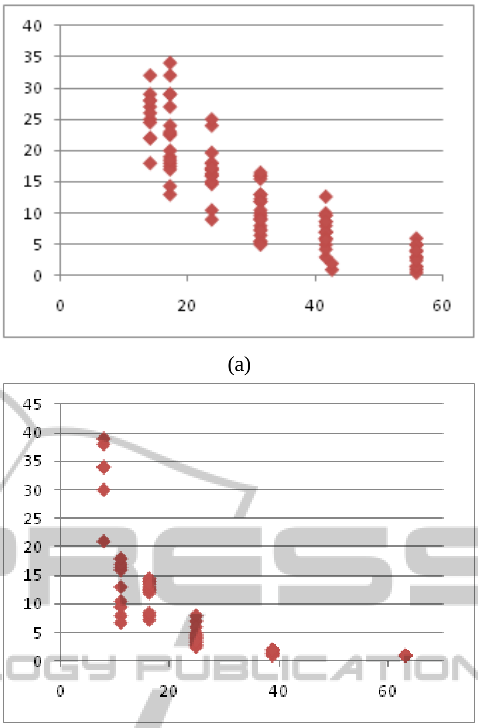
Our proposal stems from the idea that the number of interest points per person only depends on the distance of the person from the camera. This assumption is supported by experimental observations as it can be verified by considering the plots in Figure 1, which report the number of points per person versus the distance of the person from the camera. The plots were obtained using some sample frames from the video sequences of Table 1. The points are calculated by considering several persons at fixed distances from the camera and by counting the number of moving SURF points associated to each of them. From the figures, it is evident that the number of points per person strongly depends on the distance from the camera. This is more evident from the plot of Figure 1.b, obtained using a camera with a wide field depth that magnifies this dependence.

The total number of persons P into the scene can be estimated as:

$$P = \sum_{i=1}^{N} \omega\left(d\left(p_i\right)\right) \qquad (1)$$

where N is the number of detected moving SURF points and $\omega\left(d\left(p_i\right)\right)$ is the weigth attributed to the i-th point p_i. The weight depends on the distance $d\left(p_i\right)$ from the camera. The above formula requires that the distance $d\left(p_i\right)$ is calculated in the 3D-scene, but we do not have information about the 3D coordinates of the points. The authors in (Conte et al., 2010) implicitily solve this problem by attributing each detected point to a person, but, as observed by themselves, this operation can be easily performed only when persons are well isolated in the scene; so in cases of crowded scenes with persons at different distances from the camera, this procedure is tipically complex and the results unreliable.

In order to solve the problem of perspective normalization, we partition the generic frame in several longitudinal non overlapping bands, as proposed in (Chan et al., 2009). The height of each band is fixed equal to the height in pixels of an average person whose feet are on the base of the band. So, the bands represent classes of equivalence with respect to the value of the weight attributed to a detected point, thus attributing the same weight to all the points falling in the same band. According to this assumption, Equation 1 can be modified as:



(a)

(b)

Figure 1: Plots of the number of points per person (y-axis) versus the distance of the person from the camera in meters (x-axis). The points in the plots were obtained using some sample frames from view 1 (a) and view 2 (b) of the PETS2009 dataset.

$$P = \sum_{i=1}^{N} \omega\left(B_{p_i}\right) \qquad (2)$$

where B_{p_i} is the band the point p_i belongs to.

The partition of the scene in bands reconduces the counting problem in presence of perspective effects to n simpler counting problems, each for any band; in a band the perspective is not considered. Consequently, we have to set only the values of the weight for each band. For the generic i-th band, this is done by selecting some sample frames with persons that are all perfectly contained. Then, the weight for the band is obtained as the ratio between the total number of points in the band over the selected frames and the number of persons in the band. Once the set of the weigths ($\Omega = \{\omega\left(B_k\right)\}$), for all the bands, has been determined, it is possible to calculate the total number of persons in the scene by adopting Equation 2.

Finally, the output count is passed through a low-pass filter to smooth out oscillations due to image noise.

The set up procedure of the method primarily requires the determination of the height of the bands; these are depending on the geometrical parameters of the systems, as the focal lenght and the relative position of the camera in the environment. Once these ones have been properly evaluated, it is necessary to complete the procedure by estimating, for each band, the corresponding counting coefficient $\omega(B_k)$. It is worth noting that the knowledge of the perspective function f, giving the height in pixel of a person (having an average height) as a function of its position in the image, is sufficient to obtain the bands as a result of an iterative process. The perspective function f is linear and can be approximated by an automatic procedure, applied to a video of a few seconds: a person is required to cross the scene, moving in different directions, so as to obtain a good coverage of the visual area. In each frame, we can automatically determine its position p_i and the corresponding height h_i; once a sufficient number of these couples (p_i, h_i) have been extracted, it is possible to obtain, by an approximation method, the analytical expression of f. An example of the obtained results on PETS database is shown in the Figure 2.

3 EXPERIMENTAL RESULTS

The performance of the proposed method has been assessed using the PETS2009 dataset (PETS, 2009). The latter is organized in four sections, but we focused our attention primarily on the section S1 used to benchmark algorithms for the ”Person Count and Density Estimation” PETS2009 and 2010 contests. The main characteristics of the used video sequences are summarized in Table 1.

The videos refer to two different views obtained by using two cameras that contemporaneously acquired the same scene from different points (see Figure 3 for an example frame of each view). For our experimentations, we used four videos of 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 view 2 which is characterized by a wide field depth that makes the counting problem more difficult to solve. An example frame for each view of the test videos is shown in Figure 3.

Tests of the proposed method have been carried out by partitioning the frames of the video sequences in bands as shown in the Figure 2. The training of the system, aimed at determining the set Ω, was performed by using about 10 sample frames for each band. The frames were selected from other sequences available in the PETS2009 dataset that where not used for the tests. 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:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |G(i) - T(i)| \qquad (3)$$

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

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 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 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.

The performance of the proposed method on the adopted dataset is reported together with that of Albiol’s and Conte’s methods, using the results reported in (Conte et al., 2010). The motivation behind the choice of comparing our technique with respect to these two methods is twofold. First, both methods belong to the category of the indirect approaches. Secondly, both methods have already been compared to other algorithms based either on the direct or the indirect approach, in the PETS 2009 and 2010 contests on people counting, consistently outperforming them (Ellis and Ferryman, 2010). 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 performs well with respect to those other algorithms. From the results reported in Table 2, it is evident that the proposed method in almost all cases outperforms Albiol’s technique with respect to both MAE and MRE performance indices, while its performance is always very close to that obtained by Conte’s method. This aspect is more evident if we refer to the results obtained on view 2.

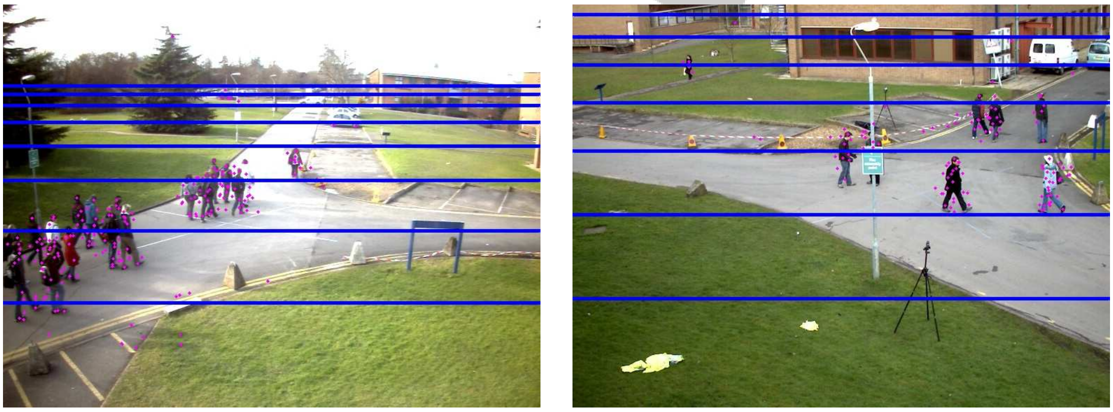
In order to have a deeper insight into the behavior of the considered algorithms, Figure 4 shows the es-

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

				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, 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 2: Performance of the Albiol's algorithm, of the Conte's and of the proposed ones. In each cell there are reported the values of the MAE and of the MRE (in parenthesis) performance indices.

Video (view)	**Albiol**	**Conte**	**Our**
S1.L1.13-57 (1)	2.80 (12.6%)	1.92 (8.7%)	1.37 (6.9%)
S1.L1.13-59 (1)	3.86 (24.9%)	2.24 (17.3%)	2.58 (15.6%)
S1.L2.14-06 (1)	5.14 (26.1%)	4.66 (20.5%)	5.44 (20.7%)
S1.L3.14-17 (1)	2.64 (14.0%)	1.75 (9.2%)	2.74 (15.1%)
S1.L1.13-57 (2)	29.45 (106.0%)	11.76 (30.0%)	9.13 (23.9%)
S1.L2.14-06 (2)	32.24 (122.5%)	18.03 (43.0%)	17.74 (43.6%)
S1.L2.14-31 (2)	34.09 (99.7%)	5.64 (18.8%)	6.61 (21.7%)
S3.MF.12-43 (2)	12.34 (311.9%)	0.63 (18.8%)	1.60 (34.6%)



(a)

(b)

Figure 2: Subdivision of the frames of the video sequences for the test: a) S1.L1.13-57 (view 1), b) S1.L2.14-31 (view 2). The height of each band approximatively corresponds to the height of a person in real world coordinates.

timated number of people as a function of time. The behavior of the considered algorithms with respect to the video sequences of Figure 4 can be explained by recalling the main hypothesis at the basis of each of them. 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 tested in working conditions that are similar to those present in the training videos. Conversely, the method by Conte et al. takes specifically into account both the perspective and the density issues, thus globally it provides better results. The proposed method uses the same hypothesis of Albiol, using a linear relation between points and persons, but the adopted proportionality

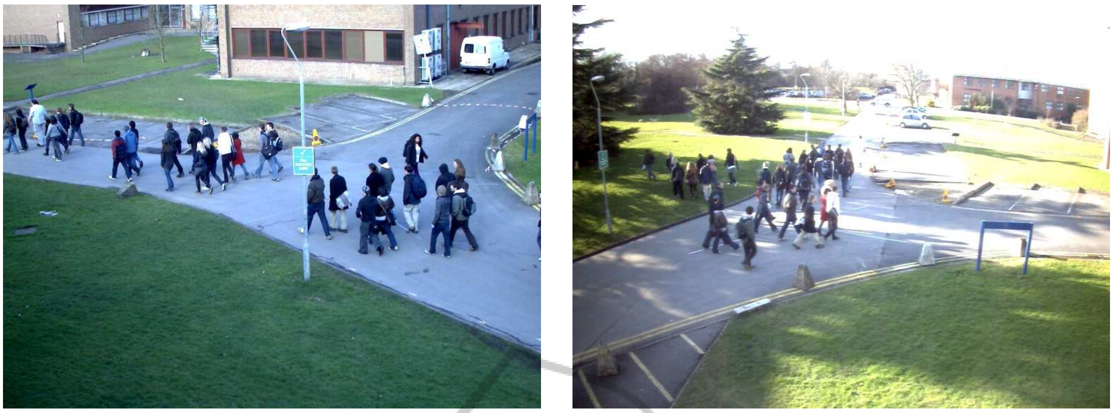
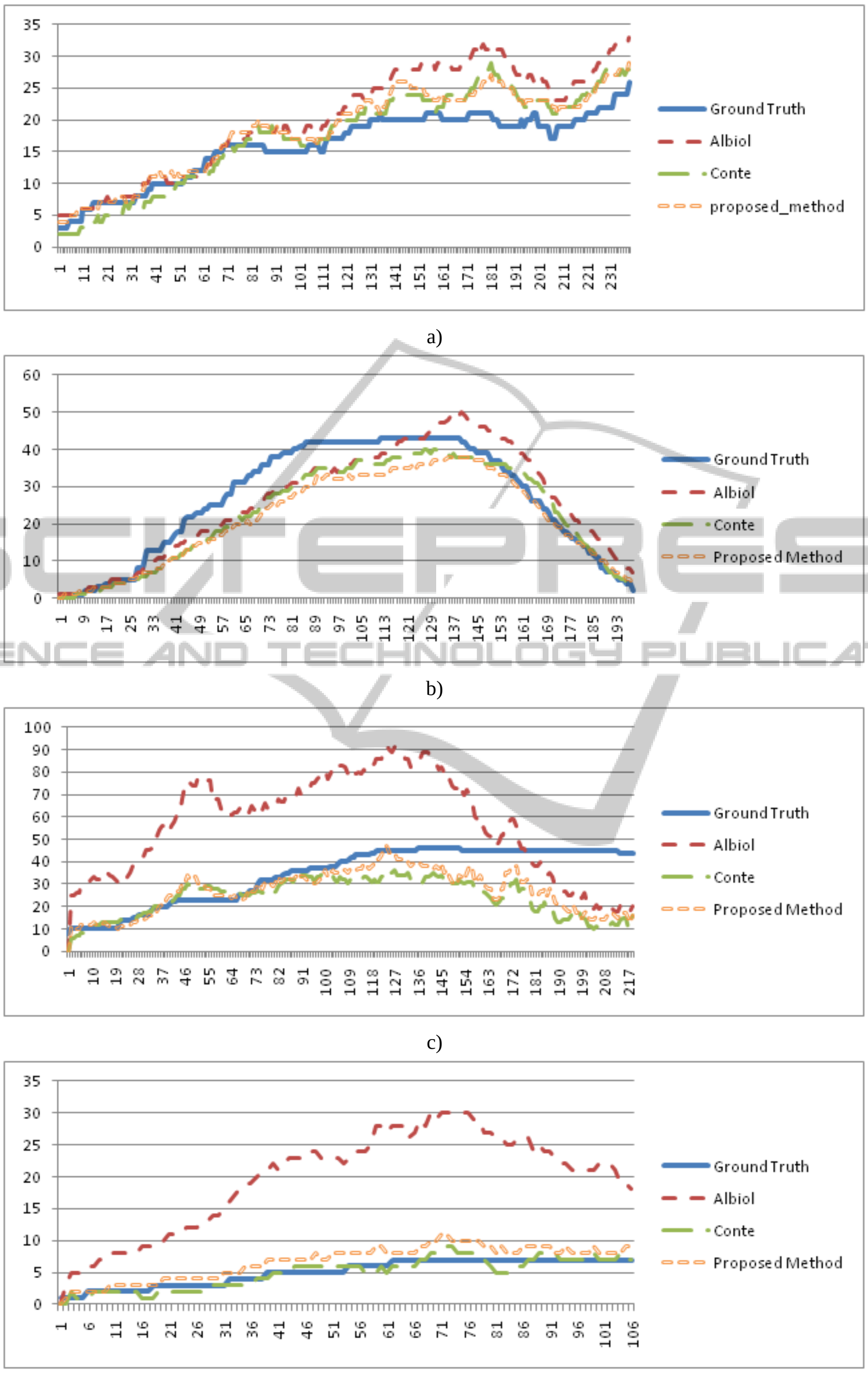

(a) (b)

Figure 3: Examples of the frames of the video sequences used for the test: a) view 1, b) view 2.

factor depends also on the distance from the camera in order to cope with perspective effects. Thus, good performance have to be expected also in cases where perspective is more evident, as in view 2. The Figure 4.a refers to the view 1 of the video sequence S1.L1.13-59. This video is characterized by isolated persons or very small groups of persons that gradually enters and crosses the scene with no or very small occlusions. The Figure 4.b refers to the same camera view sequence S1.L2.14-06, but in this case the persons cross the scene in a large and compact group, resulting in a high degree of occlusions among them. In both sequences all the persons move in a direction that is orthogonal to the optical axis of the camera, so that their distance from the camera does not change significantly during their permanence in the scene. In this regard, the perspective effect is not the main issue. If we consider these sequences, it is possible to observe that the proposed algorithm shows different behaviors if compared to the remaining two techniques: in fact, in one case it provides the lowest value of the absolute estimation error, while in the other one performs the worst. The presence of occlusions affect the performance of the proposed method; the higher is the degree of occlusion the higher is the estimation error. This can be simply explained by taking into account the fact that the proposed method has been trained by considering more samples of isolated persons than samples of groups of persons. However, it should also be noted that if we consider the relative estimation error the above described behavior changes quite significantly as the performance of the proposed method are much better. This fact is very interesting: this means that even when the absolute estimation error is higher in the average, this error is better distributed with respect to Albiol's approach and comparably with respect to Conte's one.

Figures 4.c and 4.d are related to view 2. 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 and Conte's one are able to keep the estimation error low along almost all the sequence. The exception is represented by the last part of the sequence S1.L1.13-57 where all approaches tend to underestimate the number of the persons: 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. The sequence S1.L1.13-57 is characterized by a quite large and dense crowd that crosses the scene in a direction that is almost parallel to the optical axis of the camera. Interestingly, in spite of the high degree of occlusions that characterizes the sequence, the proposed method performs better than Conte's method (Figure 4.c). This can be explained by considering the fact that the latter method infers the number of persons for each group obtained after the clustering procedure assuming that the bottom points of the cluster lie on the ground plane. This is a valid 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: in these cases, the error in the estimation of the distance of the people from the camera is negligible. As highlighted by the same authors, when several persons at different distances from the camera are aggregated in a single cluster, the dis-



a)

b)

c)

d)

Figure 4: Curves of the number of people in each frame estimated by the Albiol's algorithm, Conte's and the proposed ones together with the ground truth on the video sequence S1.L1.13-59 view 1 (a), S1.L2.14-06 view 1 (b), S1.L1.13-57 view 2 (c) and S3.MF.12-43 view 2 (d). On the x-axis it is reported the frame number.

tance estimation error can be significant. On the contrary, the proposed method is able to better cope with this situation due to the fact that the contribution of each interest point to the final estimation of the people number depends on the band which it belongs to. The curve reported in Figure 4.d, related to view 2 of the sequence S3.MF.12-43, shows that when there are few isolated persons in the scene Conte's method can provide more accurated results.

4 CONCLUSIONS

In this paper we have presented a method for counting people in video surveillance applications. The method has been experimentally compared with the algorithm by Albiol et al. and by Conte et al. that were among the best approaches of the PETS 2009 and 2010 contests on people counting. These methods are also the base from which we started to define our proposal. The experimentation on the PETS 2009 database has confirmed that the proposed method is in several cases more accurate than Albiol's one while retaining robustness and computational requirements that are considered the greatest strengths of the latter. On the other side our method obtains results comparable to those yielded by the more sophisticated approach by Conte et al. also on very complex scenarios as that proposed by the view 2 of the PETS2009 dataset, but differently from the latter it it does not require a complex set up procedure. 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

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

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

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

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

Chan, A. B., Morrow, M., and Vasconcelos, N. (2009). Analysis of crowded scenes using holistic properties. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 101–108.

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

Conte, D., Foggia, P., Percannella, G., Tufano, F., and Vento, M. (2010). A method for counting people in crowded scenes. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*.

Ellis, A. and Ferryman, J. (2010). Pets2010 and pets2009 evaluation of results using individual ground truthed single views. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*.

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

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

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

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

PETS (2009). *http://www.cvg.rdg.ac.uk/PETS2009/*.

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

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

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