Machine Vision and Applications (2013) 24:1029–1042
DOI 10.1007/s00138-013-0491-3

ORIGINAL PAPER

Counting moving persons in crowded scenes

Donatello Conte · Pasquale Foggia ·
Gennaro Percannella · Mario Vento

Received: 13 February 2012 / Revised: 15 October 2012 / Accepted: 13 February 2013 / Published online: 3 March 2013

Abstract The paper presents a method for estimating the number of moving people in a scene for video surveillance applications. The method performance has been characterized on the public database used for the PETS 2009 and 2010 international competitions; the proposed method has been compared, on the same database, with the PETS competitions participants. The system exhibits a high accuracy, and revealed to be so fast that it can be used in real time surveillance applications. The rationale of the method lies on the extraction of suited scale-invariant feature points and the successive selection among them of the moving ones, under the hypothesis that the latter are associated to moving people. The perspective distortions are taken into account by dividing the input frames into smaller horizontal zones, each having (approximately) the same perspective effects. Therefore, the evaluation of the number of people is separately carried out for each zone, and the results are summed up. The most important peculiarity of the proposed method is the availability of a simple training procedure using a brief video sequence that shows a person walking around in the scene; the procedure automatically evaluates all the parameters needed by the system, thus making the method particularly suited for end-user applications.

Keywords People counting · Crowd density estimation · Video interpretation and understading

1 Introduction

Knowing the number of people present in an area is such an important issue in the framework of video analysis applications that an increasing number of papers on this topic have been proposed in the recent past. Among the applications where this issue is encountered, we can cite video surveillance (an excessive number of persons in an area may constitute a security or safety hazard), public transportation monitoring, and business intelligence (e.g. determining how customers are distributed within a large shopping mall).

Despite the fact that recently some pioneering systems have been made commercially available, further improvements are still necessary, especially concerning their generality and flexibility. Many aspects make the problem really challenging: systems are required to work in real time on general purpose computers, possibly in parallel on different video streams coming from megapixel cameras, so as to supply up to date information on crowd density. Of course, the computational load is crucial but it is a less important issue than the expected accuracy, especially if the output is used for safety issues. The estimation accuracy of the number of people must be sufficiently high, even in the presence of dense crowds. To this concern, it is worth pointing out that in these situations only parts of people bodies appear in the image; the occluded parts generally cause significant underestimation in the counting process; it means that the partial occlusions must be forecast and suitably taken into

This research has been partially supported by A.I.Tech s.r.l., a spin-off company of the University of Salerno (http://www.aitech-solutions.eu).

D. Conte (✉) · P. Foggia · G. Percannella · M. Vento
Dipartimento di Ingegneria dell'Informazione,
Ingegneria Elettrica e Matematica Applicata,
Università di Salerno, Via Ponte don Melillo,
84084 Fisciano (SA), Italy
e-mail: dconte@unisa.it

P. Foggia
e-mail: pfoggia@unisa.it

G. Percannella
e-mail: pergen@unisa.it

M. Vento
e-mail: mvento@unisa.it

account starting from the information about the crowd density. Another crucial point is the unavoidable presence of perspective distortions: people far from the camera appear small while the near ones are significantly bigger. Therefore, counting methods must deal with these perspective issues, in order to obtain an estimation independent of the local scale of the image. Moreover, it is convenient that the system is able to work with uncalibrated cameras, as a fine calibration is generally time consuming and demands for suitable technical skills, not always possessed by the end user. Consequently, the availability of simple tuning procedures that require no knowledge about the internal organization of the algorithm and depend only on simple geometric properties derivable from the scene, is an extremely desirable feature.

The literature presents two different approaches. The direct approach (also called detection-based), relies on the individual detection of the single persons, using adequate segmentation and object detection algorithms; in this way, the number of people is then trivially obtainable. On the other hand, in the indirect approach (also called map-based or measurement-based), the number of people is estimated by measuring the occurrence of suitably defined features that do not require the separate detection of each person in the scene; these features are then somehow put in relation to the number of people.

The direct approach presents the advantage that people detection is often already performed on a scene for other purposes; as long as people are correctly segmented, the count is not affected by perspective, people densities and, to some extent, partial occlusions. On the other hand, people segmentation is a complex task, often providing unreliable outputs, especially in crowded conditions, which are of primary interest for people counting. Recent and well-known examples of the direct approach are [3,16] and [18].

Indirect approaches are based on the extraction of suitably defined measurements and raise the problem of finding an accurate correspondence between these measurements and the number of people. Some methods belonging to this category propose to base the people estimation on the amount of moving pixels [6], blob size [9], fractal dimension [11] or other texture features [15]. Despite their simplicity, promising performance in terms of estimation accuracy are obtained by the approach proposed in [1,4,7]; all them have been submitted to the PETS 2009 and 2010 contests on people counting, and achieved very encouraging results. In particular, in Albiol's paper [1], the authors use the corner points (detected using the Harris' algorithm [8]) as features. Although Albiol's method has proved to be quite robust, confirming the validity of its rationale, its accuracy decreases in presence of highly complex scenes, with large depth variations (people moving in the direction of the camera) and highly crowded moving groups. The authors in [7] explicitly deal with the perspective effects and occlusions; a trainable regressor (the ϵ-SVR algorithm) is used to obtain the number of people as a function of the moving points, a function made complex by the above mentioned effects. Experimental results demonstrated the 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 present a method that is able to obtain performance comparable to those obtained by the method in [7], but at the same time it is much simpler to implement and to set up. The proposed approach deals with the perspective effects on the estimation by subdividing the entire scene in horizontal stripes; the latter have a size depending on their distance from the camera, justifying the hypothesis of a linear relationship between the number of feature points and the people contained. Moreover, a fully automated procedure for training all the needed parameters is presented; a brief video sequence taking a person walking around in the scene is analyzed for directly obtaining all the parameters needed by the system.

The organization of the paper is the following: in the next two sections, we describe the proposed approach and the procedure for automatic training. Finally, we discuss experimental results and draw some conclusions.

2 Rationale of the method

In the literature, there are several interpretations of the problem of people counting, frequently with significantly different assumptions on the conditions under which the counting system is expected to operate. Therefore, before starting to describe our method, it is important to highlight the specific assumptions underlying our method.

We work under the following hypotheses (which, anyway, are those used by most of the approaches in the literature): the camera is stationary; the only objects present in the scene are people; we are interested in measuring only the density of the flow of moving people, not the direction, speed or any other information. The generality of these assumptions is proven by the presence of numerous datasets and benchmarking activities, including contests, sharing these hypotheses.

It is also useful to clarify what we mean by crowded scenes. From the viewpoint of the difficulty of the task, a scene in which individual people images do not overlap with each other is easy to deal with: a tracking algorithm can be applied, and then the count can be easily derived from its outputs. Seldom occlusions also do not present a great problem, since many tracking algorithms can effectively follow a person across a short occlusion. The problem become tougher when the people are partially occluded for a significant portion of the time they appear in the scene. Therefore, we consider as crowded scenes those in which most of the people in the scene are occluded for most of the time they appear.

Now, let us turn our attention to the proposed algorithm. The underlying idea of our approach is that each person framed by a camera can be represented through a small number of salient points. Such points might be located at the boundary of the silhouette and in some other points with high discontinuities as eyes, mouth, nose, clothes, etc. Under the assumption that people are the only moving objects in the scene, the total number of persons can be estimated from the set of the detected salient points exploiting their motion information.

A first attempt at expressing the relation between the salient points and the number of people was proposed by Albiol et al. [1]:

$$P = \omega \cdot N \tag{1}$$

where P is the estimated number of persons in the scene, N is the number of detected salient points, which can be associated to persons by using motion information, while ω is a proportionality constant.

Actually, the hypothesis of representing a person by a constant number of salient points is too simplistic as it does not take into account several issues as, just to name a few, the optical set up of the camera, the position of the person in the scene with respect to the camera, different appearances of the same person across the sequence but also among different persons, obstacles between the person and the camera causing occlusions, variations of the background in time and space, etc.

Thus, a generalization of Eq. 1 is required in order to take somehow into account the above-mentioned factors. Of course, the more general formulation would be an expression of the number of people as a function of the set of the salient points (i.e. depending not only on their number N but on all the information associated to them):

$$P = \phi(\{p_1, \ldots, p_N\}) \tag{2}$$

where $p_1, \ldots, p_N$ are the detected salient points. However, since function ϕ is unknown, and must be constructed through a learning process, its structure must be somewhat constrained. In order to have a formulation that is more general than Eq. 1 while keeping a simple structure that allows for the application of a learning algorithm, we can reinterpret Eq. 1 as stating that each salient point p_i gives an additive contribution to the count P, which is assumed to be independent of the point itself:

$$P = \omega \cdot N = \sum_{i=1}^{N} \omega \tag{3}$$

This reformulation lends itself very well to a generalization that extends considerably the ability to incorporate other factors in the counting without loosing the structural simplicity of the equation: instead of assuming that the contribution of each salient point is a constant, it can be considered as a function of the point itself. Thus, the equation is reformulated as:

$$P = \sum_{i=1}^{N} \omega(p_i) \tag{4}$$

Notice that this formulation makes very easy to include information that is local to each single point (and to its neighborhood); the underlying assumptions are that local information is sufficient for the counting problem, and that this information can be combined additively. Of course, the $\omega(.)$ function may not use all the available information associated with each p_i, but only a suitably chosen subset.

An approach using this formulation is presented in [7] where the $\omega(.)$ functions depends on the distance of the persons from the camera and on the local salient points density. The latter information is adopted as an implicit estimate of the amount of occlusion due to high crowd density. An ϵ-SVR regressor is used to learn $\omega(.)$ from a set of training frames in which the points belonging to each person are manually given a different label. This method showed to be much more effective than [1] when used on difficult scenes. Nevertheless higher performance is obtained at the cost of complex training procedures that have to be carried out for each camera installation. Notice that the performance of this method cannot be improved straightforwardly by incorporating more information in $\omega(.)$: in fact, while in theory, the accuracy of $\omega(.)$ would be increased given an infinite training set; in real cases, the higher complexity of the ϵ-SVR estimator would require a significant increment of the actual training set as demonstrated in the framework of the Statistical Learning Theory by Vapknik and Chervonenkis. This in turn would make the system unsuitable for actual use, due to the higher costs for its training.

In this paper, we propose a method that provides a good compromise between the two opposite requirements of effectiveness and ease of deployment. The result is a method that is able to perform as well as the sophisticated method in [7] but maintaining the overall deployment simplicity of the method [1].

In order to achieve this result, our method is based on the following ideas:

- The $\omega(.)$ function depends only on the distance of the salient point from the image plane:

$$P = \sum_{i=1}^{N} \omega\left(d\left(p_i\right)\right) \tag{5}$$

where $d(p_i)$ is the distance. Note that implicitly $\omega(.)$ depends also on the camera settings, since it has to be trained separately for each camera.

- $\omega(.)$ is modeled as a piecewise constant function, by having the scene divided into horizontal bands and using a constant value for each band; this simplifies both the computation (avoiding the need to perform an Inverse Perspective Mapping) and the training.
- The learning of $\omega(.)$ is performed using an original automatic procedure, that only requires the acquisition of a short video sequence; the training does not require that the persons in the video are manually segmented, as needed by other techniques.

In the following subsections, we will present the details of our method. First, we will describe the salient point extraction and classification, and then estimation of the people count. The automatic training procedure will be described in Sect. 3.

2.1 Salient points extraction and classification

In the literature, there is a wide variety of salient points detectors and descriptors. Some comparative studies in [12] and [13] have demonstrated that Hessian-based detectors have to be preferred with respect to other approaches as they provide better performance in terms of both stability and repeatability. Drawing from the observation that real-world objects are composed of different structures at different scales, Hessian-based detectors, as many other ones, find salient points by analyzing the image at different scales. In the category of the Hessian-based detectors, the SURF algorithm in [2] has gained a large popularity since its first appearance, because of its effectiveness and efficiency. The interest points found by SURF are more independent of scale and, thus, of distance from camera than the ones provided by other detectors. 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.

The points detected are successively classified as static or moving. Under the assumption that persons are the only moving elements into the scene, the classification is aimed at pruning the static points, as they are not associated to people.

We explore two different types of approaches for the classification of salient points: the *classification by motion vector estimation* and the *classification by local difference*. With the first one, we estimate the motion vector associated to each salient point and discriminate between static and moving ones on the basis of the vector magnitude. With the second approach, we rely on the color intensity variations between a set of homologous pixels around the salient point in the current and in a previous reference frame. We expect that the first approach should assure better classification performance, but at a higher computational cost, than the second approach.

2.1.1 Classification by motion vector estimation

Each salient point $p(\mathbf{x})$ detected in the position $\mathbf{x}$ in the frame at time t is attributed a motion vector $\mathbf{v}(\mathbf{x})$, calculated with respect to a reference frame at time t-Δ, and is consequently classified:

$$p(\mathbf{x}) = \begin{cases} \text{moving point} & \text{if } |\mathbf{v}(\mathbf{x})| > 0 \\ \text{static point} & \text{if } |\mathbf{v}(\mathbf{x})| = 0 \end{cases} \tag{6}$$

The motion vector $\mathbf{v}(\mathbf{x})$ is obtained using a block matching technique, by which the block $\mathbf{x}$ is matched to a set of candidate blocks in a reference (earlier) frame. Then, $\mathbf{v}(\mathbf{x})$ is determined as the displacement of the best matching block in the reference frame with respect to the location of the block in the current frame. Matching is based on a criterion that measures the dissimilarity between two blocks.

We used squared blocks (with sides of $2n + 1$ pixels) and evaluated the dissimilarity of the block centered in $\mathbf{x} = (x, y)$ in the current frame I_{curr} and a block shifted by $\mathbf{s} = (k, l)$ with respect to $\mathbf{x}$ in the reference frame I_{ref}, as the mean of the absolute values of the color differences of the pixels in the two blocks:

$$M(\mathbf{x}, \mathbf{s}) = \frac{\sum_{i=-n}^{n} \sum_{j=-n}^{n} \sum_{c=1}^{C} \left| I_{\text{curr}}^{c}(x+i, y+j) - I_{\text{ref}}^{c}(x+k+i, y+l+j) \right|}{(2n+1)^2} \tag{7}$$

where C is the number of the image color channels and $I^c(\cdot, \cdot)$ is the intensity value of the pixel in the cth color channel.

Furthermore, block matching requires the definition of suitable searching algorithms for exploring a search area; the latter, possibly containing the candidate blocks in the previous frame, can be made more or less wide. A fully exhaustive approach extends the search everywhere in the frame with a significant computational expense without considering that the motion of the objects of interest (the persons) is much smaller than the frame size. Simply limiting the search procedure to a window centered on the considered block with a size slightly larger than the maximum possible motion of the persons would significantly reduce the number of candidate blocks without introducing estimation errors. Considering a squared search area with side m, the number of candidate blocks analyzed following this approach is $\Theta(m^2)$. Hereinafter, we will refer to this motion estimation algorithm as *window search*.

A further reduction of the processing time is possible by adopting other search methods [10] (as *three step search*, *2D-logarithmic search*, *cross search*, ...) which determine a

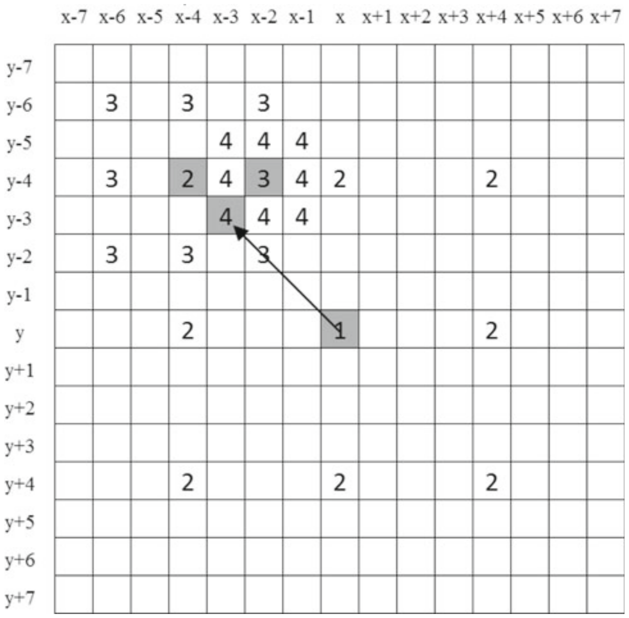

Fig. 1 Example of the *three step search* algorithm with step size $s = 4$. Each number represents the location of candidate blocks at a given iteration. The *shaded boxes* are the location of the best match at each iteration, while the *arrow* is the estimated motion vector

suboptimal solution to the problem reducing the number of candidate blocks to analyze to $\Theta(\log s)$ where s is a parameter called the step size; the search area has a size of $2^{2\cdot s}$ pixels. In Fig. 1, it is shown an example of the analysis performed by the *three step search* algorithm.

In order to reduce the effect of noise, it is possible to incorporate zero-motion biasing into the block matching technique. The current block is first compared with the block at the same location in the previous frame, before doing the search: if the difference between these two blocks is below a threshold (γ_{ZM}), the search is terminated resulting in a zero motion vector without analyzing the neighbor points. Zero-motion biasing allows to reduce the false motion, due to image noise, and the processing time, by eliminating searches; unfortunately, it may produce some false negatives by assigning a zero motion vector to a non static point. Hence, the right value of γ_{ZM} has to be determined as the best trade off between the two opposite effects.

2.1.2 Classification by local difference

Since we are not interested to the exact value of the motion vector, but only in discriminating between moving and static points, we propose to adopt an approach that simply classifies a point as static or moving if the difference between the blocks centered on it in the current and in the reference frame is below or above γ_{ZM}. In particular,

$$p(\mathbf{x}) = \begin{cases} \text{moving point} & \text{if } M(\mathbf{x}, 0) > \gamma_{\mathrm{ZM}} \\ \text{static point} & \text{if } M(\mathbf{x}, 0) \leq \gamma_{\mathrm{ZM}} \end{cases} \tag{8}$$

where $p(\mathbf{x})$ is the interest point, while $M(\mathbf{x}, 0)$ measures the dissimilarity of the block in $\mathbf{x}$ in the current frame and the homologous block in the reference frame. We expect that this approach should preserve the same classification accuracy of the previous approaches, but could significantly reduce processing time as it has to analyze just one candidate block.

2.2 People number estimation

According to the assumptions made at the beginning of this section, the total number P of persons into the scene is estimated using Eq. 5, which relates the contribution of each salient point p_i to its distance from the image plane $d(p_i)$ through the use of the $\omega(.)$ function. The distance from the image plane has been considered because, assuming that the camera lens has a negligible nonlinear distortion, the apparent size of the objects depends on this measure.

The value of $d(p_i)$ could be computed from the position of the point within the image using the Inverse Perspective Mapping (IPM) [17], assuming that the points lie approximately on a common ground plane; however, this technique would require an accurate calibration of the camera parameters, which would complicate the deployment of the system. In order to overcome this problem, we adopt an approach based on two considerations exposed below.

First, if the camera is properly aligned with the ground plane, the set of points having a given distance from the image plane will lie on a horizontal line of the image; this alignment is very easy to obtain during the camera installation if a great accuracy is not required, and it is routinely performed on cameras because a human viewing the scene would find it somewhat disturbing if it is badly aligned with the horizon. Thus, if we consider the points that are within a same narrow horizontal band of the image, we can assume that they have a very similar value for $d(p_i)$, and so we do not incur in a significant error if we use the same $\omega(.)$ value for all of them.

Second, while the above-mentioned approximation is more accurate the narrower the band is, there is no real advantage in having a band whose height is smaller than the apparent height of a person. In fact, in that case, the accuracy on the estimate of $d(p_i)$ would be limited anyway by the error due to the fact that the salient points do not lie all on the ground plane, and, to a lesser extent, to the imperfect alignment of the camera and to the nonlinear distortion of the lens.

Based on this considerations, our method partitions the image into horizontal bands whose height corresponds to the apparent height of an average person. Note that this height is not uniform across the image: it is larger at the bottom, that corresponds to an area closer to the image plane, and becomes smaller when approaching the horizon line. The value of $\omega(.)$

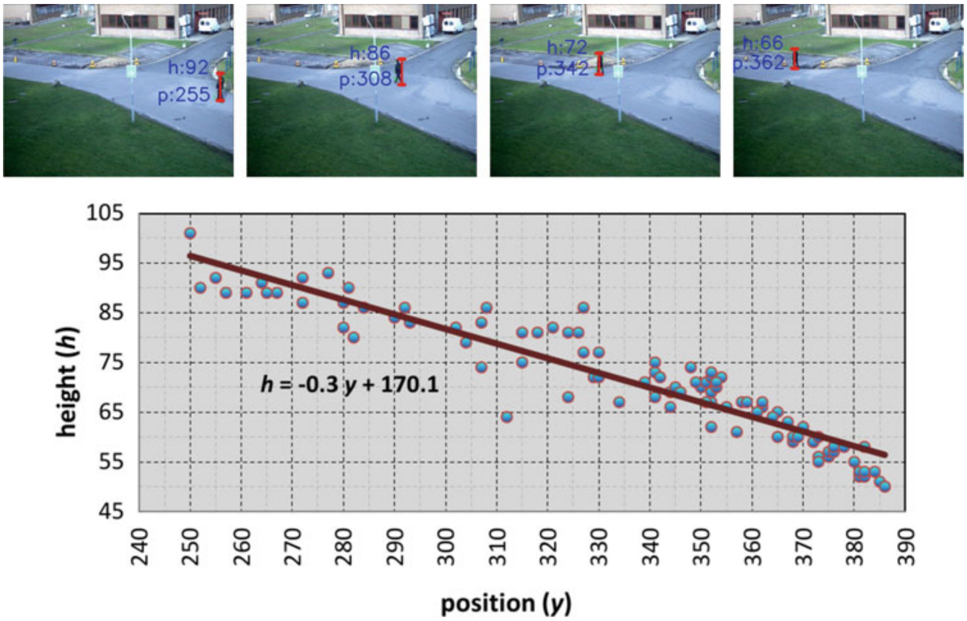

Fig. 2 On *top*, sample frames with (y_i, h_i) couples; on *bottom*, the analytical expression of the function $h = f(y)$

is assumed to be constant for all the salient points lying in a same band, and is determined using a training algorithm that will be presented later.

Accordingly, Eq. 5 is modified as:

$$P = \sum_{i=1}^{N} \omega\left(B_{p_i}\right) \tag{9}$$

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

Notice that since the values of $\omega(.)$ are obtained using a training procedure, there is no need to compute explicitly the distance from the image plane corresponding to each band, and thus to obtain the calibration parameters for the IPM.

Once the set of the weigths $\Omega = \{\omega(B_k)\}$ of the bands have been determined, it is possible to calculate the total number of persons in the scene by Eq. 9.

3 Automatic training procedure

The set up procedure of the method can be decomposed in two phases: the partition of the image into a set of horizontal bands, each having the height corresponding to the apparent height of an average person in the corresponding portion of the scene; and the computation for each band B_k of the value of $\omega(B_k)$ to be used in Eq. 9.

The first phase requires the determination of the height of the bands; these are depending on the geometrical parameters of the systems, such as the focal length and the relative position of the camera in the environment; a closed formula is obtainable, at least in the more general case, if the camera has been suitably calibrated.

However, camera calibration is a costly procedure that requires skilled personnel, not always available at installation time. To overcome this problem, we propose here an automatic procedure using a short video sequence with a person (with height between 1.6 and 1.8 m) that randomly walks within the scene in different directions, so as to obtain a good coverage of the visual area. From each frame, we extract the moving salient points and automatically determine the vertical position y_i of the person in the image coordinate system, and the corresponding apparent height h_i; once a sufficient number of these couples (y_i, h_i) have been extracted, it is possible to obtain, by a regression method, the analytical expression of the function $h = f(y)$ that gives the height in pixel of the person in the position y of the image. More precisely, if we assume that the nonlinear distortion of the lens is negligible, and the camera angle with respect to the horizontal plane is not very large, f is reasonably well approximated by a linear model:

$$h = f(y) \approx a \cdot y + b \tag{10}$$

The a and b coefficients can be determined by linear regression of the (y_i, h_i) data, using the method of least squares. Fig. 2 shows an example with the obtained (y_i, h_i) pairs and the corresponding optimal f coefficients.

The f function is successively used for partitioning the frame in bands by an iterative procedure; the first band, say B_0, is by definition located at the bottom of the frame, and so $y_0 = 0$ (in our reference system the y-axis has the origin at the bottom of the frame); its height is so calculated as $h_0 = f(y_0)$. By iterating the process, the second band B_1 is positioned immediately on top of B_0, at row $y_1 = y_0 + h_0 = 0 + h_0 = h_0$ of the image, and its height is $h_1 = f(y_1)$. In the

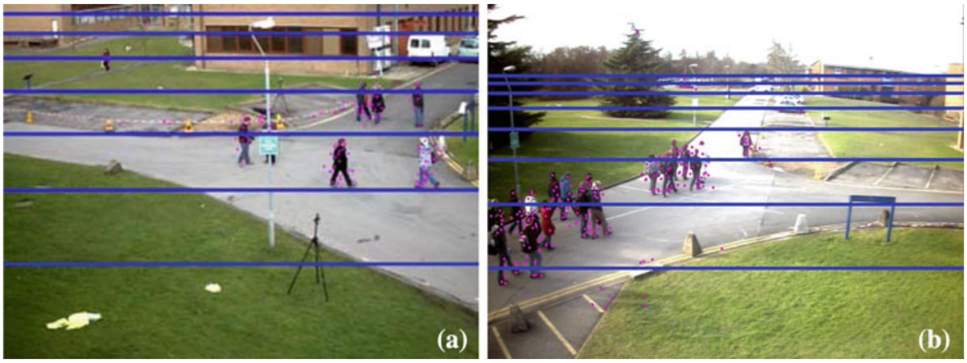

Fig. 3 Subdivision of the frames of the video sequences for the test coming from the PETS 2009 dataset, view 1 *(a)* and view 2 *(b)*. The height of each band approximatively corresponds to the height of a person in real world coordinates

most general case, the position and the height of the ith band are $y_i = y_{i-1} + h_{i-1}$ and $h_i = f(y_i)$. The iterative process is terminated when either the image has been completely scanned or the height of a band is below a certain threshold. The latter situation occurs in installations characterized by a large field depth; in this case the upper part of the frame is excluded from the analysis.

An example of the division of the frame in bands is shown in Fig. 3 where it is possible to visually verify that the height of some bands is perfectly coincident with the height of a person.

The computation of the set of weights Ω is carried out by acquiring another short video taking a group of persons, whose number is known, walking at random across the scene. Actually, it is not required that the number of persons is constant, but this makes the training procedure easier for the user, since he/she has to input this information to the training software only once. Given the video, the values of $\omega(B_k)$ are computed using the least square method, as detailed below. For simplicity of notation, in the remainder of this section we will use the shorthand ω_k to represent $\omega(B_k)$.

Given the training video sequence, the system computes for each frame f the salient points, and then classifies them into static and moving points. The moving points are compared with the previously defined bands, counting how many points lie in each band; let N_{fk} be the number of moving salient points present at frame f in band B_k. The system also knows the number of persons in frame f, that we will denote as P_f (as previously said, P_f is usually constant, although the training method does not require this).

From these data, the training algorithm finds the optimal values for the ω_k by minimizing the following quadratic error measure:

$$E = \sum_f \left(P_f - \sum_k \omega_k \cdot N_{fk} \right)^2 \tag{11}$$

Since the error term is quadratic with respect to the unknowns, ω_k, we can find the minimum by computing the gradient of E and setting it to 0; in this way we obtain for each band B_k an equation:

$$\sum_j \omega_j \cdot \left(\sum_f N_{fj} \cdot N_{fk} \right) = \sum_f P_f \cdot N_{fk} \tag{12}$$

Having as many equations as unknowns, the equation system is easily solvable as long as the corresponding matrix is well conditioned. The matrix will be ill conditioned if not all the bands have been crossed by the people in the scene, or the scene is so crowded that the number of points in some of the bands remains almost constant. The training system detects these situations by computing the condition number of the matrix, and prompts the user for the acquisition of more frames if the provided ones are not sufficient.

4 Experimental results

The performance of the proposed method has been assessed on the PETS2009 [14] dataset and the UCSD Pedestrian Dataset [5] (UCSD in the following).

The PETS2009 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 and 2010 contests. The videos used for the experimentations refer to two different views obtained by using two cameras that contemporaneously acquired the same scene from different points of view. The videos in the dataset were framed at about 7 fps with a 4 CIF resolution (704 × 480 pixels). We used four videos of view 1, namely S1.L1.13-57, S1.L1.13-59, S1.L2.14-06 and S1.L3.14-17, and four videos of view 2, namely S1.L1.13-57, S1.L2.14-06, S1.L2.14-31 and S3.MF.12-43. The videos related to view 1 are the same ones used in the people counting contest held in PETS2009. The videos related to view 2 are 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. The overall size of the UCSD dataset is 2000 frames acquired at 10 frames per second. Each frame is an 8-bit grayscale image, with dimensions

Table 1 Some characteristics of the sequences of the datasets used for assessing the performance of the proposed method

Video sequence	View	Avg # of people	Max # of people
S1.L1.13-57	1	22	34
S1.L1.13-59	1	15	26
S1.L2.14-06	1	26	43
S1.L3.14-17	1	24	41
S1.L1.13-57	2	34	46
S1.L2.14-06	2	37	46
S1.L2.14-31	2	35	43
S3.MF.12-43	2	5	7
UCSD	–	28	52

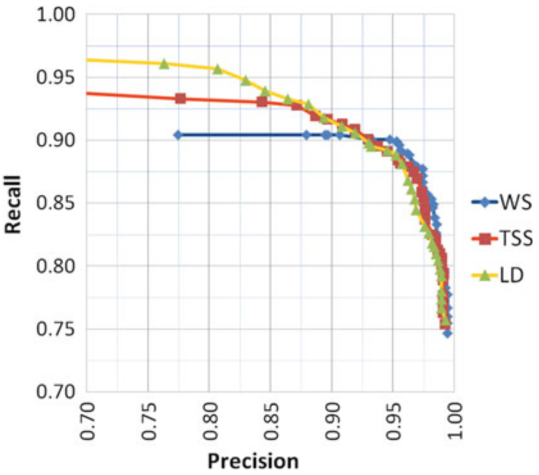

Fig. 4 Moving point classification performance of the WS, the TSS and the LD methods for different values of the bias γ_{ZM} given in terms of *Precision* and *Recall*

238 × 158. The average and the maximum numbers of subjects in the scene for each considered video are reported in Table 1.

On these sequences, we have performed two groups of tests. The tests in the first group were aimed at analyzing the impact of the choice of the algorithm used for recognizing the moving SURF points with respect to counting accuracy and computational load. Specifically, we considered three approaches for points classification: the *window search* and the *three step search*, both based on the motion vector estimation, and the one providing *classification by local difference*. Hereinafter, the above approaches are indicated as WS, TSS and LD. In order to analyze the performance of the proposed method when using the above three points classification approaches, we carried out two types of tests: the first test was aimed at evaluating their accuracy in static/moving points classification and the respective processing times, while in the second test we assessed the estimation error of the proposed people counting method when the above approaches are adopted.

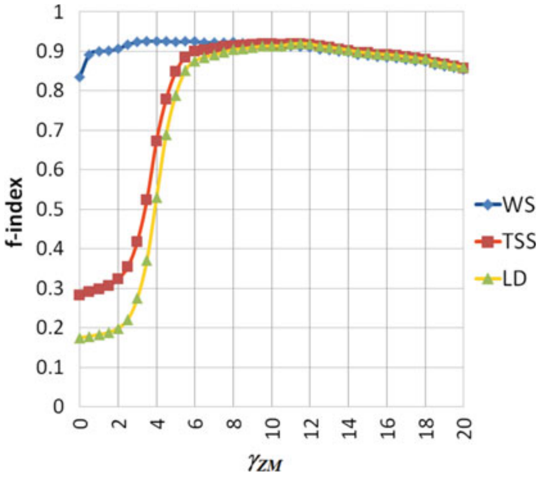

Fig. 5 Classification performance of the WS, the TSS and the LD methods in terms of the *f-index* for different values of the threshold γ_{ZM}

In the second group of tests, we compared the proposed method with respect to other state of the art people counting approaches on the considered datasets. In order to perform this comparison, we used a ground truth reporting the number of visible persons in each frame. The output of each considered method was confronted with the ground truth, computing both an absolute and a relative error measure. The following subsections provide more details on the performed tests and a discussion of the obtained results.

4.1 Accuracy of moving point classification

We collected few dozens of sample frames equally distributed from view 1 and view 2 of the PETS 2009 dataset, using videos that are distinct from those that have been used for evaluating the algorithm performance. The SURF points within these frames have been manually classified as moving or static. The resulting dataset was composed by almost 8.000 points of which about 10 % were moving ones.

We evaluated the classification performance of the WS, the TSS and the LD methods in terms of *Precision* and *Recall*, as a function of the bias threshold γ_{ZM}. Results are shown in Fig. 4. It is interesting to note that for low values of γ_{ZM} (on the left side of the plot), the three approaches tend to have low values of the Precision and high values for the Recall. This can be explained by considering the fact that the lower is γ_{ZM} the lower is the immunity to the noise introduced by the zero motion biasing. The extreme case is represented by $\gamma_{\mathrm{ZM}} = 0$, that is zero motion biasing is not used, for which we obtain the minimum value of the Precision for each approach. However, the most interesting aspect is represented by the fact that for higher values of γ_{ZM} the differences among the three curves are not appreciable. This behavior is more evident when considering the plots in Fig. 5 where the classifica-

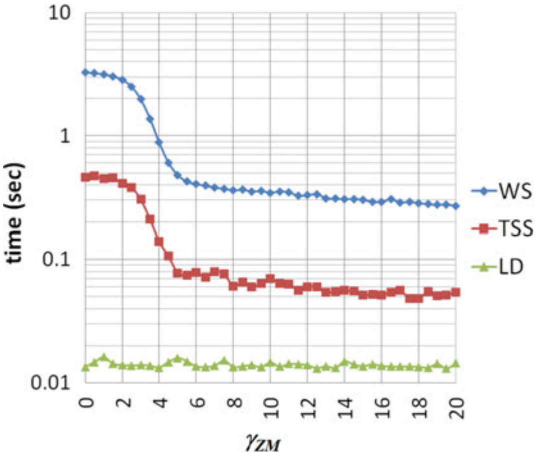

Fig. 6 Processing time per frame of the WS, the TSS and the LD methods for different values of the threshold γ_{ZM}. The y-axis is in log scale

tion performance is reported in terms of the *f-index* (the harmonic mean of Precision and Recall), that allows to compare directly the considered methods. The maximum value of the *f-index* = 0.925 is obtained by WS with $\gamma_{\mathrm{ZM}} = 4$, while TSS and LD have respectively *f-index* = 0.919 with $\gamma_{\mathrm{ZM}} = 11$ and *f-index* = 0.918 with $\gamma_{\mathrm{ZM}} = 11.5$. This means that, at least on the considered dataset, the WS guarantees a slightly better accuracy in moving point classification with respect the other approaches, while TSS does not provide any significant advantages with respect to LD.

In Fig. 6 the plots of the processing time of the WS, the TSS and the LD for different values of the threshold γ_{ZM} are shown. These results were obtained using a notebook with an Intel(R) Core(TM)2 Duo CPU L9400 @1.86 GHz and the following configurations for the moving point classification algorithms: each point was represented through a 9 × 9 pixels block and 21 × 21 pixels search area (this second parameter is only for WS and TSS). It is possible to note that the processing time of WS is one and two orders of magnitude higher than TSS and the LD, respectively.

From the above-described experiment, we can draw the conclusion that for values of $\gamma_{\mathrm{ZM}} \geq 8$ the results of the three approaches are quite similar in terms of Precision and Recall (see Figs. 4, 5). From the point of view of processing time, instead, LD is extremely faster than the other two search strategies (see Fig. 6).

4.2 Accuracy of people number estimation

The training of the system was performed using the proposed automatic training procedure on a video obtained by collecting some short clips from the two datasets containing just one person walking into the scene at different distances from the camera. The frames were selected from other sequences available in the PETS2009 dataset that where not used for the tests. Similarly, we used the first 100 frames of the UCSD dataset for training and the remaining 1900 frames for the test.

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:

$$\mathrm{MAE} = \frac{1}{Q} \cdot \sum_{i=1}^{Q} |G(i) - T(i)|\,,$$

$$\mathrm{MRE} = \frac{1}{Q} \cdot \sum_{i=1}^{Q} \frac{|G(i) - T(i)|}{T(i)} \qquad (13)$$

where Q 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 ith frame, respectively.

In Table 2, we have reported the performance of the proposed method when the WS and the LD methods for points classification are adopted. Performance are reported in terms of the two indices MAE and MRE; we have also reported the

Table 2 Counting estimation error and processing time (in seconds) per frame of the method with the WS and the LD searching strategies

Video (view)	WS			LD		
	MAE	MRE (%)	Time	MAE	MRE (%)	Time
S1.L1.13-57 (1)	1.37	6.9	1.730	1.36	6.8	0.208
S1.L1.13-59 (1)	2.58	15.6	1.395	2.55	16.3	0.201
S1.L2.14-06 (1)	5.44	20.7	1.678	5.40	20.8	0.208
S1.L3.14-17 (1)	2.74	15.1	1.629	2.81	15.1	0.218
S1.L1.13-57 (2)	9.13	23.9	0.952	4.45	15.1	0.207
S1.L2.14-06 (2)	17.74	43.6	0.871	12.17	30.7	0.203
S1.L2.14-31 (2)	6.61	21.7	1.347	7.55	23.6	0.222
S3.MF.12-43 (2)	1.60	34.6	0.637	1.64	35.2	0.206
USCD	4.44	15.1	0.244	3.20	10.9	0.046

Table 3 Counting estimation error of the Albiol's algorithm, of the Conte's and of the proposed ones on the considered datasets

Video (view)	Albiol [1]		Conte [7]		Our	
	MAE	MRE (%)	MAE	MRE (%)	MAE	MRE (%)
S1.L1.13-57 (1)	2.80	12.6	1.92	8.7	1.36	6.8
S1.L1.13-59 (1)	3.86	24.9	2.24	17.3	2.55	16.3
S1.L2.14-06 (1)	5.14	26.1	4.66	20.5	5.40	20.8
S1.L3.14-17 (1)	2.64	14.0	1.75	9.2	2.81	15.1
S1.L1.13-57 (2)	29.45	106.0	11.76	30.0	4.45	15.1
S1.L2.14-06 (2)	32.24	122.5	18.03	43.0	12.17	30.7
S1.L2.14-31 (2)	34.09	99.7	5.64	18.8	7.55	23.6
S3.MF.12-43 (2)	12.34	311.9	0.63	18.8	1.64	35.2
UCSD	4.57	16.58	13.96	50.71	3.26	10.88

average processing time per frame (the experimental settings are the same used for the experiments in Fig. 6).

It is worth noting that although LD is simpler than WS, the people estimation accuracy still remains practically unchanged; surprisingly, there are some cases with a significant performance improvement (videos S1.L1.13-57, S1.L2.14-06 of view 2 from PETS2009 and the video sequences from USCD). Furthermore, it is possible to note that using LD allows to reduce drastically the computational charge making possible to process the video sequences in real time.

4.3 Comparison with other methods

Table 3 presents the comparison between the counting accuracy of our system and that of two other systems participating to the PETS competition; in particular, methods in [1] and [7], both belonging to the category of indirect methods and top ranked as regards the counting accuracy. From the results reported in Table 3, 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 often very close to those obtained by Conte's method. This aspect is more evident if we refer to the results obtained on view 2. On ths UCSD dataset Conte et al. method has particularly poor results (high underestimation), due to the low resolution of the videos. In this case the proposed method has better performance than both the others methods.

In order to have a deeper insight into the behavior of the considered algorithms, Fig. 7 shows the estimated number of people with respect to time for our algorithm, Albiol's and Conte's over four video sequences of PETS dataset while Fig. 8 shows the same estimations for the UCSD dataset.

The behavior of the considered algorithms with respect to the video sequences of Fig. 7 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 on videos characterized by 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 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 of the PETS dataset.

Figure 7a refers to view 1 of video sequence S1.L1.13-59 of the PETS dataset. This video is characterized by isolated persons or very small groups of persons that gradually enter and cross the scene with no or very small occlusions. Figure 7b refers to the same camera view sequence S1.L2.14-06 of the PETS dataset, 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 the other one performs the worst. The presence of occlusion affects 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

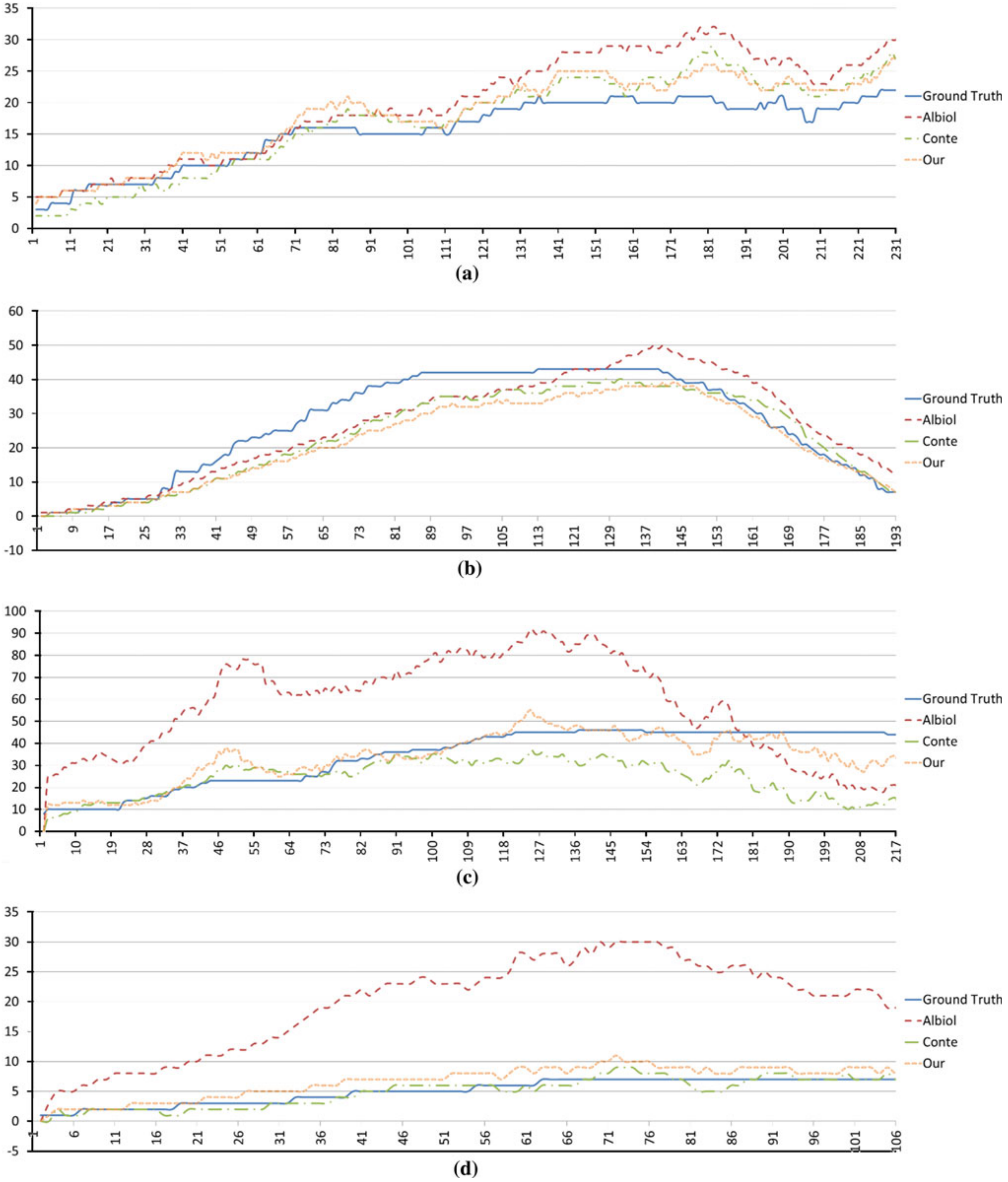

Fig. 7 Curves of the number of people in each frame estimated by the Albiol's algorithm, Conte's and the proposed ones (using LD) together with the ground truth on the video sequence S1.L1.13-59 view 1 (**a**), S1.L2.14-06 view 1 (**b**), S1.L1.13-57 view 2 (**c**) and S3.MF.12-43 view 2 (**d**). On the x-axis, it is reported the frame number

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

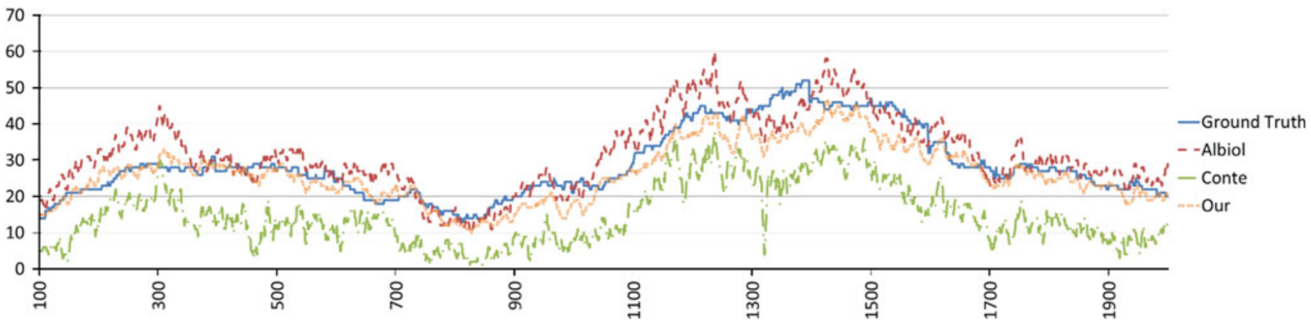

Fig. 8 Curves of the number of people in each frame estimated by the Albiol's algorithm, Conte's and the proposed ones (using LD) together with the ground truth on the video sequences of the UCSD dataset

Table 4 Counting estimation error and processing time (in s) of the proposed approach (using LD) at 4 CIF and CIF resolutions

Video (view)	CIF			4CIF		
	MAE	MRE (%)	Time	MAE	MRE (%)	Time
S1.L1.13-57 (1)	2.31	11.4	0.052	1.36	6.8	0.208
S1.L1.13-59 (1)	2.83	17.4	0.052	2.55	16.3	0.201
S1.L2.14-06 (1)	5.60	23.0	0.053	5.40	20.8	0.208
S1.L3.14-17 (1)	2.53	11.2	0.062	2.81	15.1	0.218
S1.L1.13-57 (2)	10.26	26.8	0.053	4.45	15.1	0.207
S1.L2.14-06 (2)	20.89	51.5	0.051	12.17	30.7	0.203
S1.L2.14-31 (2)	8.53	26.4	0.057	7.55	23.6	0.222
S3.MF.12-43 (2)	2.27	42.9	0.051	1.64	35.2	0.206

respect to Albiol's approach and comparably with respect to Conte's one. More precisely, Albiol's algorithm shows a significant counting error even on frames with few people, while both Conte's and our method are always more accurate in this situation.

Figure 7c, d is related to view 2 of the PETS dataset. 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 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. 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 occlusion that characterizes the sequence, the proposed method performs better than Conte's method (Fig. 7c). 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 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: 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 distance 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 Fig. 7d, related to view 2 of sequence S3.MF.12-43, shows that when there are few isolated persons in the scene Conte's method can provide more accurate results.

Figure 8 refers to the video sequence from the UCSD dataset. This video does not present strong perspective effects. Therefore, Albiol's method does not suffer from the above cited problems, but anyway the proposed method performs better. Because of the low resolution, Conte's method underestimates the number of people on almost all frames.

In Table 4, we report the results of a test aimed at assessing the robustness of the approach with respect to frame resolution. In particular, we considered the same sequences of the PETS dataset used in the previous tests, but at CIF resolution 352 × 240 pixels). The results show that a lower resolution causes a reduction in the counting accuracy, which still remains acceptable, especially if we consider the relevant decrease of the processing time.

5 Conclusions

In this paper, we have proposed 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 performing ones at the PETS 2009 and 2010 contests. The proposed approach is in several cases more accurate than Albiol's one while retaining the same robustness and low computational requirements. On the other hand, our method obtains accuracy results comparable to those yielded by the more sophisticated approach by Conte et al., even on very complex scenarios as the one occurring in view 2 of the PETS2009 dataset; however, differently from the approach of Conte et al. the proposed method does not require a complex set up procedure.

In this paper, we addressed also the complexity of the set-up procedures during installation. In particular, we have proposed a procedure for the automatic training of the system that simply requires the acquisition of two short sequences with a known number of persons that randomly cross the scene.

Among the future works on this topic, we will also investigate recognition-based approaches, using descriptors suited to the recognition of the human shape (such as the Shape Context descriptors), to perform the counting in scenes with the simultaneous presence of different kinds of moving objects (e.g. pedestrians and vehicles), which are out of the scope of the assumptions of our current method.

References

1. Albiol, A., Silla, M.J., Albiol, A., Mossi, J.M.: Video analysis using corner motion statistics. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 31–38 (2009)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
3. Brostow, G.J., Cipolla, R.: Unsupervised bayesian detection of independent motion in crowds. In: IEEE Conf. on Computer Vision and, Pattern Recognition, pp. 594–601 (2006)
4. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–7 (2008)
5. Chan, A.B., Vasconcelos, N.: Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Trans. Pattern Anal. Machine Intell. 30:909–926 (2008). http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70738
6. Cho, S.Y., Chow, T.W.S., Leung, C.T.: A neural-based crowd estimation by hybrid global learning algorithm. IEEE Trans. Syst. Man Cybern. B **29**(4), 535–541 (1999)
7. Conte, D., Foggia, P., Percannella, G., Tufano, F., Vento, M.: A method for counting people in crowded scenes. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 225–232 (2010). doi:10.1109/AVSS.2010.78
8. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988)
9. Kong, D., Gray, D., Tao, H.: A viewpoint invariant approach for crowd counting. In: International Conference on, Pattern Recognition, pp. 1187–1190 (2006)
10. Love, N.S., Kamath, C.: An empirical study of block matching techniques for the detection of moving objects. Tech. Rep. UCRL - TR - 218038, University of California, Lawrence Livermore National Laboratory (2006)
11. Marana, A.N., da F. Costa, L., Lotufo, R.A., Velastin, S.A.: Estimating crowd density with mikowski fractal dimension. In: Int. Conf. on Acoustics, Speech and, Signal Processing (1999)
12. Mikolajczyk, K., Schmid, C.: Scale and Affine Invariant Interest Point Detectors. Int. J. Comput. Vision **60**(1), 63–86 (2004). doi:10.1023/B:VISI.0000027790.02288.f2
13. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Machine Intell. **27**(10), 1615–1630 (2005)
14. PETS: http://www.cvg.rdg.ac.uk/PETS2009/ (2009)
15. Rahmalan, H., Nixon, M.S., Carter, J.N.: On crowd density estimation for surveillance. In: The Institution of Engineering and Technology Conference on Crime and Security (2006)
16. Rittscher, J., Tu, P., Krahnstoever, N.: Simultaneous estimation of segmentation and shape. In: IEEE Conf. on Computer Vision and, Pattern Recognition, pp. 486–493 (2005)
17. Tan, S., Dale, J., Anderson, A., Johnston, A.: Inverse perspective mapping and optic flow: a calibration method and a quantitative analysis. Image Vision Comput. **24**, 153–165 (2006)
18. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. IEEE Trans. Pattern Anal. Mach. Intell. **30**(7), 1198–1211 (2008)

Author Biographies

Donatello Conte received in 2002 a Laurea degree in Computer Engineering from the "Federico II" University of Naples (Italy), and in 2006 a Ph.D. in Information Engineering from University of Salerno, Italy. Since 2006 he is an Assistant Professor of Computer Science at the University of Salerno. His research interests include Structural Pattern Recognition, Realtime Video Analysis and intelligent video surveillance systems; he is the author of several research papers on these subjects. He is a member of the IAPR and of the IAPR Technical Committee 15 (Graph-based Representations in Pattern Recognition) since 2002. He is a reviewer in several international conferences and journals.

Pasquale Foggia received in 1995 a Laurea degree (cum laude) in Computer Engineering, and in 1999 a Ph.D. in Electronic and Computer Engineering from the "Federico II" University of Naples, Italy. From 2004 to 2008 he has been an Associate Professor of Computer Science at the Department of Computer Science and Systems of the same university, while since 2008 he is at the University of Salerno. His research interests include basic methodologies and applications in the fields of Computer Vision and Pattern Recognition; he is the author of several research papers on these subjects. He is a member of the IAPR, and has been involved in the activities of the IAPR Technical Committee 15 (Graph-based Representations in Pattern Recognition) since 1997.

Gennaro Percannella received in 1998 a Laurea degree (cum laude) in Electronic Engineering, and in 2002 a Ph.D. in Electronic and Computer Engineering, both from the University of Salerno, Italy. Currently, he is an Assistant Professor of Computer Science and Artificial Vision at the University of Salerno, where he is a member of the Artificial Vision Research Group. His interests cover the areas of pattern recognition, video and audio analysis and machine learning in artificial vision, with applications like medical and biological image analysis, robotic vision and intelligent video surveillance. He is a member of IAPR. He authored more than 60 research papers in international journals and conference proceedings in the field of Computer Vision and Pattern Recognition. He serves as referee for many relevant journals and conferences and is in the program committee of several relevant international conferences.

Mario Vento is a fellow scientist of the International Association of Pattern Recognition (IAPR). Currently he is a Full Professor of Computer Science and Artificial Intelligence at the University of Salerno (Italy), where he is the coordinator of the Artificial Vision Lab. From 2002 to 2006 he served as the chairman of IAPR Technical Committee TC15 on "Graph Based Representation in Pattern Recognition", and from 2003 as an associate editor of the "Electronic Letters on Computer Vision and Image Analysis". His research interests fall in the areas of Artificial Intelligence, Image Analysis, Pattern Recognition, Machine Learning and Computer Vision. More specifically, his research activity covered Real time Video analysis and interpretation for traffic monitoring and video surveillance applications, Classification Techniques, either Statistical, Syntactic and Structural, Exact and Inexact Graph Matching, Multi-Expert Classification and Learning Methodologies for Structural Descriptions. He has authored over 170 research papers in International Journals and Conference Proceedings and serves as referee for many relevant journals in the field of Pattern Recognition and Machine Intelligence.