Pedestrian head detection and tracking using skeleton graph for people counting in crowded environments

Kheir-Eddine AZIZ, Djamel MERAD and Bernard FERTIL
LSIS - IM
UMR CNRS 6168,
163, Avenue of Luminy
13288 Marseille Cedex 9.
kheir-eddine.aziz, djamal.merad, bernard.fertil@univmed.fr

Nicolas THOME
UPMC - LIP6
Box courier 169
4 place Jussieu
75252 PARIS Cedex 05.
Nicolas.Thome@lip6.fr

Abstract

This paper describes a new head detection method for people counting in crowded environments from a single camera. Our method adopts skeleton graph to distinguish person among people in crowded enviroments. The usage of skeleton graph is the main difference between this method and the traditional ones. Firstly, the skeleton graphs are calculated for each selected blob in the scene after foreground estimation. Then, we explore the structural property of each blob for a head detection and to predict a number of people. Each detected head in a skeleton silhouette is identified as independent state or partially occluded state, and during tracking every state is updated. Finally, the experimental results are shown to demonstrate the robustness of our method.

1 Introduction

People-counting systems have been widely studied in many commercial and public locations, such as theaters, shopping centers, stations, etc. Many people are passing in these areas so it is important to recognize aspects of their movements. Various methods which estimate the number of people in input images have been previously proposed. They can be divided into three approaches:

Trajectory clustering approach. In this approach people are counted by tracking and identifying visual features over time. The feature trajectories which exhibit coherent motion are clustered and a number of clusters gives an estimation of the number of pedestrians. For example, Antonini et al.[3] proposed a people counting method in which the trajectories are obtained by a tracking algorithm. The trajectories are clustered The trajectories are clustred according to their lengths and spatial locations. This approach estimates the number of pedestrians who passed within a specific time.The inconvinient of this approach is that a real-time processing is difficult.

Feature-based regression approach. This approach estimates a number of pedestrians by a regression on features extracted from an input image, e.g. neural networks,[4][5]. Nevertheless, pedestrian positions in the input image cannot be estimated by these methods which even despite this fact cannot be executed in real-time processing.

Individual pedestrian detection. In this scheme, the proposed algorithm estimates the number of pedestrians who were detected in input images. For example: [16][8][2][15]. Moreover, these methods cannot be applied to very crowded scenes with significant occlusion because all pedestrians need to be detected and segmented.

In other approches such as [7][12][13][1], a human silhouette is segmented before analyzing its shape properties so as the body parts are extracted. Davis et al.[7], Fujiyoshi et al [9] perform the labeling by first determining a human pose among a set of predefined ones in the first moment. Nevertheless, this preprocessing scheme is inevitably determined to fail in some cases, decreasing the overall system performances. The approach proposed by Mori et al.[12] identically retrieves the human pose before performing the labeling. Among pre-stored sets of exemplar 2D views for where key points are manually identified. Fujiyoshi et al. [9] proposed a motion analysis of a human"'star skeleton'" in a video stream. This approach is used only to treat the independent target state and not for testing in a crowded area. Thome et al.[13] used a properly labeling human body parts in video sequences for tracking and motion interpretation. Alahi et al. [1] proposed an approach based on generative model. They tryied to minimize a differentce between a synthetic image and an observed image.

Our proposed method explores a property of the graph skeleton and labeled body parts from the silhouette to deal with occlusions among people in motion. In this paper, we propose a new skeleton-based head detection approach which can count people with a good accuracy. This method does not need to segment all pedestrians. The experimental results are shown to demonstrate the robustness of our method.

2 People counting method using skeleton graph

The proposed system is illustrated in the Fig.1. A input image is segmented into blobs of moving objects, using background subtraction. We extract a skeleton graph for each blob. Finally, the number of people is estimated in each blob by a head detection in the skeleton graph.

Figure 1: The counting people system. (a) Input image, (b) Background subtraction, (c) Skeleton graph, (d) Head detection and pose estimation, (e) Head tracking.

2.1 Background subtraction

For a pre-recorded sequence such as the PETS and CAVIAR dataset, to compute foreground masks we use a forward-backward approach devloped by Ge et al.[10]. The adaptive background subtraction is run forward in time from a first frame to a last one, and a second process runs backward in time from the last frame to the first one. We assume a rapid change in illumination occurs at time T, with gradual illumination changes before and after that. The forward background subtraction pass will produce clean foreground masks for times less than T and then suffer degraded performance at time T before a gradual recovering. Likewise, the backward pass will produce clean foreground masks for times greater than T, but suffer degraded results for a short period of time before T.

2.2 Skeleton graph computing

For each detected region (individual/group human), we compute the graph skeleton for each of them using the approach developed by Thome et al [13]. The first step in order to detect visible segments corresponding to body parts in the image consists thus in determining the skeleton points. Whatever the strategy used, the main difficulty is related to its sensitivity to noise. To overcome this shortcoming, the silhouette is smoothed. This is achieved by computing the Fourier Descriptor of its outer contours. At this stage, the skeleton is determined by computing the Delaunay triangulation of the smoothed reconstructed silhouette. This approach is the most adapted for the following reasons. First, the computation is fast and accurate. Moreover, the Delaunay triangle structure is isomorph to the graph by containing neighborhood information.

For getting a set of segments, the skeleton point sequences are polygonalized afterwards. This step consists in identifying a set of points and the link between them, representing the segments. We point out that each skeleton point corresponds to the center of the circumscribed circle to each Delaunay triangle. Each link between two skeleton points is associated with a quantity corresponding to a mean radius of the segment along the skeleton points. This quantity is the number of skeleton points between two centers of the circumscribed circle.

2.3 Head detection

The skeleton points may be classified in depence on their neighborhood degree. Points having a single neighbor corresponds to end points. Points having more than two neighbors define starting points for segments. Points having exactly two neighbors correspond to points on a continuous

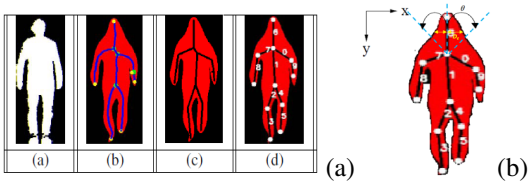

Figure 2: (a) Skeleton graph computing. (b) Head detection.

curve among the starting points and the end points. We can notice that the end points and the strating points of segments corresponding to body parts. Fig. 2-a illustrates the skeleton computation. An extracted silhouette, the first set of segments after the polygonalization and in the last set after removing small edges are represented in Fig. 2-a.(a, b, c, d) respectively.

In the crowded area, the heads are considered as the most apparent parts of the skeleton. At this step, for detecting head we are interested only in the points' set having a single neighbor, the segment corresponding to the extreme node is subsequently taken and its degree inclination compared to the vertical axis is calculated. If the degree tilt is included in $[-\theta, \theta]$, the segment is classified as a head of a person (Fig. 2-b, Algorithm 1).

Algorithm 1 Head detection

Require: $skeleton\_list$: list of the graph skeleton

```
for i = 0 to skeleton_list.size do
  for j = 0 to segments_list.size_i do
    x_neighbor ← neighbor(segment_k)
    if x_neighbor = 1 then
      y_begin_seg ← cordinate(segment_k)
      y_end_seg ← cordinate(segment_k)
      if y_begin_seg < y_end_seg then
        angle ← symetric_angle(segment_k)
        if angle ∈ [−θ, θ] then
          head ← true
        end if
      end if
    end if
  end for
end for
```

2.4 Head pose estimation

In the previous step, the shape of detections can be corrupted by the noise which induces consequently the false detections. To verify the validity for each detected head, we must estimate the distance between the local reference model of a head in the world coordinate system $\{x_h, y_h, z_h\}$, which his size is assumed known $(20cm \times 20cm)$, and a reference detection in the camera coordinate system $\{x_c, y_c, z_c\}$ which is supposed to be calibrated. This distance is according to a pre-determined threshold (two meters in our case) between the two references to accept or reject one detection (see Fig.3) .

This estimation consists in finding the rigid transformation (R, T) minimizing calculated error (as the sum of error squares) of the one of two collinearity equations (in the

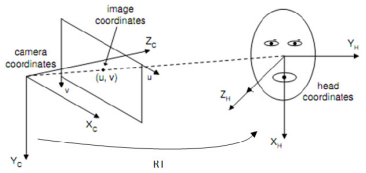

Figure 3: Head pose estimation.

image space or in the object space). We used the method proposed by Lu et al.[11]. named Orthogonal Iteration (OI) algorithm. To estimate the objects pose, this algorithm used an appropriated error function defined in the objects space. The error function is rewritten in order to accept an iteration based on the classical solution of the 3D pose estimation problem, called absolute orientation problem. This algorithm gives exact results and converges quickly enough.

2.5 Head tracking

In this work, we adopt a framework based particle filter similar to the one used in [14]. Specifically, a head is modeled as an rectangle centered at (x, y) and with size (H_x, H_y). In the first frame, we use the head detection algorithm (described in section 2.3) to detect the location and size of the head rectangle. The dynamics of the (moving) head at time t are described by a state vector St consisting of the following eight components $\{x, y, X_v, Y_v, H_x, H_y, H_{vx}, H_{vy}\}$ where (x, y) represent a center location of the head rectangle, (X_v, Y_v) represent the motion velocity, (H_x, H_y) are the lengths of the half axes, and (H_{vx}, H_{vy}) are the corresponding scale changes on the axes.

We use a set of properly weighted random samples. The weight for each sample will be computed according to the new observations, this weight is based on the color histogram difference between the measured color distribution and the model color distribution (computed at the previous time instant). Then the mean state, which specifies the trackedhead, is estimated by $E[S_t] = \sum_{n=1}^{N} w_t^{(n)} s_t^{(n)}$ where $s_t^{(n)}$ is a sample of the state vector and $w_t^{(n)}$ is the corresponding weight.

3 Experiments

In the begining, we applied the proposed method to the experiment of detecting head of people for finding the number of pedestrians passing an outdoor area. We evaluate the single-view counting on the PETS crowd counting task. People counts for the S1.L1.13-57 sequence are shown in Figure.4.

We observed the robustness of our method for people's head detection in most cases and in different situations (independent human, partial occlusion humans), except the case of complete occlusion where the heads have the same abscissa in the coordinate system of the image.(Figure.6-Frame 73).

Fig.5 provides the evaluation of the counting people per region task. Note that the y axis on this graph represents the average error in number of people per frame, where the

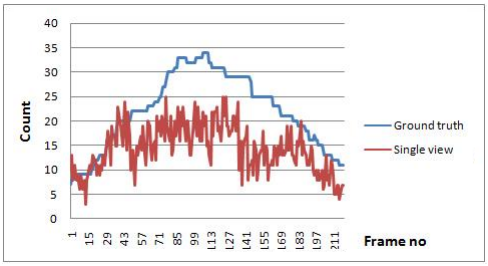

Figure 4: People count in the individual frames from PETS sequence S1.L1 using single view.

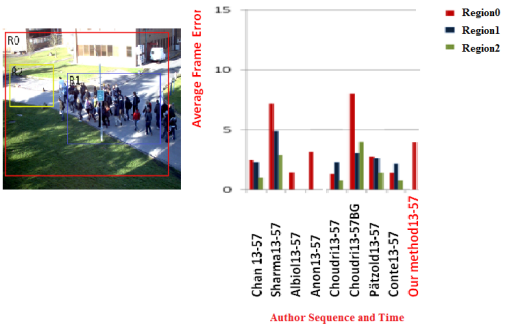

Figure 5: Counting People Evaluation

lower the value, the better the performance per frame. Our method is compared to a wide variety of methods, which have been proposed and tested in this category and from Fig.5 it can been seen that the most of the methods and their variants have consistent and comparable performance.

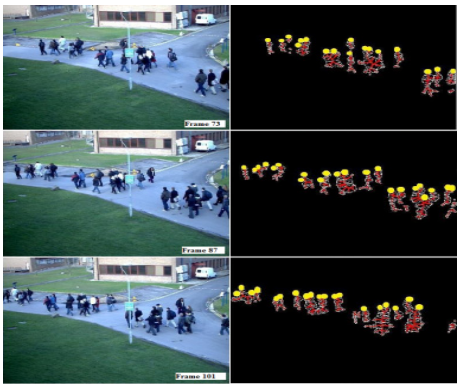

Figure 6: Sample head detection results on PETS data in image plane View1.

The integration of tracking in the process of counting reduces the average error due to complete occlussions. The main idea is to launch a tracking process on the individual cases of missing his head in fustionned with others blobs. We are able to distinguish the head before, during and after the fusion of the people. In this case, the counting performance evaluation is done with the accuracy in the group of human extracted in different videos CAVIAR dataset. To verify the performance of the proposed method, the results are compared to the people detection method proposed by *Dalal et al* (Individual pedestrian detection approach, see

Video	Number of the group of humans	Total people	People detection (Dala & al)	Proposed method
1	2 (3 people + 2 people)	5	50%	80%
2	2 (2 people + 2 people)	4	50%	100%
3	2 (2 people + 2 people)	4	50%	75%
4	1 (2 people)	2	50%	100%
5	1 (2 people)	2	50%	100%
6	1 (3 people)	3	66%	100%
7	7 (2 + 4 + 3 + 4 + 2 + 4 + 2)	21	50%	85%
8	2 (2 people + 2people)	4	50%	100%
9	3 (2 people + 2 people + 2 people)	6	33%	83%
10	1 (5 people)	5	40%	80%
11	11(3 + 4 + 4 + 2 + 2 + 3 + 4 + 2 + 2 + 4 + 3)	33	56%	81%

Figure 7: Accuracy of the proposed method.

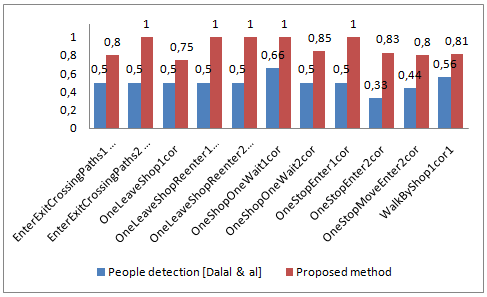

Figure 8: Counting people accuracy for the differente group of humans.

section 1).

For an independent person, the performance of both methods are the same. Table 7 compares the accuracy of the proposed method and the people detection method. In the table, we can see that the results of the method based on people detection and the proposed method are significantly better for all sequences which are containing crowded scenes. Moreover, the proposed method outperformed more than the people detection method as shown in Fig.8. Our method successfully estimates the number of people when the scene is not busy and when the number of people in heavily crowded scenes due to the occlusions among individuals. As the number of people increases, we can easily expect that the amount of occlusions will also increases.

Taking into account the relatively low image resolution (Fig.6), our counting people performances are good. However, the head tracking in this case is more difficult due to their small size. Hence, with the low-resolution videos, the counting people process based just on the head detection.

4 Conclusions

A method to count the number of pedestrians even in crowded situations is proposed in this paper. We proposed a new people counting based on head detection which can be used to count people in indoor/outdoor areas where there are many moving people. Therefore, this method is useful for surveillance purposes, building management, obtaining marketing data, and other purposes. For the robustness counting, the proposed algorithm can track a head reliably in cases of temporal occlusions by dealing with multiple hypotheses for the pose. Experiments on a crowded scene showed that the number of people passing through the indoor/outdoor area was successfully estimated by using head-based detection. We have conducted a first evaluation for application to global counting people system with the high and low resolution videos. In addition, this yielded very promising counting people performances and it makes the realistic algorithm for multi-camera systems.

References

[1] A.Alahi, L.Jacques, Y.Boursier and P.Vandergheynst: Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In: PETS-Winter, 2009.

[2] A.Albiol and M.J.Silla: Video analysis using corner motion statistics. In: CVPR 2009, pp. 31 – 37, Miami, Florida, 2009.

[3] G.Antonini and J.Thiran: Counting pedestrians in video sequences using trajectory clustering. In: TCSVT, pp. 1008 – 1020, 2006.

[4] A.B.Chan, Z.S.J.Liang and N.Vasconcelos: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: CVPR, pp. 1 – 7, 2008.

[5] D.Conte, P.Foggia and G.Percannella and M.Vento: A method based on the indirect approach for counting people in crowded scenes. In: AVSS, IEEE Conference, pp. 111 – 118, 2010.

[6] D.Navneet and B.Triggs: Histograms of oriented gradients for human detection. In: CVPR, pp. 886 – 893, Washington, 2005.

[7] L.S.Davis, D.Harwood and I.Haritaoglu: A human body part labeling system using silhouettes. In: ICPR, pp. 77 – 78, 1998.

[8] Y.Do: Region based detection of occluded people for the tracking in video image sequences. In: CAIP, pp. 829 – 836, Korea, 2005.

[9] H.Fujiyoshi and A.J.Lipton: Real-time human motion analysis by image skeletonization. In: WACV, Washington, 1998.

[10] W.Ge and R.T.Collins: Evaluation of sampling-based pedestrian detection for crowd counting. In: PETS-Winter, pp; 1 – 7, 2009.

[11] C.P.Lu, G.D.Hager and E.Mjolsness: Fast and globally convergent pose estimation from video images. PAMI, pp. 610 – 622, June 2000.

[12] G.Mori and J.Malik: Estimating human body configurations using shape context matching. In: ECCV, pp. 666 – 680, London, 2002.

[13] N.Thome, D.Merad and S.Miguet: Learning articulated appearance models for tracking humans: A spectral graph matching approach. Image Commun., pp. 769 – 787, November 2008.

[14] K.Nummiaro, E.Koller-Meier and L.J.V.Gool: Object tracking with an adaptive color-based particle filter. In: DAGM, pp. 353 – 360, London, UK, 2002.

[15] P.K.Sharma, C.Huang and R.Nevatia: Evaluation of people tracking, counting, and density estimation in crowded environments. In: CVPR, Miami, Florida, 2009.

[16] P.Viola, M.J.Jones and D.Snow: Detecting pedestrians using patterns of motion and appearance. In: IJCV, pp. 153 – 161, Hingham, July 2005.