Fast People Counting Using Head Detection From Skeleton Graph

Djamel MERAD[1], Kheir-Eddine AZIZ[2]
LSIS - IM
[1,2] UMR CNRS 6168, ESIL-Case 925
163, Avenue of Luminy
13288 Marseille Cedex 9.
[1]Djamal.Merad@univmed.fr
[2]Kheir-Eddine.Aziz@univmed.fr

Nicolas THOME
UPMC - LIP6
Bote courrier 169
Couloir 26-00, Floor 5, Office 526
4 place Jussieu
75252 PARIS cedex 05.
Nicolas.Thome@lip6.fr

Abstract

In this paper, we present a new method for counting people. This method is based on the head detection after a segmentation of the human body by skeleton graph process. The skeleton silhouette is computed and decomposed into a set of segments corresponding to the head , torso and limbs. This structure captures the minimal information about the skeleton shape. No assumption is made about the viewpoint, this is done after the head pose process. Several results present the efficiency of the labelling process , particularly its structural properties for the detection of heads within a crowd. A proposed method has been tested with an experiment of counting the number of pedestrians passing in a specific area.

1. Introduction

People-counting systems have been widely studied in many commercial and public locations, such as theaters, shopping centers, stations, etc. There are many passing persons in these areas so it is important to recognize aspects of their movements. This information can be used to determine the value of a lease, to decide on effective advertising and to display relevant types of merchandise for high sales. In addition, in public places, this information can be applied to arrange safety and subsidiary facilities effectively. For all these reasons, researchers have been concerned with studying methods of counting passing persons.

People counting through image processing is to estimate the number of pedestrians in input images. The information of pedestrians including their number and the positions, from a people counting system based on image processing is expected to reduce the surveillance cost and the observers' fatigue. Pedestrian information can be used in a variety of potential applications. Various methods which estimate the number of people in input images have been previously proposed. They can be divided into three approaches:

1. Visual feature trajectory clustering.
2. Feature-based regression.
3. Individual pedestrian detection.

A trajectory clustering approach. In this approach peoples are counted by tracking and identifying visual features over time. The feature trajectories that exhibit coherent motion are clustered and the number of clusters correspends an estimated number of pedestrians. For example, Antonini et al. [1] proposed a people counting method in which the trajectories obtained by tracking algorithm. The trajectories are clustered based on their lengths and spatial locations. Whereas this approach estimates the number of pedestrians who passed within a specific time, while real-time processing is difficult.

Feature-based regression approach. This approach estimates the number of pedestrians by regressing the features, extracted from an input image, using a regression function. These methods typically work in the three steps, wich are following: 1) background subtraction; 2) extracting various features of the foreground region; and finally 3) estimating the number of pedestrians by a regression function of extracted feature values, e.g. neural networks.

Kong et al. [5] proposed a people counting method in which a neural networks as regression function is used to regress the features, such as edge orientation and blob size histograms obtained by applying background subtraction and edge detection to input image, for estimate the number of pedestrians. Chan and al. [2] proposed a method in which a Gaussian process is used as regression function to regress 28 features extracted from crowd segment obtained by the mixture of dynamic texture.

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.77

However, these methods cannot estimate the positions of pedestrians in the input image and they cannot be executed in real-time.

Antonini et al. [1] proposed a people counting method for estimating the size of inhomogeneous crowds, composed of pedestrians that travel in different directions without using explicit object segmentation or tracking. First, the crowd is segmented into components of homogeneous motion, using the mixture of dynamic textures motion model. Second, a set of simple holistic features is extracted from each segmented region, and the correspondence between features and the number of people per segment is calculated by using a Gaussian Process regression.

Individual pedestrian detection. In this scheme, an algorithms estimates the number of pedestrians that were detected in input images. For example, Viola et al. [13] proposed a people counting method based on boosting appearance and motion features. Zhao et al. [15] proposed a method using a Bayesian model-based segmentation.

However, these methods cannot be applied to very crowded scenes with significant occlusion because they need to detect and segment all pedestrians.

Park et al. [9] proposed a new area-based decision rule that can count people more accurately and can also measure direction of their paths. There are two main ways to approach an area-based decision rule. First, the image is divided into 72 sectors and a size of a person is trained to calculate the mean and variance values for each divided sector. Changes in the size of persons in each sector are estimated by the length variation of the projected person in each of the sectors within the FOV (Field of View). The person's body was approximated as a rectangular form and the lengths of the projected line in each sector were calculated. These values refer to the real lengths of the persons and were also applied to the image because of their proportional relations. Second, various movements of people (which occur in the real world) are analyzed to treat merging and splitting relations among them.

Do.Y. [3] proposed to use simple features of the bounding boxes of target people such as position and size. When two people are in partial occlusion and in the same box, they are segmented into each independent person by analyzing the shape of the binary foreground within the bounding box. Each foreground was identified in independent, partially occluded, or completely occluded state, and the state is updated during tracking.

There are other approaches where they segment the human silhouette before analyzing its shape properties for extract the body parts such as Haritaoglu et al. [4], Mori et al. [8], Thome et al. [12]. Haritaoglu et al. [4]. perform the labeling by first determining the pose among a set of predefined ones.

However, this preprocessing scheme is inevitably prone to fail in some cases, decreasing the overall system performances. The approach proposed by Mori et al. [8] identically retrieves the human pose before performing the labeling. Among a pre-stored set of exemplar 2D views for which key points are manually identified, they first determine for the test shape TS the best match in the shape context meaning, before transferring the key points to TS. Thome et al. [12] used a properly labeling human body parts in video sequences for tracking and motion interpretation. They proposed to perform this task by using Graph Matching. The silhouette skeleton is computed and decomposed into a set of segments corresponding to the different limbs. A Graph capturing the topology of the segments is generated and matched against a 3D model of the human skeleton. The limb identification is carried out for each node of the graph, potentially leading to the absence of correspondence. The method captures the minimal information about the skeleton shape. No assumption about the viewpoint, the human pose, the geometry or the appearance of the limbs is done during the matching process, making the approach applicable to every configuration.

In our the proposed method, we explore the 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 that can count people with a good accuracy. There are four main ways to approach a skeleton-based head detection. 1) the background subtraction; 2) skeleton graph computing; 3) head detection; and 4) estimating the heads pose. The rest of the paper is organized as follows. In Section 2, the proposed method is presented in detail. Experimental results are shown in Section 3. Finally, in Section 4, conclusions and suggestions for further work are presented.

2. People counting method using skeleton graph

The proposed system is illustrated in the Figure 1. The 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 head detection in the skeleton graph.

2.1. Background subtraction

We adopt the Grimson and Stauffer method [11] for foreground segmentation. The authors introduce a method to model each background pixel by a mixture of K Gaussian distributions (K is a small number from 3 to 5). Different Gaussians are assumed to represent different colours. The weight parameters of the mixture represent the time proportions that those colors stay in the scene. The background components are determined by assuming that the

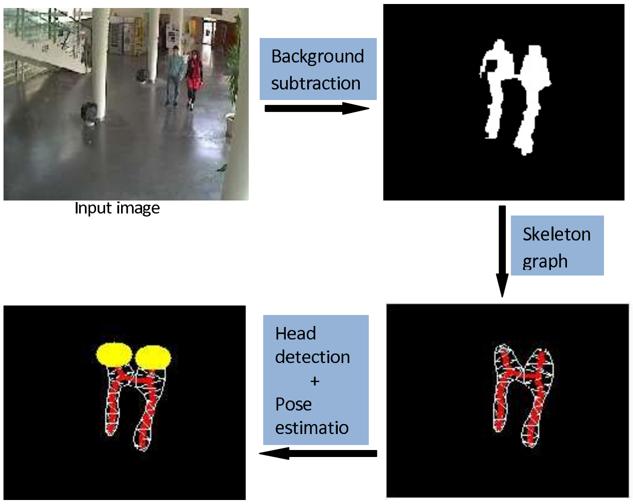

Figure 1. The counting people system.

background contains B highest probable colours.

Every new pixel value is checked against existing model components in order of fitness. The first matched model component will be updated. If it finds no match, a new Gaussian component will be added with the mean at that point and a large covariance matrix and a small value of weighting parameter.

- Adaptive Gaussian Mixture Model

Each pixel in the scene is modelled by a mixture of K Gaussian distributions. The probability that a certain pixel has a value of X_N at time N can be written as $P(X_N) = \sum_{j=1}^{k} W_j * \eta(X_N, \theta_j)$. where k_w is the weight parameter of the k^{th} Gaussian component. $\eta(X_N, \theta_j)$ is the Normal distribution of k^{th} component represented by

$$\begin{aligned}\eta(\mathrm{X},\theta_{\mathrm{k}}) &= \eta(X, \mu_k, \sum\nolimits_k) \\ &= \frac{1}{(2\pi)^{\frac{D}{2}} \, |\sum_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_k)^T \sum_k^{-1} (X-\mu_k)}\end{aligned}$$

where $_{\mathrm{k}}$ is the mean and $\sum_k = \sigma_k^2 I$ is the covariance of the k^{th} component.

The K distributions are ordered based on the fitness value w_k/σ_k and the first B distributions are used as a model of the background of the scene where B is estimated as $B = \arg\min_b(\sum\limits_{j=1}^{b} W_j) > T$.

The threshold T is the minimum fraction of the background model. In other words, it is the minimum prior probability that the background is in the scene. Background subtraction is performed by marking a foreground pixel any pixel that is more than 2.5 standard deviations away from any of the B distributions. The first Gaussian component that matches the test value will be updated by the following update equations,

$$W_{j,t} = (1-\alpha)W_{j,t-1} + \alpha M_{j,t}$$

where $M_{j,t} = \begin{cases} 1\text{;if W is the first match Gaussian component} \\ 0\text{;otherwise} \end{cases}$

$$\begin{aligned}\mu_{j,t} &= (1-\rho)\mu_{j,t-1} + \rho X_{j,t} \\ \sigma_{j,t}^2 &= (1-\rho)\sigma_{j,t-1}^2 + \rho(X_{j,t} - \mu_{j,t-1})^T (X_{j,t} - \mu_{j,t-1})\end{aligned}$$

where W_k is the k^{th} Gaussian component. $1-\alpha$ defines the time constant which determines change. If none of the K distributions match that pixel value, the least probable component is replaced by a distribution with the current value as its mean, an initially high variance, and a low weight parameter.

2.2. Graph skeleton computing

For each detected region (individual/group human), the heads are considered as parts of the skeleton silhouette. 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 related to the skeleton computation corresponds to its sensitivity to noise. To overcome this shortcoming, we first smooth the silhouette. This is achieved by computing the Fourier Descriptor of its outer contours. The Fourier Descriptors provide a discriminative signature of the contour of an object (Zahn et al. [14], Persoon et al. [10]). The $A(k)$ coefficients are determined by the computation of the DFT (Discrete Fourier Transform) for the N contours points considered as complex numbers $X(i)$:

$$A(k) = \frac{1}{N} \sum_{i=0}^{N-1} X(i) e^{-j2\pi ki/N}$$

The coefficient $A(k)$ represent the discrete contour of a shape in the Fourier (frequency) domain. The general shape of the object is represented by the lower frequency descriptors, whereas high frequency coefficients capture details of the object. Thus, smoothing the silhouette is carried out by only keeping a subset of the lowest frequency descriptors.

At this stage, the skeleton is determined by computing the Delaunay triangulation of the smoothed reconstructed silhouette. This approach is the most adapted to our purpose for the following reasons. First, the computation is fast and accurate. Moreover, the Delaunay triangle structure is isomorph to a graph by containing neighborhood information.

The skeleton point sequences is then poligonalized. This step consists in identifying a set of N points $P_i, i \in [1; N]$

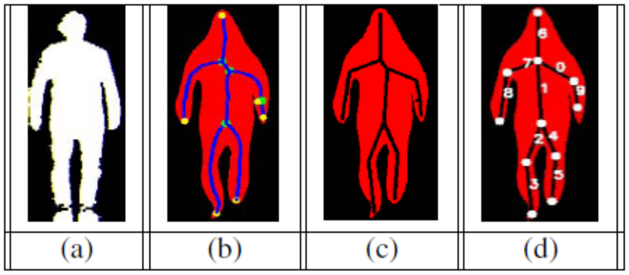

Figure 2. Graph skeleton computing.

and the link between them, representing the $N-1$ segments. We point out that each skeleton point corresponds to the center of the circumscribed circle to each Delaunay triangle. To each link between P_i and P_j $(i \neq j)$ is associated a quantity M_r corresponding to the mean radius of the segment along the skeleton points and computed by the following way :$M_r = \frac{1}{M}\sum_{i=1}^{M} r_i$, where M is the number of skeleton points between P_i and P_j and r_i corresponds to the radius of the i^{th} point. We come back in section 2.3 on the use of M_r for the purpose of the head detection.

2.3. Head detection

The skeleton points may be classified depending on their neighborhood degree. Points having a single neighbor (S) correspond to end points. Points having more than two neighbors (M) define starting points for segments. Points having exactly two neighbors (C) corresponds to points on a continuous curve between (M) and (S) points.

We can notice that (S) and (M) skeleton points must be end points of segments corresponding to body parts. In addition, some (C) points are also likely to belong to the P_i set. Thus, we split each (C) sequence into k segments, so that the mean curvature for the corresponding skeleton points is 0 (in practice under a given small threshold). Figure 2 illustrates the skeleton computation. An extracted silhouette,the skeleton computation (with (S) points in yellow, (M) points in green), the first set of segments after the poligonalization and in the last set after removing small edges are represented in Figure.2.(a, b, c, d) respectively.

At this step, for detecting head we interest only the points' set having a single neighbor (S), we subsequently takes the segment corresponding to extreme node and calculate its degree inclination compared to the vertical axis. If the degree tilt is included between $[-\theta, \theta]$, whereas the segment in question could possibly be considered as head.

2.4. Head pose estimation

In the previous step, the shape of detections can be corrupted by the noise which induces consequently the

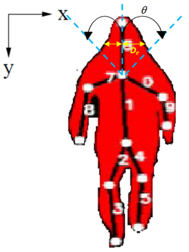

Figure 3. Head detection.

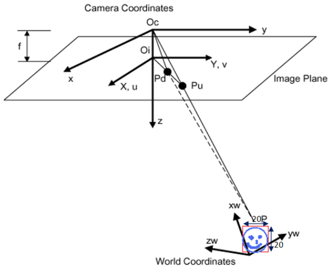

Figure 4. Head pose estimation. In the figure, the world coordinate system is xw, yw, zw; the camera coordinate system is x, y, z. The origin of the camera coordinate system is centered at Oc, and the z-axis coincides with the optical axis. (X, Y) is the image coordinate system at Oi (intersection of the optical axis with the front image plane) and is measured in pixels. (u, v) are the analog coordinates of the object point in the image plane. f is the distance between the front image plane and the optical center.

false detections. To verify the validity for each detection, we must estimate the distance between the local reference model of a head in the world coordinate system $\{x_w, y_w, z_w\}$, which his size is assumed known (20×20), and a reference detection in the camera coordinate system $\{x, y, z\}$ supposedly 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 Figure 4) .

The 3D pose estimation consists in finding the rigid transformation (R, T) minimizing some calculated error (as the sum of error squares) of the one of two collinearity equations (in the image space or in the object space). The two methods generally used to solve this problem are the Gauss-Newton and the Levenberg-Marquardt methods [6]. We used the method of Lu et al. [7] named Orthogonal Iteration (OI) algorithm. Contrary to classical methods, used to solve the optimization problems on the whole, the OI algorithm cleverly exploits the specific structure of the 3D pose estimation problem. To estimate the objects pose, this algorithm uses 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, therefore it is very interesting for real-time applications.

3. Experiments

We applied the proposed method to the experiment of detecting head of people in a scene for finding the number of pedestrians passing an indoor area. We processed video image sequences in the size of 240×320 pixels. One of following three states is assigned to each region detected as foreground. Figure.5 shows the examples of states assigned to blobs.

1. Independent human: if a single human is located within a detected region (blob),
2. Partial Occlusion: if two or more people are together within the seem region but they can be separated by a skeleton-based head detecting method and tracking them by the head tracking algorithm,
3. Complete Occlusion: is the case when two or more people in the seem region are and they cannot be separated. This occurs not only when actually one is behind the other and located vertically on the same linear .

Experiments were done for four video sequences and the results are summarized in figure (resized) 6, 7, 8, 9 and 10.

We observe the robustness of our method for detecting people's heads in most cases and in different situations (independent human, humans partial occlusion, complete occlusion humans), except the case of complete occlusion where people who are on the same vertical lineare which the structural information of the person who's in the face fussioned with another located in behind (fig 9, frame 192). This may be remedied with a tracking process. The main idea is to launch a tracking process on the individual cases of missing his head in fustionned with others blobs. Another interesting point of our method is that the method is able to distinguished the head before, during and after the fusion of the people, which allow first places to robust tracking algorithms and also confirms the effectiveness of structural information incorporated into the skeletons of the interest objects existing in a scene.

4. Conclusions

In this paper, we propose a new people counting method based on head detection which can be used to count people in an area where there are many people moving. Therefore, this method is useful for surveillance purposes, building management, obtaining marketing data, and other purposes. To implement the method, a new head-based detection was applied to extract the head of each person crowded with other persons in the same blob. Further, the head pose estimation was estimated by finding the rigid transformation between the reference system of the model head and the reference system of the camera. This method can be made more robust with an integration of the tracking process.

References

[1] G. Antonini and J.-P. Thiran. Counting pedestrians in video sequences using trajectory clustering. *IEEE Trans. Circuits Syst. Video Techn.*, 16(8):1008–1020, 2006. 1, 2

[2] A. Chan, Z. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. CVPR,. pages 1–7, 2008. 1

[3] Y. Do. Region based detection of occluded people for the tracking in video image sequences,CAIP05. page 829, 2005. 2

[4] I. Haritaoglu, D. Harwood, and L. S. Davis. Ghost: A human body part labeling system using silhouettes. volume 1, page 77, Los Alamitos, CA, USA, 1998. IEEE Computer Society. 2

[5] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *BMVC05*, 2005. 1

[6] D. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:441–450, 1991. 4

[7] C. Lu, G. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *PAMI*, 22(6):610–622, June 2000. 4

[8] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part III*, pages 666–680, London, UK, 2002. Springer-Verlag. 2

[9] H. H. Park, H. G. Lee, S.-I. Noh, and J. Kim. An area-based decision rule for people-counting systems. In *MRCS*, pages 450–457, 2006. 2

[10] E. Persoon and K. S. Fu. Shape discrimination using fourier descriptors. volume 8, pages 388–397, Washington, DC, USA, 1986. IEEE Computer Society. 3

[11] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:747–757, 2000. 2

[12] N. Thome, D. Merad, and S. Miguet. Human body part labeling and tracking using graph matching theory. In *AVSS '06: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, page 38, Washington, DC, USA, 2006. IEEE Computer Society. 2

[13] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IEEE Computer Society*, page 734, 2003. 2

[14] C. T. Zahn and R. Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Trans. Comput.*, 21(3):269–281, 1972. 3

[15] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:459, 2003. 2

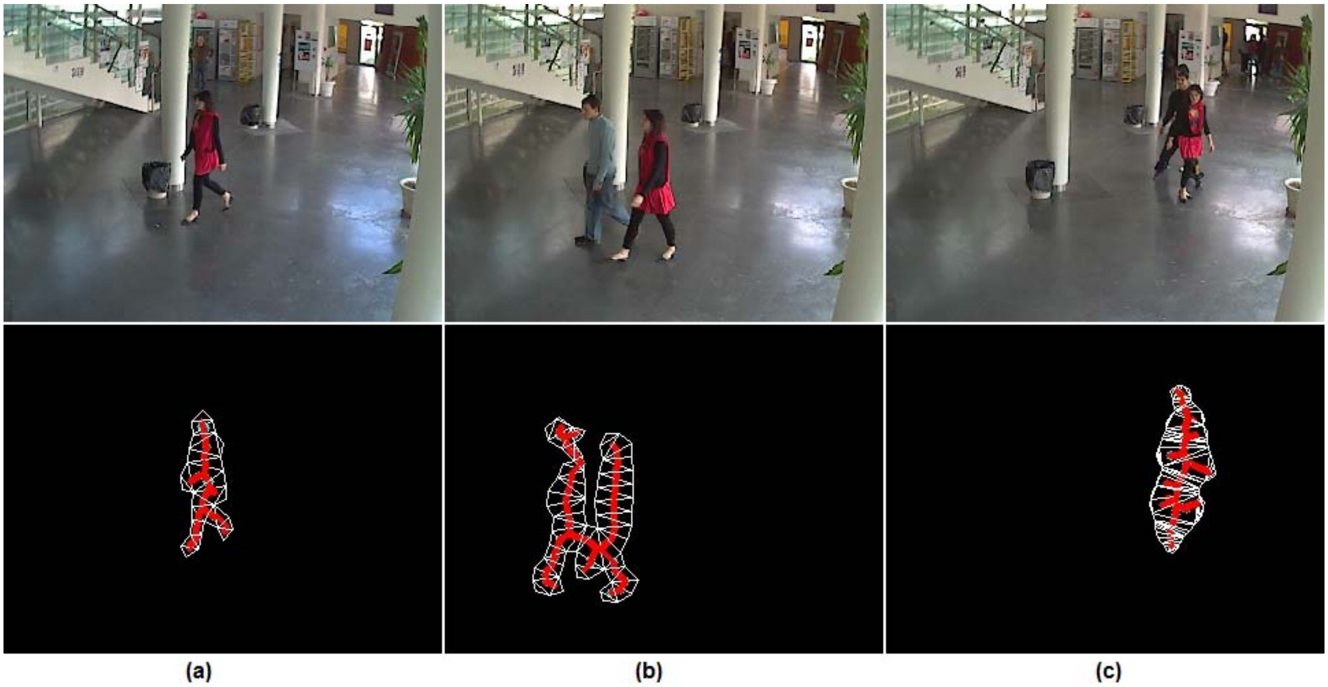

Figure 5. States assigned to blobs: (a) Independent human, (b) Partial Occlusion, (c) Complete Occlusion.

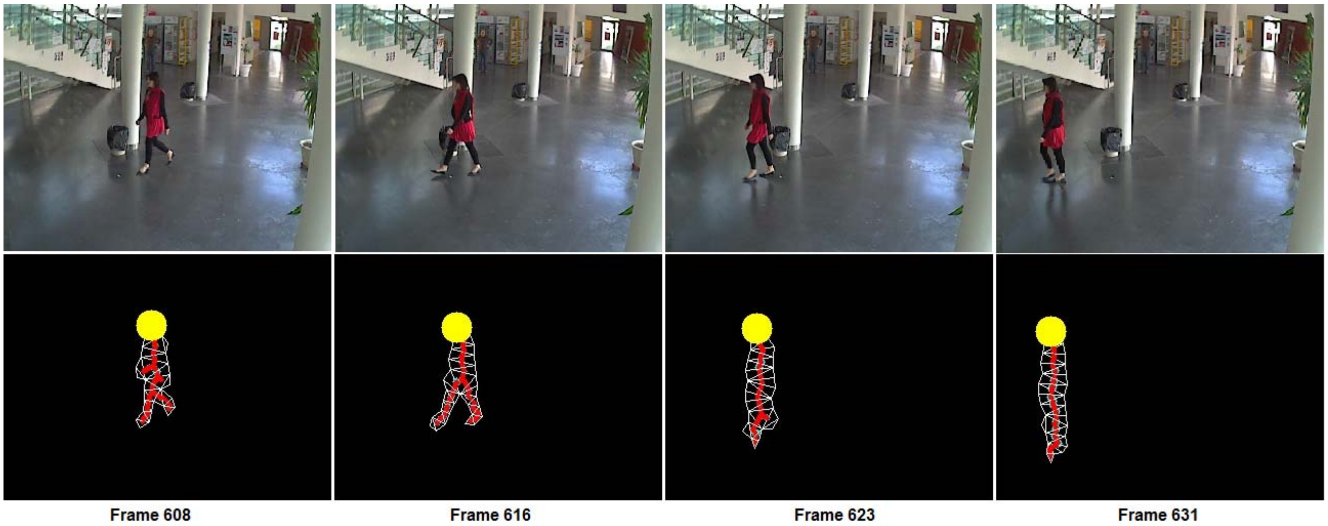

Figure 6. Single human.

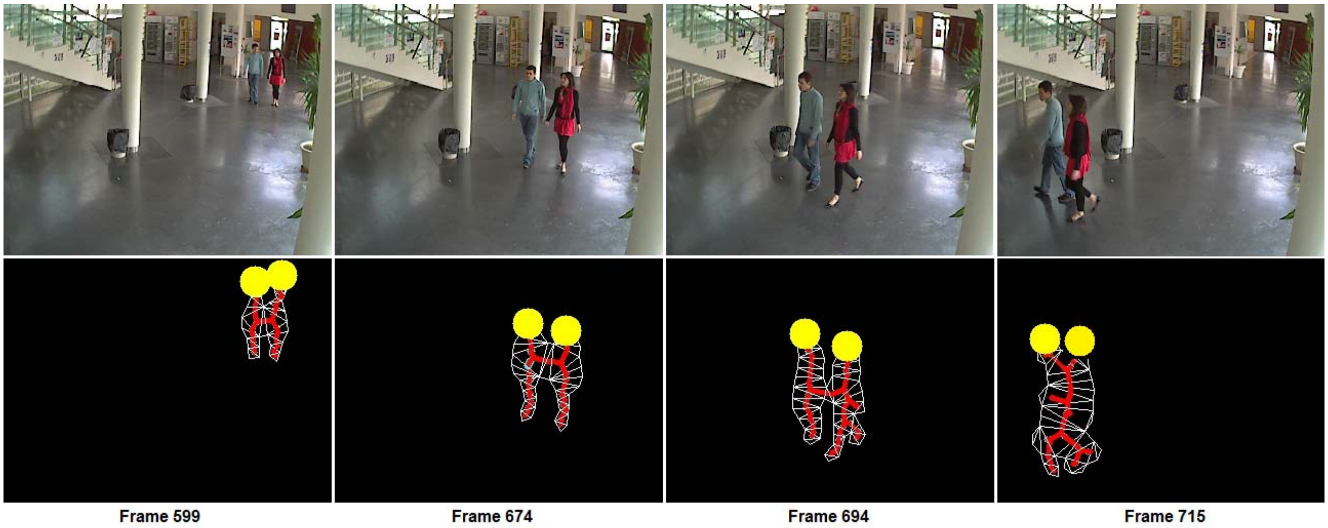

Figure 7. Couple of humans.

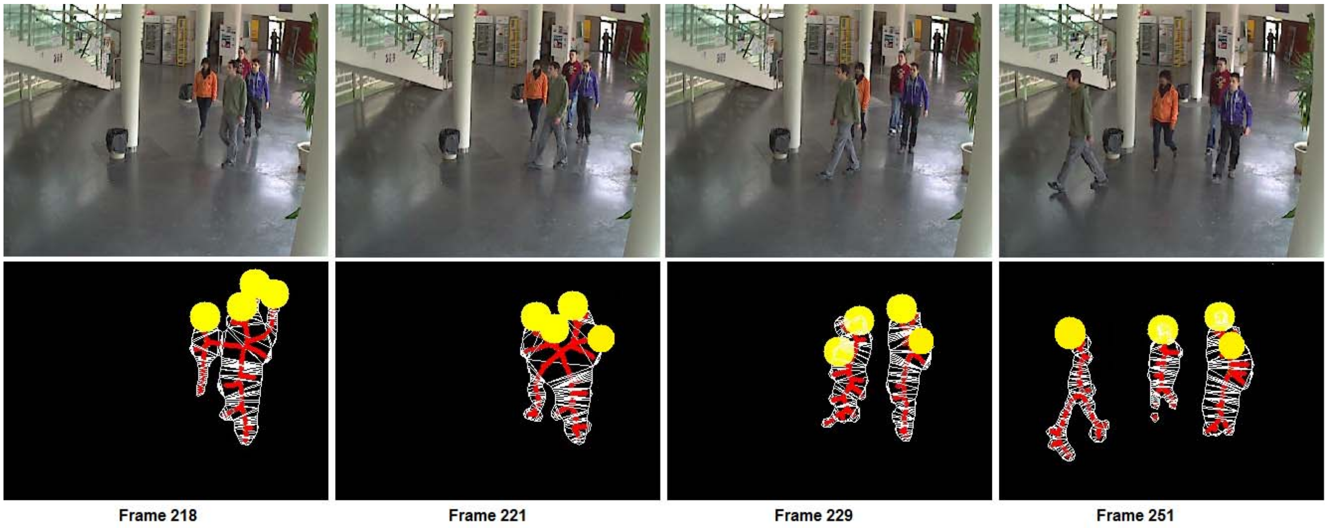

Figure 8. Group of humans.

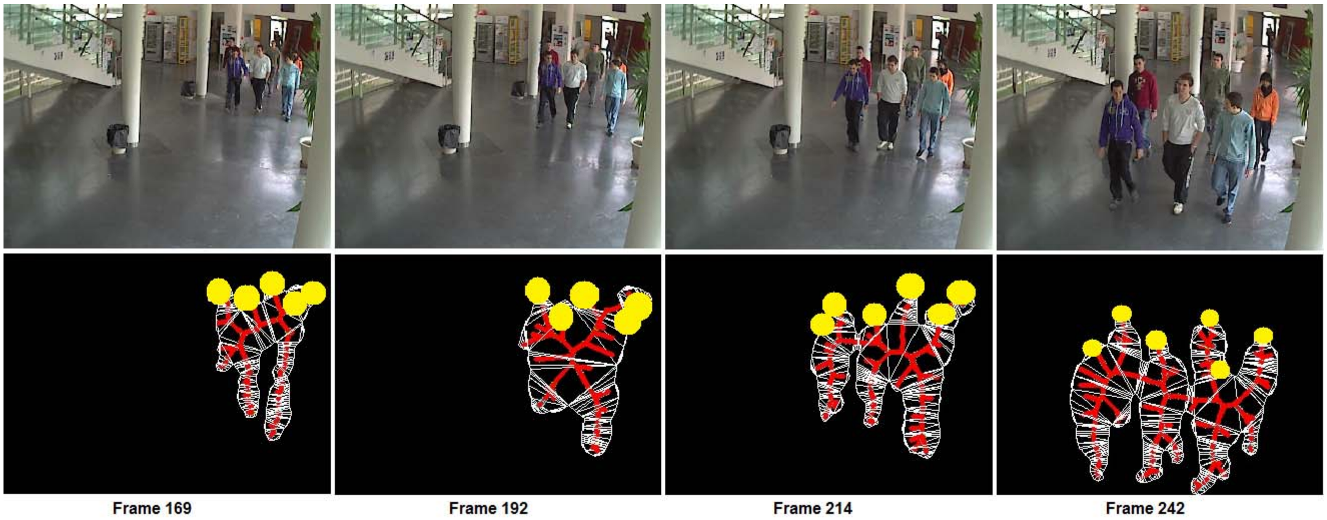

Figure 9. Group of humans.

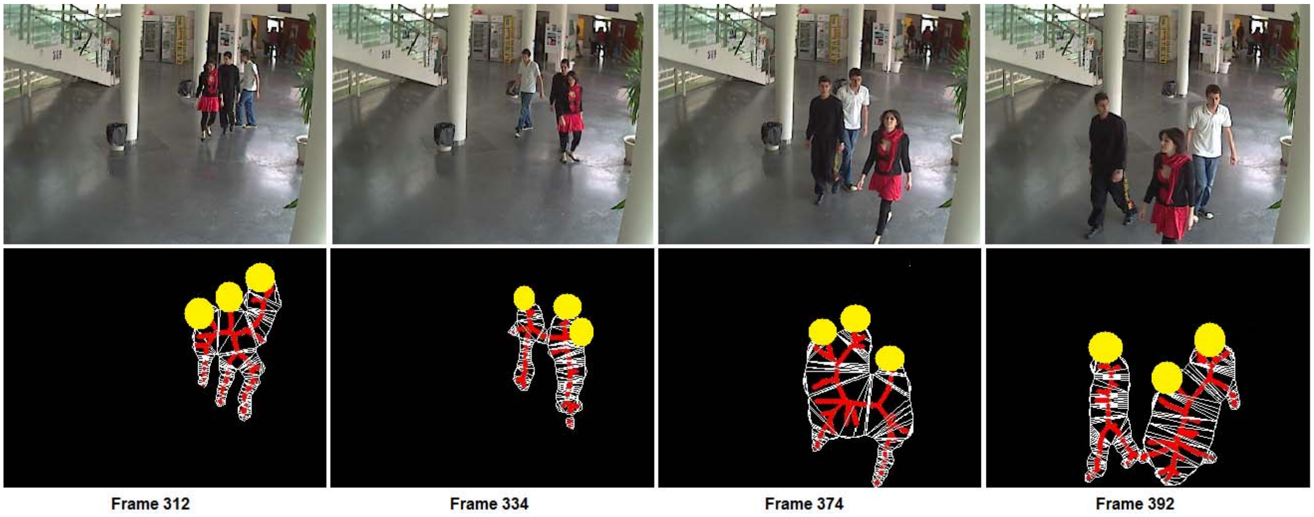

Figure 10. Group of humans in motion with occlusion.