PAPER Special Section on Machine Vision and its Applications

Real-Time Counting People in Crowded Areas by Using Local Empirical Templates and Density Ratios

Dao-Huu HUNG[†a)], **Gee-Sern HSU**[††], **Sheng-Luen CHUNG**[†††], *Nonmembers*, *and* **Hideo SAITO**[†], *Senior Member*

SUMMARY In this paper, a fast and automated method of counting pedestrians in crowded areas is proposed along with three contributions. We firstly propose Local Empirical Templates (LET), which are able to outline the foregrounds, typically made by single pedestrians in a scene. LET are extracted by clustering foregrounds of single pedestrians with similar features in silhouettes. This process is done automatically for unknown scenes. Secondly, comparing the size of group foreground made by a group of pedestrians to that of appropriate LET captured in the same image patch with the group foreground produces the density ratio. Because of the local scale normalization between sizes, the density ratio appears to have a bound closely related to the number of pedestrians who induce the group foreground. Finally, to extract the bounds of density ratios for groups of different number of pedestrians, we propose a 3D human models based simulation in which camera viewpoints and pedestrians' proximity are easily manipulated. We collect hundreds of typical occluded-people patterns with distinct degrees of human proximity and under a variety of camera viewpoints. Distributions of density ratios with respect to the number of pedestrians are built based on the computed density ratios of these patterns for extracting density ratio bounds. The simulation is performed in the offline learning phase to extract the bounds from the distributions, which are used to count pedestrians in online settings. We reveal that the bounds seem to be invariant to camera viewpoints and humans' proximity. The performance of our proposed method is evaluated with our collected videos and PETS 2009's datasets. For our collected videos with the resolution of 320x240, our method runs in real-time with good accuracy and frame rate of around 30 fps, and consumes a small amount of computing resources. For PETS 2009's datasets, our proposed method achieves competitive results with other methods tested on the same datasets [1], [2].

key words: *local empirical templates, local density ratios, density ratio bounds, and people counting*

1. Introduction

An open vision problem [3] is to real-time count people from a monocular camera in crowded areas. A great deal of practical applications i.e. public transport security [4]–[6], building security [3], [7], pedestrian traffic management [4], [5], [7], and consumer estimation [7], etc. need the knowledge of the number of people walking through a scene. The prevalence of cameras on the doorstep and the increasing permeation of vision-based techniques to many corners of life make vision-based people counting approaches become potential solutions.

One of the most challenges of counting pedestrians under the view of surveillance cameras is due to humans' swarm behavior. Pedestrians usually gather to form groups walking together, in turn, exhibiting occluded-pedestrian patterns that pose a difficult recognition problem not only for computer vision techniques but also sometimes even for human eyes. It seems to be a bottleneck in applying the advances of computer vision to people counting, especially in crowded areas. To overcome the bottleneck, some existing methods are restricted to work in specific camera viewpoints, i.e. top-view viewpoint [8]–[12]. In surveillance, oblique settings of cameras are preferable due to a wider field of view. Others need a significant amount of work for settings and initialization when camera viewpoints change [4], [5], [13]–[15]. In practice, camera-viewpoint changing during the operation likely occurs due to both human and natural factors, such as carelessness in cleaning and maintaining cameras, and earthquakes, etc. In this context, these methods are prone to poor performance unless repeating such initial setups. In addition, dynamic lighting conditions also cause detrimental effects.

Recently, the common methodology to infer the information of a crowd is based on the local features of individuals appearing in the same local image patches with the crowd. This methodology is successfully applied to tracking people in extremely crowded environments [16], [17]. In the similar methodology, this paper presents a method that is able to simply self-discover unknown scenes captured from an uncalibrated surveillance camera and subsequently to count pedestrians. The method can cope with various viewpoints, ranging from oblique to top views using neither complicated nor manual initial setups. It is essential in practice when we want to apply the method to count pedestrians in many places without repeating such initial settings. During the operation, the method continues to explore the working scene to adapt to unexpected changes of camera viewpoints. Pedestrians are either separated individuals or groups of several occluded people under the view of camera, moving in unconstrained manner. The method can be performed in real-time, consuming a small amount of computing resources. It is applicable to a counting system with several cameras and only one processing device.

To realize the method, we propose to use *local empirical templates* and *density ratios*. The former is the foregrounds induced by single pedestrians in the local image

Manuscript received November 5, 2011.
Manuscript revised January 23, 2012.

[†]The authors are with the Graduate School of Science and Technology, Keio University, Yokohama-shi, 223–8522 Japan.
[††]The author is with the Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan.
[†††]The author is with the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan.
a) E-mail: hungdaohuu@hvrl.ics.keio.ac.jp

DOI: 10.1587/transinf.E95.D.1791

patches. They are clustered upon their features of similar silhouettes along trajectories because pedestrians are different in sizes. Each image patch has the most appropriate LET, resulting tens of empirical templates in the viewing window. Roughly speaking, the empirical templates can outline the foregrounds typically made by single pedestrians. By using LET, perspective effect of cameras can be handled. Given a scene, the empirical templates most appropriate for the scene can be determined when single pedestrians are spotted in the scene. By its nature, the size of single pedestrians is always smaller than that of groups of pedestrians captured in the same image patches with the single pedestrians. Therefore, the clustering process to extract LET for unknown scenes can be performed automatically. During the operation[†], LET are updated when single pedestrians are found, leading the capability of adaptation to unexpected camera-viewpoint changes.

The latter is the ratio between the size of a foreground made by a group of pedestrians and the size of LET considered the most appropriate for the image patch where the group foreground is captured. Because of the local scale normalization between sizes, each density ratios appears to have a bound closely related to the number of pedestrians that induce the group foregrounds. We extract the bounds of density ratios for groups of different numbers of pedestrians in the offline learning phase, and reveal that the bounds seem to be invariant to camera viewpoints and humans' proximity. Both LET and the bounds are used to count the pedestrians in the online counting phase, captured from unknown scenes.

To extract the bounds of density ratios, both empirical approach and the proposed approach in this paper, 3D human models based simulation, can be used. In the former approach reported in our previous work [18], occluded-pedestrians patterns are extracted from a large collection of surveillance videos under different viewpoints and weather conditions. These patterns are manually classified into categories according to the number of people they contain, for computing the density ratios. It is shown from the empirical data that density ratios of the patterns in each classified category are both lower and upper bounded. Both lower and upper bounds of each category are so-called *density ratio bounds*. It seems that the bounds are invariant to camera viewpoints, lighting conditions, and the pedestrians' proximity [18]. The limited observed data, however undoubtedly, cannot cover all situations in reality to prove these above observations.

The latter approach, proposed in this paper, is a 3D human models based simulation in which camera viewpoints and interpersonal distances of pedestrians in groups are easily adjusted. Google Sketchup, a powerful tool to model our world in 3D, is the most suitable simulation environment for our approach. Camera positions in Google Sketchup are changed on the surface of hemispheres whose radii are the distances between the camera and pedestrians, to capture occluded-pedestrians patterns as shown in Fig. 1. In addition, the proximity of pedestrians is able to vary from dense to sparse by manipulating the interpersonal distances between 3D human models. By doing so, the data obtained from the simulation is more sufficient than the data obtained from the former approach, since it is possible to cover most typical situations in reality.

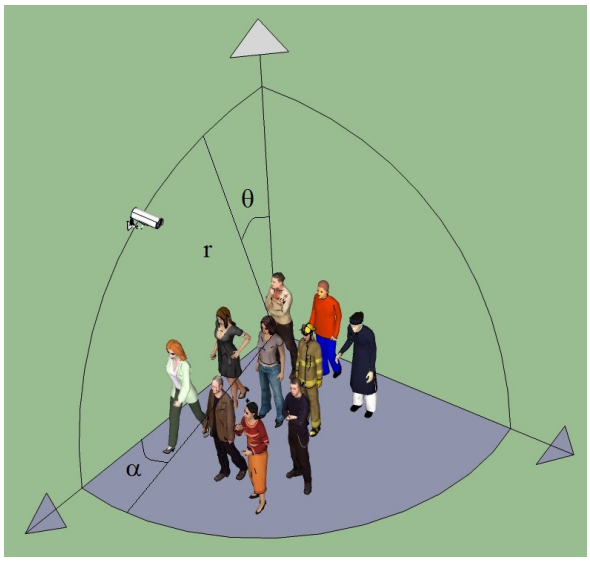

Fig. 1 Our 3D human models based simulation setup in Google Sketchup.

Two assumptions are made for the proposed method of people counting to work: (1) pedestrians must be in upright pose; (2) no vehicles or other moving objects appear in the scene, that is, only people are moving objects taken into consideration. A few cases are also excluded in our study: (1) pedestrians far away from the camera so their sizes appear very small. Quantitatively, the height of the pedestrian is less than 5% height of the view. (2) Groups with serious occlusion, which even challenge human eyes to count.

Since recently, there is a growing interest focusing on the topic of people counting, an increasingly urgent problem is how to evaluate fairly these algorithms' performance. It is recommended to test algorithms on the same Benchmark, in turn, leading fair comparisons. Thus, one dataset of PETS 2009 Challenge was designed for person count and density estimation to meet the growth in the development of the field [3]. Therefore, our proposed method are evaluated on both PETS 2009 benchmark datasets and our video samples collected from real scenes with various parameters, such as viewpoints, lighting conditions, and occlusion.

The rest of this paper is organized as follows. Section 2 provides a review of related works. Our proposed method consisting of offline learning and online counting phases is presented in Sect. 3. In the offline learning phase, we will focus on the proposed 3D human models based simulation to extract occluded-people patterns and density ratio bounds. Experimental results and performance evaluation are reported in Sect. 4. Conclusions and future works are made in Sect. 5.

[†]The online counting phase.

2. Related Works

In this section, we delineate the context of our proposed method in the backdrop of related works, which can be broken into two categories. One trying to segment a group of pedestrians into individuals is quite dominant so far. The second treating a group of pedestrians *en bloc* or as a single entity [4] is considered by few studies.

To count people in entrance/exit gates and in elevator zones, the top-view viewpoint was usually used [8]–[10], [12]. There are almost no occluded people under this viewpoint thus it is easy to segment and count individuals. However, the region of interest is limited by the constraint of ceilings. Through the training, Park *et al.*, [11] obtained the mean and variance values of persons size in each sector of 72-sector-devided images, which are sensitive to the height of camera positions and are used to count people later. His work only reported results of top-view viewpoint cases. In practice, oblique settings of camera are more preferable in spite of the difficulty of occluded people.

Haritaoglu *et al.*, [19] developed W4 system for real-time detecting and tracking multiple people. It counts groups of people by roughly finding heads through corresponding peaks of horizontal projected histogram of group foregrounds. Human shape models were used to interpret the foreground in a Bayesian framework that was implemented by Markov Chain Monte Carlo method [20]. It could segment a group of people into individuals at the expense of high computational cost. In their latter work [21], human motion was globally tracked by using 3D ellipsoid human shape models. Robust results are shown on some difficult sequences, successfully tackling situations of a small number of people moving together, having occlusion, cast shadow and reflection. Cast shadow is removed by the geometrical analysis with a seemingly reasonable assumption of known sun light direction.

People positions were found by searching rectangular-patch pedestrian models across foreground regions in a heuristic way to maximize overlap area [22]. Distance between two legs is represented by periodic and quasi-periodic signals which aid to count people in a group [6], [23]. They show that these signals extracted from a single person and from occluded people are of different patterns. Very small groups of people who perhaps must walk across camera could be detected. It is realized that so far these methods seem to rely on obtaining good quality foreground.

Human appearance models were used to detect people in occlusion. Appearance models for unoccluded people entering the scene are built and subsequently are tracked in the conditions of occlusion [24]–[26]. Xi Zhao *et al.*, [27] presented a people counting approach based on face detection and tracking. A standard face detector located faces tracked them. Free camera viewpoint is obviously achieved but people need to turn their faces to camera. Li *et al.*, [28] trained offline Adaboost HOG (Histogram of Oriented Gradients) features of heads and shoulders to detect people in each frame. They argued that people in occlusion even in crowded environments often presented features of heads and shoulders. Okuma *et al.*, [29] introduced a combination of Adaboost for object detection and particle filter for tracking. However, both of them [28], [29] have two weaknesses in common: stringent requirement of extensive tracking data and inaccurate estimated results in cluttered environments.

Multiple-camera and stereo solutions are another potential class of approaches to resolve the problem of occluded people. Kettnaker and Zabih [14] combined visual appearance matching with mutual content constraints between cameras to identify a same person from different cameras. M2 Tracker system [15] could segment and track people in cluttered environments by using region-based stereo from up to 16 cameras. However, multiple-camera solution is limited to applications of small spatial areas. Kelly *et al.*, [5] discussed a stereo solution for people counting in both indoor and outdoor crowded scenes under various viewpoints. They developed 3D clustering process by using bio-metrically inspired constraints for people detection and track matching process in a weighted maximum cardinality matching scheme. However, in general, multiple-camera and stereo solutions require prior deliberate camera calibration and significant amount of work for registration. If there is any change in camera installation in practice, calibration and registration must be repeated again.

The other promising class of techniques to alleviate occlusion is clustering. Rabaud and Belongie [30] clustered Kanade-Lucas-Tomasi (KLT) features and tracked them across frames in crowded scenes. The KLT tracker, by its very nature, seems to tackle occlusion well in crowded areas. Brostow and Cipolla [31] applied a Bayesian framework for clustering of feature point trajectories to detect individuals in a crowd.

Segmenting a group of occluded pedestrians into individuals provides both the number of pedestrians and the additional information of their locations in the group that is not the gist of the problem of people counting. For dense groups in crowded areas, foreground may not be easily segmented. Therefore, some authors treated a group *en bloc* to estimate its count, leading good performance. Davies *et al.*, [32] used linear fitting to find the relationship between the number of edge pixels and the number of people in a region of interest. Texture measured as different qualitative labels since they argued that image of sparse and dense crowds was often made up of low and high frequency patterns, respectively [33]. The link between these qualitative labels and the counts depends on specific applications. The classification was done by self-organizing map that involved an intensive training. Kilambi *et al.*, [4] provided a solution in the light of using geometric projections, dealing with the entire area occupied by a group as a whole rather than trying to detect individuals separately. Estimated occupied areas were combined with some social statistics of interpersonal distance to determine the count. The results reported usually inaccurate and seemed to be sensitive to the distance from people to cameras. The reason is that these social statistics are not

true, in general, for real applications. In the same spirit, Arai *et al.*, [13] analyzed quantitatively the geometrical relationships between image pixels and their intersection volumes in the 3D world to estimate the number of people. Both these methods must repeat the camera calibration and registration if they want to count people in unknown scenes. Chan *et al.*, [1], [34] adopted Gaussian process regression for segment, internal edge, and texture features which are manually normalized to account for perspective to estimate the number of people. Albiol *el al.*, [2] analyzed moving corners to count people. They assumed that each person, on average, exposes a particular number of moving corners. However, it does not hold in general since the average moving corner per person may vary in accordance with camera viewpoints. Lee and Kim [35] count pedestrians entering and leaving a virtual gate. Foreground detected in the virtual gate is accumulated in a period of time to build a foreground map and is distinguished by two moving directions of pedestrians based on motion vectors. They assume that pedestrians only move in two directions to enter and leave the virtual gate. They argue that one person passing the virtual gate will exhibit a particular number of foreground pixels in the foreground map, depending on humans' speed.

In this paper, we take advantage of treating a group of occluded people *en bloc*. Our system consists of low-cost but effective modules in order to ensure real-time implementation, good accuracy, and adaptation to various camera viewpoints without a complicated or manual initial setup.

3. Proposed Method

The proposed method is comprised of an offline learning phase and an online counting phase. In the offline learning phase, the proposed 3D human models based simulation is performed to extract patterns of both single pedestrians and groups of pedestrians under a variety of camera viewpoints and with different degrees of pedestrians proximity. Foregrounds of these patterns are segmented to produce single foregrounds and group foregrounds, respectively. Clustering single foregrounds, induced by single pedestrians, leads to the generation of Local Empirical Templates. Density ratios of group foregrounds, induced by groups of pedestrians, are defined as the ratio of the size of the foregrounds to that of appropriate LET, considered in the same image patches with the group foregrounds. We take the density ratio calculation for all group foregrounds extracted from the simulation in order to build the distributions of density ratios with respect to the number of pedestrians. Upper and lower bounds of the distributions are extracted as the density ratio bounds. Since these occluded-people patterns obtained from the simulation under different viewpoints and distinct degrees of human proximity, the bounds seem to be independent of these factors. Both Local Empirical Templates and Density Ratio Bounds are used to count pedestrians in a foreground captured online from unknown scenes.

3.1 Local Empirical Templates, Local Density Ratios, and Density Ratio Bounds

The local empirical templates of single foregrounds are represented by their width, height and trajectories or their positions in the image. Depending on different settings, especially the viewpoints of camera, the number of LET in a fixed-view window can be as few as a couple or as many as tens. Experiments on the single LET and group foregrounds reveal the following observations:

- The LET of single foregrounds can be used to discriminate single foregrounds from group foregrounds. The relative sizes of the extracted foregrounds from each other reveal the corresponding crowd densities in many cases, and therefore the foregrounds with sizes smaller than most of the others are likely to be caused by single pedestrians. The decision can be made using a distance measure between the foreground and the LET.
- If the viewing window is divided into $M \times N$ cells (image patches) by a grid, the *local density ratio*, $D(m,n)$, can be defined for each cell (m,n), $n = 1,\ldots,N$; $m = 1,\ldots,M$, as follows,

$$D(m,n) = \frac{S_g(m,n)}{\mathbf{T}_s(m,n)} = \frac{S_g(m,n)}{H_{temp}(m,n) \times W_{temp}(m,n)} \quad (1)$$

  where $S_g(m,n)$ is the size of a group foreground captured at cell (m,n) and its neighbors because a group foreground may not appear in one cell, and $\mathbf{T}_s(m,n)$ is the size of the local empirical template, measured by its width $W_{temp}(m,n)$ and height $H_{temp}(m,n)$ at the cell (m,n). It is observed that although both $S_g(m,n)$ and $\mathbf{T}_s(m,n)$ vary across the viewing window, the variation in $D(m,n)$ appears limited by a bounded range when the crowd density in the group foreground is kept a constant. In other words, the following bounds can be observed,

$$D_M(N_p) > D(m,n,N_p) > D_m(N_p) \quad (2)$$

  where D_M and D_m are the upper and lower bounds of density ratio $D(m,n,N_p)$ of a group foreground containing N_p pedestrians at cell (m,n).

Equation (2) shows that the local density ratios are independent of the cell's location (m,n), and depend on N_p only. Because N_p can be considered an absolute crowd density of a group foreground which has different sizes over the viewing window, the local density ratio $D(m,n,N_p)$ normalizes its size variation to that of the LET.

3.2 Offline Learning and 3D Human Models based Simulation

In the offline learning phase, we setup a simulation environment in Google Sketchup (see Fig. 1). Various 3D human models[†] are manipulated standing on the ground with dif-

[†]Up to ten.

ferent proximity, ranging from dense to sparse groups. The human models wear many kinds of fashion and are in distinct poses. Camera positions are changed on the surface of hemispheres to capture groups-of-pedestrians samples occluded under the views of camera. Radius of the hemisphere is adjustable to generate samples in which pedestrians are far away and close to the camera. This simulation makes sure that most typical camera viewpoints and configurations[†] of groups of pedestrians are covered.

To model the camera viewpoints, a spherical coordinate system is used as shown in Fig. 1 since the camera is moved on the surface of a quarter of a hemisphere.

$$\begin{aligned} &P_{camera} = P(r, \alpha, \theta) \\ &r \approx const \\ &\alpha = [0, 90^o], \Delta\alpha = 10^o \\ &\theta = [0, 60^o], \Delta\theta = 15^o \end{aligned} \tag{3}$$

where, r the radial distance, α the azimuth angle, θ the inclination angle. The angles are measured in degrees. In surveillance, the inclination angle usually varies between 0 and 60 degrees since the camera is mounted well above the human's head. When the inclination angle is 0 degree, we have the top view. In the simulation, given a fixed number of pedestrians N_P, the azimuth angle α is kept in constant and the inclination angle θ is adjusted between 0 and 60 degrees by a step of 15 degrees to capture the occluded-people patterns. Similarly, when the inclination angle θ is kept in constant, the azimuth angle α is adjusted between 0 and 90 degrees by a step of 10 degrees. The human proximity is also adjusted along with the changes of angle settings. Consequently, 50 occluded-people patterns are captured for each N_P. By changing the azimuth and inclination angles in this way, it is confident to capture typical situations of pedestrians in occlusion. In total, we generated 500 occluded-people patterns with N_P varying from 1 to 10. Figure 2 shows the description of adjusting azimuth and inclination angles in the simulation.

The radial distance is kept nearly constant in the simulation since it plays as a scale factor in both numerator and denominator of Eq. (1). If the radial distance is changed, for example, from high to small values, the density and the size of both group foregrounds and LETs increase by approximately equal scale factors. In Sect. 4.2, we will discuss how to choose the radial distance to ensure a good accuracy of the proposed method.

For each camera viewpoint and configuration of groups of pedestrians, two images are generated. One captures the group of pedestrians in occlusion under the view of camera and the other captures only one single pedestrian in the same image patch that plays as an appropriate local empirical template, as shown in Fig. 3. The counts of pedestrian groups, so-called Ground Truth, Np, are generated along with each group. We calculate local density ratios for all occluded-people patterns of these groups by Eq. (1) and split them into categories according to their ground truth, for constructing distributions of density ratios with respect to the ground truth. We aim at proving our observation in Eq. (2) via the distributions.

In this paper, we consider groups containing up to ten pedestrians, that is, the ground truth, Np, will vary from one to ten. Density ratios of all occluded-people patterns computed by Eq. (1) are drawn in a scatter plot in Fig. 4 to visualize the distributions of density ratios with respect to the ground truth. Density ratios of occluded-people patterns having the same ground truth are denoted in Fig. 4 by a same symbol. It is obvious to see that there exists a boundary to separate two categories of group foregrounds with two successive values of ground truth. The decision boundary should be a horizontal line. Given a fixed number of pedestrians N_P, we consider the distribution of local density ratios as a triangle. D_M and D_m, the bounds of local density ratios for various N_P are found at the intersections of these triangles, constructed from the data of the scatter plot in Fig. 4, and are given in the Table 1.

The training dataset used in the offline learning phase in this paper is completely different and more sufficient than the one used in [18] since it covers typical situations, in terms of camera viewpoints and the degree of human proximity. Interestingly, the obtained density ratio bounds given in Table 1 are quite similar to the ones in [18]. That is, it demonstrates the generality of density ratio bounds to cope with various viewpoints and a variety of degrees of human proximity.

The offline learning phase is summarized by the flowchart in Fig. 5.

3.3 Online Counting

The online counting consists of the following steps, as shown in Fig. 6.

1. Foregrounds are firstly extracted from the input video using the GMM [36] and enhanced by applying morphological operator i.e. open and close to remove pepper noise and eliminating moving shadow [37].
2. In the online counting phase, the input scene is unknown. It is necessary to explore the LETs, which are appropriate to the working scene. To extract LETs for the working scene, we keep the sizes of all foregrounds observed in the working scene in the cell buffers. The foregrounds are either single foregrounds or group foregrounds. It is the fact that the sizes of single foregrounds are smaller than the sizes of group foregrounds. In each cell, we choose the foregrounds whose sizes are similar and smaller than the sizes of the others for clustering to extract LET for the cell.
   In our implementation, we design a user-friendly interface to facilitate the LET exploration. Our system marks a bounding box for each detected foreground. If the users observe a few single pedestrians already passing through the scene, the users can stop the LET

[†]Arrangement of pedestrians positions on the ground.

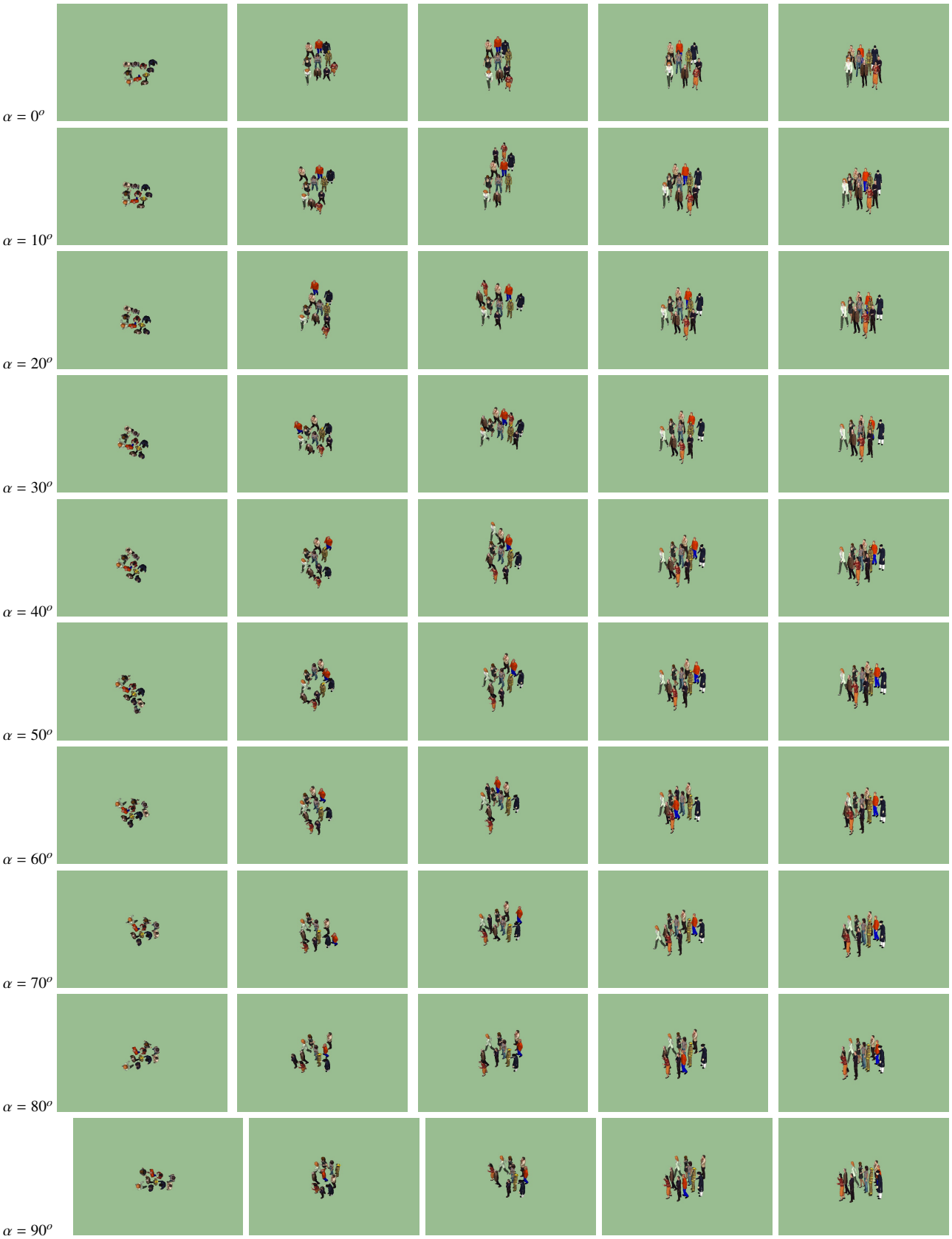

Fig. 2 Description of adjusting azimuth and inclination angles in the 3D human model-based simulation.

(a) Groups of people (b) Appropriate LETs

Fig. 3 Samples of generated images.

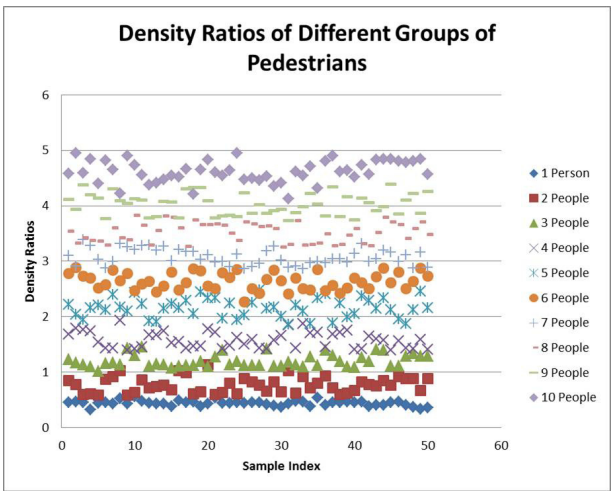

Fig. 4 Distributions of density ratios with respect to the ground truth.

Table 1 Bounds of local density ratio and corresponding ground truth (the number of people in the groups).

Local Density Bounds	N_p
0.3 ~ 0.6	1
0.6 ~ 1.1	2
1.1 ~ 1.4	3
1.4 ~ 1.9	4
1.9 ~ 2.45	5
2.45 ~ 2.87	6
2.87 ~ 3.3	7
3.3 ~ 3.78	8
3.78 ~ 4.25	9
4.25 ~ 4.9	10

exploration and proceed to the counting. This step ensures that our counting system has already observed single pedestrians passing through the scene. The sizes of these single pedestrians are clustered to extract LETs for this scene. Subsequently, our counting system performs the interpolation and extrapolation to determine the LETs for the cells observing no single pedestrians based on the LETs of adjacent cells. During the counting phase, LETs can be corrected via the scheme of scene-based template update that is described later in this section. This process is straightforward for the user without knowledge of computer vision to perform. After this step, our system is ready to count pedestrians passing through this scene.

3. A nearest neighbor classifier is trained in the offline learning phase to discriminate single foregrounds from group foregrounds in the online counting phase. The training samples of single foregrounds and group foregrounds are denoted as the positive and negative ones, respectively. In the offline learning phase, occluded-people patterns and their appropriate LETs are available. The sizes of group foregrounds are normalized to the sizes of LETs to form the negative training samples. LETs by its nature are single foregrounds. Therefore, the sizes of single foregrounds are normalized to the sizes of LETs to create the positive training samples.
In the online counting phase, the size of a given foreground is normalized to the size of its appropriate LET to create the feature vector. The nearest neighbor classifier makes discrimination decisions based on this feature vector.

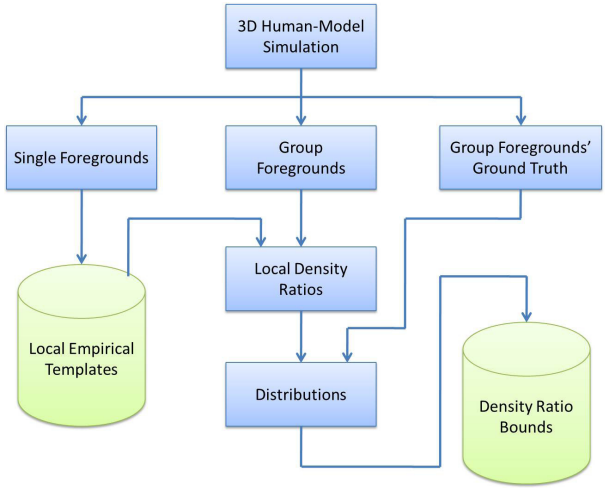

Fig. 5 Flowchart of the offline learning phase.

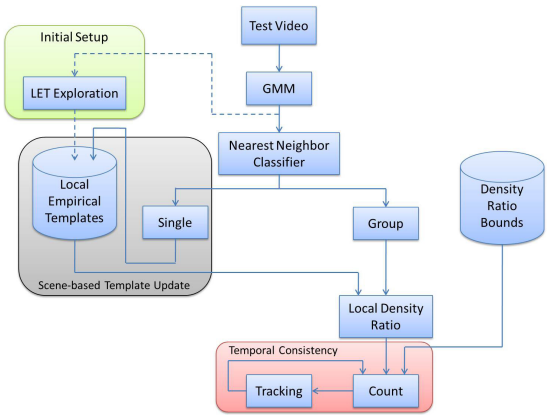

Fig. 6 Flowchart of the online counting phase.

4. The single foregrounds, their trajectories, and the LET that have validated the single foregrounds are kept in a memory buffer for the cells where the single foregrounds are captured. That is, these single foregrounds are used to update corresponding LET, so-called scene-based template update, according to the following formulae.

$$H^{new}_{temp}(m,n) = (1-\alpha)H^{old}_{temp}(m,n) + \alpha H_{sing}(m,n) \quad (4)$$

$$W^{new}_{temp}(m,n) = (1-\alpha)W^{old}_{temp}(m,n) + \alpha W_{sing}(m,n) \quad (5)$$

where, $H_{sing}(m,n)$ & $W_{sing}(m,n)$ are height and width of a detected single foreground, respectively, $H_{temp}(m,n)$ & $W_{temp}(m,n)$ are sizes of LET at the cell (m,n), and α is the learning rate. After the process of LET exploration, the initial LETs are the results of clustering the sizes of a few single pedestrians. The learning rate is set to a high value, for example 0.4, to continue learning the sizes of detected single pedestrians. For each cell, we make a statistics of how many times the LET of this cell is updated. If the number of updates is greater than a particular threshold, i.e. 20 or 30 times, the learning rate is set to a small value, for instance 0.2. The strategy of selecting the learning rate in Eq. (4) and Eq. (5) is unchanged for different scenes. Because single foregrounds may not appear all over the viewing window, interpolations and extrapolations on sections of their trajectories are performed to estimate and extend the most part of regions that foregrounds appear. It is not a rare condition that single foregrounds only appear in certain segments of a walkway because of occlusion, merging, and low contrast to the backgrounds, etc. Therefore, some cells are short of LET, and some LET's trajectories can be broken or segmented. In the online counting phase, trajectories of both single and group foregrounds will be kept in the buffer and analyzed to map out walkway regions. When single foregrounds appear in segments of these regions, interpolation and/or extrapolation based on the observed single foregrounds will be performed to fill in the cells with "virtual" single foregrounds passing through. This step helps to distinguish the regions with foregrounds from the rest without foregrounds, and establish the scene-based spatial distribution of LET with appropriate sizes.
5. With the established scene-based spatial distribution of the LET, the count of pedestrians in a foreground captured in a local cell (m,n) on the viewing window can be estimated by the local density ratio in Eq. (1) with Table 1. Because local density ratio is computed per frame at each cell, each cell will end up with one to a few local density ratios when a foreground moves through. The majority of these density ratios are averaged and considered as the density ratio of the cell. Together with the density ratios evaluated at all cells where the foreground passes, the density ratio of the foreground can be properly determined by a majority voting. To ensure the accuracy of the people count on the foreground, the current count is checked for consistency with the counts obtained along the trajectories of the foregrounds appeared in the previous frames. Possible split and merge of foregrounds are also considered in this consistency check.

It is noted that in the first step of the online counting phase, the moving shadows caused by the sunshine is nearly eliminated by the algorithm in [37]. Sometimes, the density estimation of pedestrians is sensitive to small moving shadows still existing after applying the algorithm in [37]. However, the feature used in this paper is the density ratio that is less sensitive to such small moving shadows. If the small moving shadows make the density of group foregrounds slightly surge, they also make the density of LETs slighly rise, which are regularly updated by the scheme of scene-based template update. The effects of moving shadows are somehow compressed by taking the local density ratio by Eq. (1).

4. Experimental Results and Performance Evaluation

To illustrate the online counting performance of our proposed method with density ratio bounds obtained from the offline learning phase, we have tested it on PETS 2009 Benchmark as well as our own video samples. The former is the most difficult in which contain large groups of occluded people under different illumination conditions. Comparisons with results of other methods tested on the same benchmark are also provided. The latter cover broad spectra of different parameters i.e. illumination variations, camera viewpoints, and many kinds of occluded-people patterns.

4.1 Evaluation on PETS 2009 Benchmark

PETS 2009 benchmark provides a training dataset S0, containing subsets for background model learning [3]. Frames in the dataset S0 contain pedestrians walking through the scene. Therefore, we exploit sizes of these pedestrians in dataset S0 for initializing LETs which are used in the online counting phase.

Figure 7 shows some typical visual results of testing on both subsets in view_001 of dataset S1, L1. The number in the top-left corner of the image is the total estimated count throughout entire image. Manually counted ground truth of each frame is compared with the results of our proposed method, our previos work [18], holistic properties-based method [1] and moving corner-based method [2], tested on the same dataset. These results are adapted from their papers [1], [2]. Figure 8 and Table 2 depict the comparison by graphs sketched in the same coordinate and by Mean Squared Errors, respectively.

Since PETS 2009 datasets provide short sequences for testing, we have to use the first frame of these sequences to initialize the background models of GMM. The first frames of both subsets contain a few people, generating

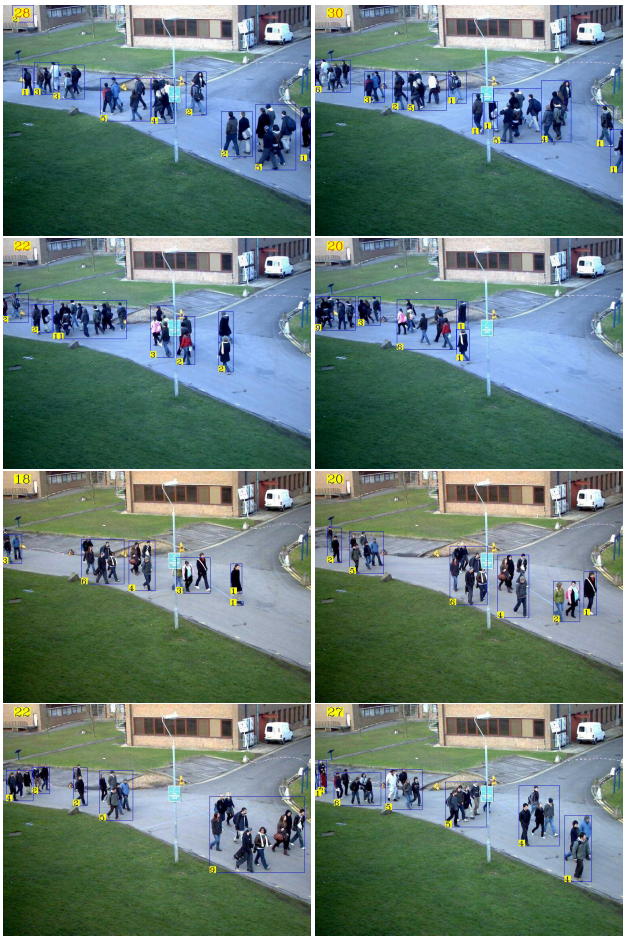

Fig. 7 Results of our proposed method tested on Dataset S1, L1 (view_001); the first two rows are of subset Time_13-59, and the last two rows are of subset Time_13-57.

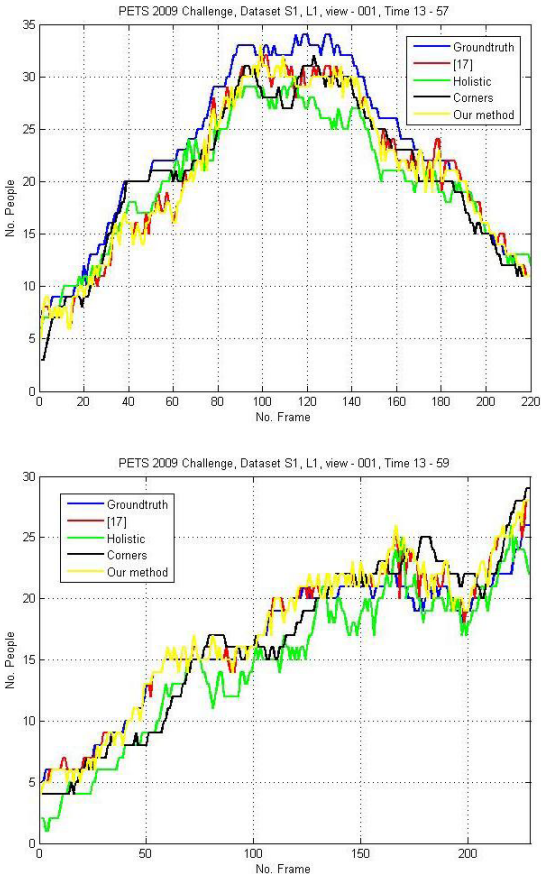

Fig. 8 Performance comparison between our proposed method, our previous work [18], other methods using holistic properties and moving corners, tested on the same datasets, and ground truth. The left graphs are for subset Time_13-57, the right graphs are for subset Time_13-59. Results of methods using holistic properties and moving corners are adapted from their papers [1], [2]

ghosts in foreground images. Since the ghost effect in subset Time_13-57 is more significant than that in subset Time_13-59, results tested on the subset Time_13-59 are better than those tested on the subset Time_13-57. In comparison with other methods tested on the same benchmark [1], [2], the results of our proposed method are competitive.

In addition, to demonstrate the effectiveness of our method, Fig. 9 illustrates the instantaneous estimated counts of a group containing 6 pedestrians during its lifetime appearing in the scene. Table 3 shows the average estimated counts for groups of different numbers of pedestrians. Given a fixed number of pedestrians N_P, we extract the total number of occurrences of groups containing N_P pedestrians from the results of our method, denoted as $O(N_P)$. We also extract the ground truth of these groups GT_i. The average estimated count of the groups containing N_P pedestrians, $\overline{EC}(N_P)$ is the following.

$$\overline{EC}(N_P) = \frac{\sum_{i=1}^{O(N_P)} GT_i}{O(N_P)} \tag{6}$$

The average estimated counts in Table 3 are shown for the method in this paper and the one in [18]. There are only some small difference between the results of our method in this paper and those of our previous work in [18] for groups of more than five pedestrians.

Table 2 Mean Squared Errors of our method and other methods [1], [2], [18] tested on PETS 2009's dataset.

	Subset Time_13-57	Subset Time_13-59
Our method	7.0073	1.7773
[18]	7.9091	2.1703
Holistic properties [1]	12.7364	6.3799
Moving corners [2]	6.6182	5.6507

4.2 Evaluation on In-House Collection

In this section, we further assess the performance of the proposed method by using six input video samples, recorded under various camera viewpoints and in different weather conditions. Our system runs in real-time with average processing frame rate being around 30 fps and consumes a small amount of computing resources. Figure 10 shows some sample frames and the counting results.

The first row of Fig. 10 shows a scene recorded at

Table 3 Average estimated counts of our methods in this paper and in our previous work [18] for groups of various pedestrians.

No. of pedestrians	Our method in this paper				Our method in [18]			
	Time_13-57		Time_13-59		Time_13-57		Time_13-59	
	No. groups	Average counts	No. groups	Average counts	No. groups	Average counts	No. groups	Average counts
1	610	1.02	379	1.01	610	1.02	379	1.01
2	383	2.17	301	2.13	383	2.17	301	2.13
3	166	3.26	42	2.93	166	3.26	42	2.93
4	135	4.07	286	3.70	135	4.07	286	3.70
5	98	5.25	138	4.88	98	5.25	138	4.88
6	49	6.40	103	5.68	38	6.36	104	5.68
7	46	6.96	48	6.29	57	6.96	47	6.30
8	22	7.86	0	0	18	7.72	0	0
9	18	8.72	1	9	19	8.63	1	9
10	34	9.85	0	0	11	9.55	0	0
11	25	11.2	0	0	51	10.59	0	0

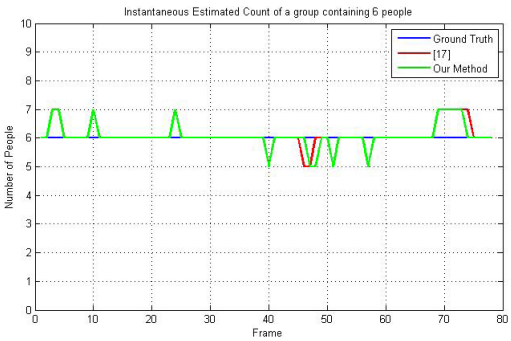

Fig. 9 Instantaneous estimated counts of a group containing 6 pedestrians during its lifetime.

noon. This scene is challenging since the left hand side of the scene is under strong sunshine and the right one is much darker. Its background contains a lot of texture. When a pedestrian moves from bright to dark regions and vice versa, the foreground patterns change considerably that cause some difficulty. Without removing the shadow by the algorithm in [37], the LET exploration will find large LETs in the left part of the scene and smaller LETs in the right part of the scene. By eliminating the moving shadow using [37], LETs in the left part of the scene, extracted from LET exploration, become smaller. Given a foreground and its appropriate LET, we take the density ratio by Eq. (1), turning out the correct numbers of pedestrians. The counting results in the same scene with and without moving shadow are presented in the first row of Fig. 10.

The next three rows of Fig. 10 show different crowded scenes in which many pedestrians move freely, resulting in many distinct patterns of occluded people. They demonstrate the ability of our proposed method in estimating the count of large groups of pedestrians in occlusion. The last three rows of Fig. 10 demonstrate the test of our proposed method against far-away and top-view viewpoints. Although good results often observed, some overestimation and underestimation still occur for groups of nearly full or full occlusion, and groups with moving bicycles.

Our system works not well when camera is mounted too high above the ground floor. In this case, the distance between pedestrians and camera is too far so that their foregrounds account for small number of pixels. Local density ratio computed in Eq. (1) is sensitive to small value of the denominator or small size of LET. Therefore, small changes in the numerator of Eq. (1) make local density ratios vary largely. In practice, broken foregrounds caused by low contrast between background and pedestrian appearance often make the variations of numerator of Eq. (1). However, if the size of LET is large enough, such variations are well tolerated. It is hard to define a specific range for the distance between the camera and pedestrians to ensure an enough accuracy of our method since cameras are different from resolutions. Empirically, the camera viewpoints are chosen so that the sizes of single pedestrians should not be smaller than 20 by 55. In our system, pedestrians far away from camera, their small foregrounds are probably filtered as noise.

In our system, pedestrians are kept tracking from the beginning of entering a scene to the end of leaving the scene. If a pedestrian completes this close process, the counter increases by one, resulting in the total number of pedestrians walking through a scene in a period of time. It is important information for many real applications, i.e. consumer estimation, retailing network planning, and public facility planning [7]. In addition, our system also provides the instant estimated count in every frame that is vital in various applications, i.e. hazardous-situation awareness systems, intelligent walking-signal systems, and public security monitoring systems [3]–[7]. Quantitative evaluation on the In-House collection is given in Table 4. The results obtained by our method in this paper and by our previous work [18] are slightly different. The reasons are that (1) density ratio bounds used in two works are different for groups of more than five pedestrians, and (2) In-House collection contains a few groups of more than five pedestrians.

5. Conclusions

We summarize the proposed concepts of local empirical templates and density ratios with some important characteristics, which are good for counting people in crowded ar-

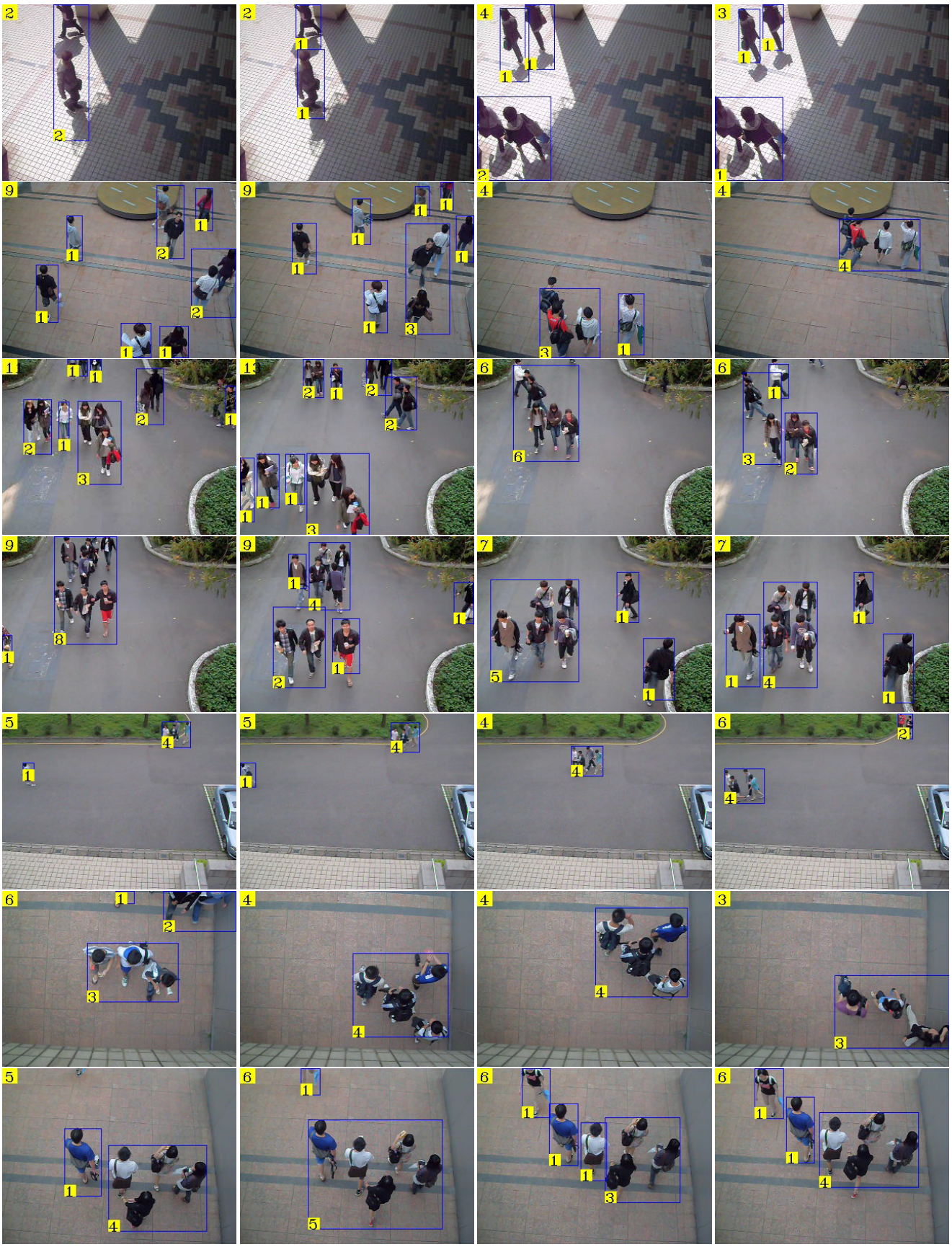

Fig. 10 Results of our proposed method tested on the In-House collection.

Table 4 Total number of pedestrians by our proposed method, [16], ground truth and counting errors, evaluated on the in-house collection. The unit of counter and ground truth is person and the unit of Length is frame.

Video sequences	Length	Ground truth	Our method		[18]	
			Counter	Error	Counter	Error
Sequence 1	30105	48	45	6.25%	45	6.25%
Sequence 2	26979	213	221	3.76%	221	3.76%
Sequence 3	26698	177	172	2.82%	170	3.95%
Sequence 4	4960	13	12	4.95%	12	4.95%
Sequence 5	18540	101	96	4.95%	97	3.96%
Sequence 6	15180	39	38	2.56%	36	7.96%

eas from an uncalibrated surveillance camera. Density of a group of pedestrians varies according to the size of LET or a single pedestrian, captured in the same image patch. The ratio of density of occluded pedestrians to that of appropriate LET is a good feature to discriminate the number of pedestrians. Due to the local scale normalization between sizes, density ratio appears to have a bound, closely related to the number of people who induce the group. More importantly, the bounds are relatively invariant to camera viewpoints and human proximity.

Our proposed method of people counting is composed of two phases: offline learning and online counting. These important characteristics are proved in the offline learning phase in which the 3D human models based simulation is conducted to collect hundreds of typical occluded-people patterns in a variety of viewpoints for extracting the bounds. Both LET and density ratio bounds are used to count pedestrians captured online from unknown scenes. Our proposed method achieves good accuracy, high adaptation to various camera viewpoints, and real-time performance[†] when it is tested on our video samples with standard resolution in surveillance. For evaluation on PETS 2009 datasets, it also shows the competitive results with other methods tested on the same datasets. Since it is low-cost, we can integrate multi-channels working simultaneously in the same PC. That is, we could count people in different places simultaneously without using extra processing devices.

[†]Run in PC with chipset Intel Core 2 Duo T9300, 4 GB Ram, frame rate: 30 fps.

References

[1] A.B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," Proc. Int'l Workshop on PETS, pp.101–108, 2009.
[2] A. Albiol, M.J. Silla, A. Albiol, and J.M. Mossi, "Video analysis using corner motion statistics," Proc. Int'l Workshop on PETS, pp.31–37, 2009.
[3] J. Ferryman and A. Shahrokni, "An overview of the pets2009 challenge," Proc. Int'l Workshop on PETS, pp.25–30, 2009.
[4] P. Kilambi, E. Ribnick, A.J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," Computer Vision and Image Understanding, vol.110, no.1, pp.43–59, 2008.
[5] P. Kelly, N.E. O'Connor, and A.F. Smeaton, "Robust pedestrian detection and tracking in crowded scenes," Image Vision Computing, vol.27, no.10, pp.1445–1458, 2009.
[6] C.J. Pai, H.R. Tyan, Y.M. Lian, H.Y.M. Liao, and S.W. Chen, "Pedestrian detection and tracking at crossroads," Pattern Recognition, vol.37, no.5, pp.1025–1034, 2004.
[7] P.K. Sharma, C. Huang, and R. Nevatia, "Evaluation of people tracking, counting and density estimation in crowded environments," Proc. Int'l Workshop on PETS, pp.39–46, 2009.
[8] J. Barandiaran, B. Murguia, and F. Boto, "Real-time people counting using multiple lines," Proc. WIAMIS, pp.159–162, 2008.
[9] A. Abiol, I. Mora, and V. Naranjo, "Real-time high density people counter using morphological tools," IEEE Trans. Intell. Transport. Syst., vol.2, no.4, pp.204–218, 2001.
[10] S. Xu, X. Chen, W. Sun, and D. Xie, "A robust method for detecting and counting people," Proc. IEEE Int'l Conf. on Aud. Lang. and Image Proc., pp.1545–1549, 2008.
[11] H. Park, H. Lee, S.I. Noh, and J. Kim, "An area-based decision rule for people-counting systems," In MRCS 2006, LNCS, vol.4105, pp.450–457, 2006.
[12] X. Yuan, X.Y. Wei, and Y.D. Song, "Pedestrian detection for counting applications using a top-view camera," IEICE Trans. Inf. & Syst., vol.E94-D, no.6, pp.1357–1361, June 2011.
[13] H. Arai, I. Miyagawa, H. Koike, and M. Haseyama, "Estimating number of people using calibrated monocular camera based on geometrical analysis of surface area," IEICE Trans. Fundamentals, vol.E92-A, no.8, pp.1932–1938, Aug. 2009.
[14] V. Kettnaker and R. Zabih, "Counting people from multiple cameras," IEEE Int'l Conf. Multimedia Computing and Systems, pp.267–271, 1999.
[15] A. Mittal and L.S. Davis, "M2 tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," Int'l J. Computer Vision, vol.51, no.3, pp.189–203, 2003.
[16] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," Proc. ECCV, pp.1–14, 2008.
[17] L. Kratz and K. Nishino, "Tracking with local spatio-temporal motion patterns in extremely crowded scenes," Proc. CVPR, pp.693–700, 2010.
[18] D.H. Hung, S.L. Chung, and G.S. Hsu, "Local empirical templates and density ratios for people counting," Proc. ACCV, vol.4, pp.90–101, 2010.
[19] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-time surveillance of people and their activities," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.809–830, 2000.
[20] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," Proc. CVPR, pp.459–466, 2003.
[21] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.9, pp.1208–1221, 2004.
[22] O. Masoud and N.P. Papanikolopoulos, "A novel method for tracking and counting pedestrians in real-time using a single camera," IEEE Trans. Veh. Tech., vol.50, no.5, pp.1267–1278, 2001.
[23] R. Cutler and L.S. Davis, "Robust real-time periodic motion detection, analysis, and applications," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.781–796, 2000.
[24] A.M. Elgammal and L.S. Davis, "Probabilistic framework for segmenting people under occlusion," Proc. ICCV, pp.145–152, 2001.

[25] A. Senior, "Tracking people with probabilistic appearance models," Proc. ECCV Workshop on PETS, pp.48–55, 2002.
[26] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Tracking people by leaning their appearance," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.1, pp.65–81, 2007.
[27] X. Zhao, E. Dellandrea, and L. Chen, "A people counting system based on face detection and tracking in a video," Proc. AVSS, pp.67–72, 2009.
[28] M. Li, Z.X. Zhang, K. Huang, and T.N. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," Proc. ICPR, pp.1–4, 2008.
[29] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. Lowe, "A boosted particle filter: Multitarget detection and tracking," Proc. ECCV, pp.28–39, 2004.
[30] V. Rabaud and S. Belongie, "Counting crowded moving objects," Proc. ICPR, pp.705–711, 2006.
[31] G. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," Proc. CVPR, pp.594–601, 2006.
[32] A.C. Davies, S.A.M. Yin, and S.A. Velastin, "Crowd monitoring using image processing," Electron. Commun. Eng. J., pp.37–47, 1995.
[33] A.N. Marana, S.A. Velastin, L.F. Costa, and R.A. Lotufo, "Automatic estimation of crowd density using texture," J. Safety Sci., vol.28, no.3, pp.165–175, 1998.
[34] A.B. Chan, Z.S. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," Proc. CVPR, pp.1–7, 2008.
[35] G.G. Lee and W.Y. Kim, "A statistical method for counting pedestrians in crowded environments," IEICE Trans. Inf. & Syst., vol.E94-D, no.6, pp.1357–1361, June 2011.
[36] C. Stauffer and W.L.R. Grimson, "Learning patterns of activities using real-time tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.747–757, 2000.
[37] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, "Improving shadow suppression in moving object detection with hsv color information," Proc. Conf. on ITSC, pp.334–339, 2001.

Dao Huu Hung was born in Hanoi, Vietnam. He received B.Sc. from Hanoi University of Technology and M.Sc. from National Taiwan University of Science and Technology, both in Electrical Engineering, in 2007 and 2010, respectively. In 2011, he joined Panasonic R&D Center Vietnam and spent one month at Panasonic Advanced Technology Development Center at Nagoya, Japan as a visiting R&D Engineer. Currently, he is Ph.D. student at Hyper Vision Research Laboratory, Graduate School of Science for Open and Environmental Systems, Keio University at Yagami, Japan. His research interests include image processing, computer vision, pattern recognition, and applications to surveillance systems.

Gee-Sern Hsu received his M.Sc. and Ph.D. in Mechanical Engineering from University of Michigan at Ann Arbor, in 1993 and 1995, respectively. He served as R&D Manager from 2001 to 2003 and Director from 2004 to 2007, both in Penpower Technology, Taiwan. Since 2008, he has been an Assistant Professor of Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taiwan. His research interests include computer vision, image analysis and recognition, intelligent video surveillance.

Sheng-Luen Chung received the B.S. degree in electronic engineering department from the National Chiao-Tung University, Taiwan, R.O.C., in 1985, and the M.S.E. and Ph.D. degrees from the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 1990 and 1992, respectively. Since 1992, he has been with the Electrical Engineering Department of the National Taiwan University of Science and Technology, Taiwan, R.O.C., where he is now a Professor. His current research interests include system identification, embedded system design, applications of supervisory control, smart home, and intelligent video surveillance. Dr. Chung was a recipient of the 1994 IEEE George E. Axelby Outstanding Paper Award from the IEEE Control Systems Society in 1994 for a paper coauthored with S. Lafortune and F. Lin. He is a senior member of IEEE.

Hideo Saito received his B.E., M.E., and Ph.D. degrees in Electrical Engineering from Keio University, Japan, in 1987, 1989, and 1992, respectively. He has been on the faculty of Department of Electrical Engineering, Keio University since 1992. From 1997 to 1999, he stayed at Robotics Institute, Carnegie Mellon University as a visiting researcher. Since 2006, he has been a Professor of Department of Information and Computer Science, Keio University. His research interests include computer vision, mixed reality, virtual reality, and 3D video analysis and synthesis. He has been an area chair of ACCV 09, ACCV 10, and ACCV'12, general co-chair of ICAT 06, ICAT 08, and general chair of MVA 09. He is a senior member of IEEE.