# Tracking People in Crowds by a Part Matching Approach

Zui Zhang
University of Technology, Sydney (UTS)
P.O. Box 123, Broadway
2007 NSW Australia
zuiz@it.uts.edu.au

Hatice Gunes
University of Technology, Sydney (UTS)
P.O. Box 123, Broadway
2007 NSW Australia
haticeg@it.uts.edu.au

Massimo Piccardi
University of Technology, Sydney (UTS)
P.O. Box 123, Broadway
2007 NSW Australia
massimo@it.uts.edu.au

## Abstract

*The major difficulty in human tracking is the problem raised by challenging occlusions where the target person is repeatedly and extensively occluded by either the background or another moving object. These types of occlusions may cause significant changes in the person's shape, appearance or motion, thus making the data association problem extremely difficult to solve. Unlike most of the existing methods for human tracking that handle occlusions by data association of the complete human body, in this paper we propose a method that tracks people under challenging spatial occlusions based on body part tracking. The human model we propose consists of five body parts with six degrees of freedom and each part is represented by a rich set of features. The tracking is solved using a layered data association approach, direct comparison between features (feature layer) and subsequently matching between parts of the same bodies (part layer) lead to a final decision for the global match (global layer). Experimental results have confirmed the effectiveness of the proposed method.*

## 1. Introduction

The interest in video tracking of objects has increased in recent times with various emerging applications such as surveillance, human-centered computing, anthropocentric video analysis, perceptual user interfaces, interactive computer games, ambient intelligence and several others.

Tracking of a moving object implies accurately locating it in each frame of a frame sequence. Tracking multiple targets simultaneously raises a further problem of probabilistic data association [4]: which object is which along the frame sequence? This question typically involves matching single objects in consecutive frames based on coherent models of shape, motion and appearance features. However, the problem becomes progressively more difficult to solve in the presence of increasing target occlusions. Such occlusions mainly occur when the view of the target object is obstructed completely or partially either by elements of the static scene or other moving objects.

While dealing with occlusions explicitly or implicitly has received considerable attention in the tracking literature, tracking humans under challenging occlusions has only recently started attracting the attention of the computer vision and pattern recognition communities. Not only tracking under occlusions must deal with all the problems of usual tracking, but also it has to tackle new problem dimensions such as frequently and severely occluded views and challenging data association. We clarify that the objective herein is not that of tracking the articulated human body, that is impractical at the typical level of resolution of wide-area surveillance videos; rather, that of tracking the location of individual people as they move around the field of view.

In recent years, a few survey papers have covered the topic of occlusion management. In 2002, Gabriel *et al*. presented a review of existing techniques and systems for tracking multiple occluded objects using single or multiple cameras [7]. In 2007, these authors have provided a comparative review of the main recent (2002-2006) tracking methods dealing with significant occlusions [21]. Moeslund et al. presented a survey of advances in vision-based human motion capture and analysis including pixel-based occlusion handling and part-based human tracking [10]. Yilmaz et al. in 2006 have provided a comprehensive survey on video tracking also addressing occlusion handling [20]. In 2002 and 2003, Tao et al. [16] and Zhou et al. [26], respectively, have proposed a method for tracking objects under occlusions by capturing the spatial and the temporal constraints on the shape, motion and appearance of the tracking objects in a dynamic layer representation. Wu and Nevatia in [18] and [19], proposed a human tracking method that takes into account the deformable nature of the human body and the effect of occlusions by modeling a human by body parts. In this paper, we follow a similar rationale by adding further emphasis to the data association problem and its solution. In 2007, Pan and Hu proposed a human tracking algorithm that explicitly models the occluder through a multiple-step

approach [11]. However, the algorithm's extensive use of template correlation operators may make it difficult for a system to meet real-time constraints.

Possibly, the main limitation across the existing tracking literature is that no clear attempts have been made to take into account the nature and statistical distribution of occlusions in the conceptual development stage of the tracking algorithm. While doing so inevitably specializes the application, it also makes expectations for accurate tracking more realistic. In particular, careful consideration must be given to the distribution of occlusions when tackling tracking of humans in very crowded environments (e.g., train stations, airports and shopping malls). Such environments are dominated by typical commuter traffic, with dozens of people walking simultaneously along various flows of directions (to various trains, exits, gates etc). Trajectory patterns tend to be individual i.e. people do not walk in formations and, within a common flow; different individuals exhibit varied speeds and paths. On the single individuals, occlusions tend to be a) repeated and frequent; b) partial, with different parts of the individual occluded at different times, and c) provided by different occluding elements at different times. Because of these conditions, a tracking approach cannot rely on even seldom entirely unoccluded observations.

Given the above framework, the main contribution of this paper is the definition of a part-based people tracking approach suitably mirroring the occlusion distribution of crowded environments. The simple part-based model used in our approach can support effective tracking through such occlusions while at the same time be realistically fitted on typical surveillance videos. This model allows us correct target association even when only a few parts of a target are visible. Moreover, its updating procedure updates the model's parts independently of one another, guaranteeing that the overall model can be kept up to date even in the absence of completely unoccluded views even for sustained periods of time.

The paper is organized as follows. In section 2, we provide an overall structure of the proposed methodology and a detailed description of each individual component. Section 3 presents the experiments carried out and the analysis of the results obtained and Section 4 concludes the paper.

## 2. Methodology

The approach we propose in this paper is a solution to the problem of effectively tracking people in crowds. The approach is based on the adoption of a simplified articulated human model to support the various stages of data association and tracking. An articulated model for a human may range from a minimum of three parts (such as head and shoulders, torso and legs) to anatomical degrees of freedom. Given that we want to retain use of typical wide-area surveillance views, the resolution of each target is of low-medium quality. This prevents us from using a high number of degrees of freedom in the human model. Therefore, we choose to limit the number of parts to a few only. Each part is characterized by a feature set including appearance (e.g., HSV color histogram etc) and spatial features (e.g., centroid of the blob, principal axis of inertia etc). Data association is provided by matching the model's parts to those of possible candidates in the current frame. We choose to restrict the predictive aspects of our approach to quantitities that can be realistically predicted in the crowd scenario. The outline of our tracking algorithm can be summarized as follows:

- *Segmentation*: Obtain object segmentation for the current frame by applying background subtraction with a Gaussian Mixture background model. Generate the foreground image and label all objects present in it.
- *Prediction*: For each currently tracked target, $t_i$, based on its model, calculate the search range for its possible new position in the current frame and select all objects falling within the search range (candidates for a match, $c_{ij}$).
- *Alignment of model on candidates*: For each tracked target, $t_i$, and for each of its candidates, $c_{ij}$, align the model on the candidate and divide this into parts; extract the relevant features for each part. Manage various occlusion cases.
- *Match features*: For each pair, $t_i$-$c_{ij}$, match their features for each part. Infer global matching from parts' matching. The candidate with sufficient, highest likelihood is considered as the current position of the target.
- *Update human model*: Update the human model for the matched and unmatched parts with separate policies.

Each of the above steps is explained in detail in the following subsections.

### 2.1. Segmentation

The initial process for most object tracking algorithms from static camera views consists of background modeling and foreground extraction. In our approach, we use a Gaussian mixture model at pixel level for the background [9]. Foreground regions are extracted by background subtraction, then morphologically closed, smoothed by a median filter, and labeled by using connected components labeling. A size filter is applied to remove noisy small regions and holes inside remaining regions are filled. As the focus of this algorithm is only human tracking, the size filter is also used to remove foreground regions which do not correspond to an acceptable human body size. The result obtained from this initial step is a set of foreground objects or "blobs". Shadow removal is then applied to such objects to filter out foreground shadow pixels and improve the blobs correspondence with the actual visual objects [13].

Various types of errors intrinsically affect the segmentation stage, such as the case of partial segmentation due to occlusions. Such errors are discussed in details in Section 2.4.

## 2.2. The human model

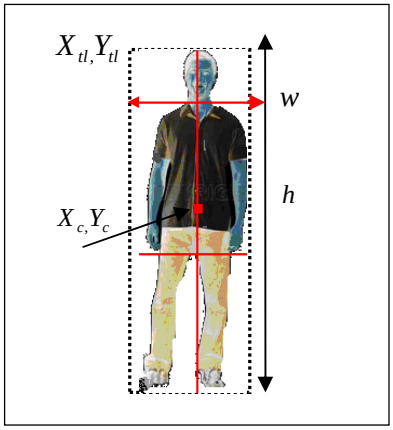


Figure 1. The human model.

The human model adopted in our approach is shown in Figure 1. It contains five parts: head, left and right arms, and left and right legs. The model has an overall rectangular shape and a total of six degrees of freedom in the image plane that are chosen as: the top left point, $X_{tl}$, $Y_{tl}$; the blob's centroid, $X_c$, $Y_c$; and the height, $h$, and width, $w$, of the rectangle. Such degrees of freedom, together with an assumption on the head part's height, subdivide the overall rectangle into five rectangular parts. The head part occupies the top and its height is assumed as a fixed percentage of $h$. The remaining four rectangles represent a body part each. As illustrated in Figure 1, the torso and the two arms are split into two body parts where each part contains one arm and one half of the torso. We decided to construct the model in this way as the torso component provides some desirable stability to the part model by compensating for the instability of the arms. A feature set for each part completes the model.

### 2.2.1 The feature set

There are many features that can potentially contribute to obtain an effective tracking and data association algorithm and the choice amongst them is essentially empirical. In our case, we aim to select features that are relatively invariant to pose and deformation, and also to imprecise fitting of the spatial degrees of freedom of the model. They should also be lightweight to extract and compare for computational reasons. Such requirements tend to exclude appearance templates as their cross-correlation is very sensitive to alignment and computationally heavy. In light of the above, we choose to use the following six features:

- *Area*: The area of a segmented body part is measured by counting the actual number of foreground pixels in the region.
- *Perimeter*: The perimeter of a body part is the length of its contour inclusive of the intersections between rectangular parts.
- *Centroid*: The part's centroid is computed over all its foreground pixels.
- *Principal axis of inertia*: The moment of inertia, also called mass moment of inertia, is the rotational analog of mass i.e. the inertia of a rigid rotating body. This feature helps with testing the rotation angle of the body part.
- *Color histogram*: Appearance is described by the HSV color histogram of the part. We base the measurement of color similarity between two HSV histograms upon the Minkowski metric [14].
- *Amount of pixel overlap*: Despite deformations, we expect the same part of a target to show significant overlap in closely successive frames.

We note that all the above features enjoy limited variance to small pose changes, deformations, imprecision in fitting of the spatial model, and light occlusions. On the other hand, major occlusions will cause significant changes to their values. Therefore, the chosen feature set, $\{f_i\}_{i=1..M}$, $M = 6$, promises a good choice towards correct data association.

## 2.3. Prediction

The first step in the prediction procedure is to find which detected objects are within the range of prediction for a target (gating). Within this search range, a matching procedure will then select the most likely object. The search range is a rectangle area centered at the centroid of the human model, with width and height proportional to the magnitude of the centroid's speed. In other words, the prediction range is based on motion magnitude only and not its direction. In our experiments, the human motion's direction proved not reliably predictable even based on nonparametric models such as particle filters [8, 25]. Unlike other types of targets, a person can suddenly change its motion towards any direction; this is especially the case with the relatively low frame rates of typical surveillance cameras.

Overall, the prediction process can be described by the following steps:

1. Estimate the search range based on the current location of the target and the magnitude of its centroid's speed filtered and predicted using a Kalman filter. The linear/Gaussian assumption of the Kalman filter holds well for these features.

2. For each blob, determine whether it consists of more than one person (under-segmentation) and, if it does, apply the head detection procedure described in Section 2.4 and [22] in order to separate every person included within the blob.
3. Confirm as candidates for matching only those people whose centroids fall within the search range.

For prediction and tracking, Kalman filters have been adopted wherever conditional probabilities appeared to respect Gaussian hypotheses and linear propagation. We use three separate Kalman filters for separate features as we assume they are independent of each other [17]. The first Kalman filter, $KF_1$:

$$KF_1: \ \mathrm{Z}_1=\{X_{c.}, Y_c\}; \mathrm{X}_1 = \{X_c, Y_c, \dot{X}_c, \dot{Y}_c\} \quad (1)$$

is used for estimating the centroid of the blob and its speed, and deriving the search window size as linearly proportional to the speed magnitude. The observation space consists of $X_{c,}Y_c$ and is represented in terms of frame-based coordinates. From empirical observations, the speed of a blob in the state space appears to be relatively stable. The second Kalman filter, $KF_2$, is used for estimating the height, *h*, and width, *w*:

$$KF_2: \ \mathrm{Z}_2=\{h, w\}; \mathrm{X}_2 = \{h, w, \dot{h}, \dot{w}\} \quad (2)$$

When the target person is side-viewed, as the human walking gait is periodic, the correct motion model has to be periodic. A basic view estimator can be used to adjust the motion model. The third Kalman filter, $KF_3$, is used to estimate the displacement between the top-left corner of the bounding box and the centroid.

$$h_c = X_{tl} - X_c \quad (3)$$

$$w_c = Y_{tl} - Y_c \quad (4)$$

$$KF_3: \ \mathrm{Z}_3=\{h_{c.}, w_c\}; \mathrm{X}_3 = \{h_c, w_c, \dot{h}_c, \dot{w}_c\} \quad (5)$$

This estimate is particularly useful to map the human model onto an observation blob in the case of under- or miss-segmentation. When such situations are detected, the centroid is not calculated using the normal approach; rather, assuming that the top-left-corner can still be estimated based on previous observations. In case of occlusion, the observed quantities cannot be reliably observed, thus, the predicted values are confirmed by all three Kalman Filters. Note that the aforementioned three Kalman filters could be straightforwardly combined into a single Kalman filter. However, we chose to separate them due to the fact that the combined matrix would be blob-based (certain features influence only certain state).

## 2.4. Occlusion analysis and management

In our system, the main assumption for handling occlusions is that people may frequently occlude each other; however, the same parts do not stay occluded for long periods of time. More specifically, our methodology does not aim to handle long-term occlusions; instead, we explicitly focus on the occlusions caused by crowds in motion, i.e. short-term, repeated, and on different parts. We categorize the occlusion situation into three cases: normal segmentation, under-segmentation and partial segmentation. The normal segmentation case occurs when the segmented blob's width and height measurements are similar to the prediction. The observation in this case is obtained based on the actual segmentation. In the under-segmentation case multiple physical objects are merged into a single blob. Therefore, the segmented blob's width (height) measurement is significantly larger than the predicted width (height). In the partial segmentation case, the segmented blob's width (height) measurement is significantly smaller than the predicted width (height). We deal with such under- or partial segmentation situations by analyzing the blob using a head detection algorithm. In our tracking approach, the core assumption we make about the scene is that the tops of the heads are visible at all times. Such an assumption is widely utilized in the tracking literature (e.g., [5]). Therefore, we use the head as an *anchor* for aligning our human model onto the candidates for matching. Empirically we found that the use of heads as anchors and some basic assignment rules between models and blobs provide equivalent information to depth ordering.

To this aim, we need a head detection algorithm that is capable of handling significantly variable conditions in terms of equatorial viewpoint (i.e. frontal, profile, back view, from -180 degrees to +180 degrees), tilt angle (i.e. from horizontal to aerial), scale and resolution. We presented one such head detector in detail in [22]. Concisely, we build a model for the head based on appearance distributions and shape constraints. The appearance distribution models the colors of hair and skin by sets of Gaussian mixtures in the XYZ and HSV color spaces. The shape constraint fits an elliptical model to the candidate region and compares its parameters with priors based on the human anatomy. In this work, we further test the shape constraint by use of the Hough transform as the head model may provide miss detections in some frames.

The head detection step provides a list of candidates for matching against the model, typically with multiple, close responses for a single head. We do not attempt to cluster such multiple responses. Instead, we use all of them as possible candidates and let the matching procedure choose the most likely. This raises a fine point about the feature set used in our approach. As stated in Section 2.2.1, the feature set is designed to be limitedly variant to pose, deformation and imprecise fitting of the spatial model onto a candidate. While this design decision supports correct data association even in challenging circumstances, an undesirable side-effect is that it may lead to inaccurate alignment of the model and the candidate. In turn, this causes pollution of the

$\{f_i\}_{i=1..M}$ features values and, in the medium-long term, unacceptable model degradation and data association failure. Therefore, in order to mitigate the impact of inaccurate matches we introduce a *correction step* after the matching procedure has identified the best candidate. The head of the human model is used as a template to search for the best matching head in the search range by template correlation. Given that such templates are very small, the correlation operator carries a negligible computational overhead. The correction is confirmed only if template matching is very good and the distance between the matched head and candidate's head is within a threshold. If the correction is confirmed, the model is rebuilt based on the new head position, and each of the body part features is re-computed.

Examples of challenging occlusions are provided and analyzed in detail in the Experiments section.

## 2.5. Matching

In the proposed system matching is achieved by using a layered data association approach. More specifically, direct comparison between features (feature layer) and subsequently matching between parts of the same bodies (part layer) lead to a final decision for the global match (global layer). Thus, the matching obtained between tracked targets and blobs is referred to as global match whereas the matching obtained between the parts of the same bodies is referred to as local match.

The overall matching process can be described by the following steps:

1. For each unoccluded candidate, divide its blob region into five parts based on the blob's bounding box and centroid. Extract features for each part.
2. For each occluded candidate, apply the human model onto the candidate and extract part features accordingly.
3. Apply part-by-part feature comparison between the human model and the candidate to estimate local matches.
4. Infer the global match from the local matches.
5. Choose the candidate that provides the best possible global match. The global match must also be above a threshold and provide an adequate match ratio with the runner-up.

Local and global matches are further described in the following subsections.

### 2.5.1 Local match

For each body part, a comparison can be obtained by calculating the difference between the features of the potential target with the features of the human model. Each difference is set within a boundary to provide a binary score or decision. The weighted sum of all the decisions provides the final decision for the local match.

Let us note each feature of the part in the model and the observation as $f_{i_m}$ and $f_{i_o}$, respectively. The score $s_{f_i}$ is then computed as follows:

$$if\ lth_{f_i} > d_{f_i} > uth_{f_i}\ then\ s_{f_i} = 1, else\ \ s_{f_i} = 0 \quad (6)$$

$$d_{f_i} = \frac{|f_{i_m} - f_{i_o}|}{f_{i_m}} \quad (7)$$

As shown in our experiments in Section 3, this metric proved sufficient to cope with the changes occurring between feature values in two successive frames, without occlusions. When occlusions occur, qualitatively, there are two situations: i) the occlusion is *minor* and features still match; ii) the occlusion is *major* and features don't match. This is a desirable effect in agreement with the feature and part rationale underpinning our approach. Instead, we decided to leave motion features out of the feature set as they appeared unreliable in this respect. We then progress from feature matching to part matching by using the scores $s_{f_i}$ calculated during feature matching. Each body part, $p_j$, is given a match score as:

$$if\ d_{p_j} > th_{p_j}\ then\ s_{p_j} = 1, else\ \ s_{p_j} = 0 \quad (8)$$

$$d_{p_j} = \sum_{i=1}^{6} w_{f_i} s_{f_i} \quad (9)$$

Weights in (8) have been chosen empirically and follow the constraint that they sum up to 1:

$$\sum_{i=1}^{6} w_{f_i} = 1$$

The head has double the weight of the other body parts due to its role as anchor in our tracking algorithm.

All weights and thresholds have been tuned empirically in our experiments to date. However, learning them automatically from labeled training data is the focus of future development.

### 2.5.2 Global Match

Global match is performed based on the weighted combination of the local matches.

$$k^* = \arg\max_k (D_k) \quad (10)$$

$$D_k = \sum_{l=1}^{5} w_{p_l} s_{p_l} \quad (11)$$

following the constraint that the weights sum up to 1.

A strong local match outputs a higher weight than a weak local match. The final decision for the data association process is made based on the results from the global match of all the potential candidates. In general, the person with the highest global match likelihood is considered as the

current position of the target person. However if multiple potential candidates have similar global match likelihoods or all the global match likelihoods are too weak, then we do not make a decision in the current frame and proceed to the next frame. In the next frame, multiple predictions are made based on the multiple potential candidates in the previous frame and a final decision is made based on the combination of the global match likelihoods from two frames.

### 2.6. Updating

The updating process is crucial for our algorithm to perform as desired. Correct updates of the human model will allow correct prediction and matching, while incorrect updates will corrupt the human model and lead to significant faults in the prediction and matching processes. In general, decision for update is based on both global and local match results. Update at the global level involves update of the geometrical features $\{X_{c.}, Y_c, h, w, h_{c.}, w_c\}$.

The update for body part depends on the body part matching decision. If there is a match, then there is a complete replacement of the spatial features (centroid and position) and partial replacement of the area, perimeter and principal axis of inertia features by using a running average on the values of the model and the observation. The histogram instead is replaced completely or left as before. In the case of *no match* only spatial features are replaced, other features remain as they are. Geometrical features, on the other hand, are always updated to keep the whole body model consistent. These rules apply to all body parts and all cases.

## 3. Experiments

While an increasing number of papers have started addressing the issue of how to perform quantitative comparison of existing algorithms (e.g., [3, 6]), performance evaluation of visual surveillance systems is still an unresolved issue. There is no commonly agreed performance evaluation criteria (i.e., how to perform objective/ comprehensive/ comparative evaluation, how to represent the complexity and range of issues handled, etc.) for tracking in crowds. We thus carried out two experiments in order to evaluate the performance and stability of the proposed approach under various occlusion conditions qualitatively. We tested our tracking algorithm on sequences from the CAVIAR dataset [2] and the AVSS 2007 dataset [1].

From the CAVIAR dataset we used a sub-sequence from the video named "WalkByShop1cor.mpg", where a couple is walking along a corridor browsing, and there are persons going inside and coming out of stores. There are no illumination changes; however there is occurrence of occlusions between the target and the rest of the people in FOV.

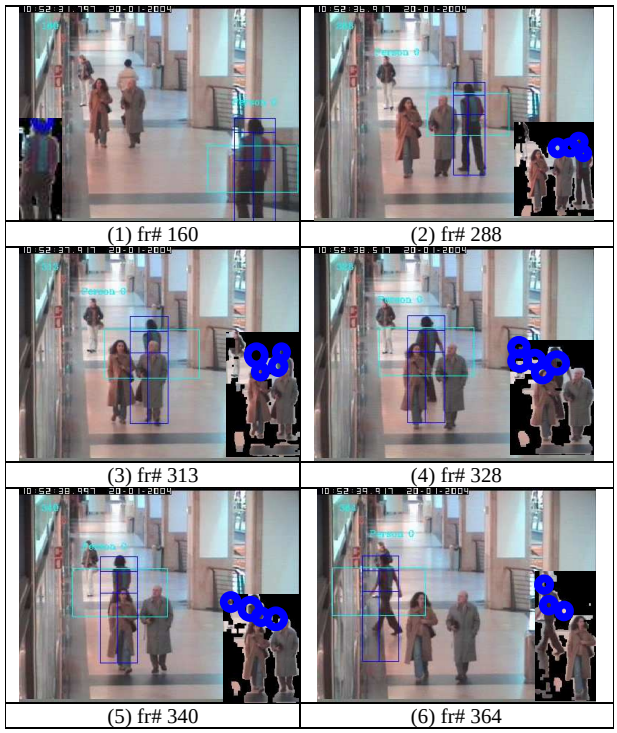


Figure 2. The proposed algorithm handling a challenging long-term occlusion from the CAVIAR dataset.

For our experiment we only used around 368 frames with 98 frames of background. The length of the occlusion is around 80 frames. Representative frames are shown in Figure 2. Please note that as we chose to use a subset, the frame numbers displayed in the figure do not correspond to the actual frame numbers in the dataset. The rectangle around the object corresponds to the target and its parts, while the bright green rectangle represents the search range. The segmented blob and the head candidates are displayed from left in the lower right corner of each image. The first image shows the target person in full view with the body model by parts fitted onto the segmented blob. The second image displays the case where the spatial fitting is as desired while the target encounters initial stages of occlusion. In the third image the target faces significant occlusion and the part matching returns low likelihood for the occluded body parts. The fourth image shows how the speed of the model adaptation is necessarily a tradeoff between stability and responsiveness. Therefore, the model gradually incorporates occlusions by obtaining a good match for the left and right legs despite being occluded, as well as the unoccluded body parts of the object, namely, head, left and right arms. The fifth image displays how the model eventually survives occlusion by another person, and the

sixth image reflects how the model of the person is retrieved after occlusions pass.

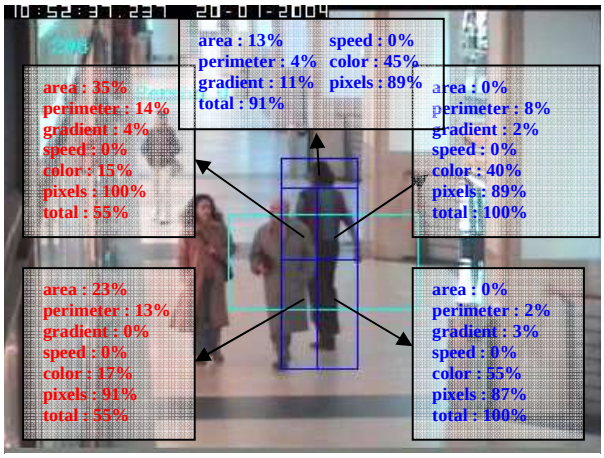


Figure 3. Illustration of the data association procedure at the feature layer. Features $f_1$ to $f_6$ are area, perimeter, gradient, speed, color histogram, and overlapping pixels, respectively, and $\{d_{p_j}\}_{j=1..6}$ is the final match for the body part represented as total.

In addition to Figure 2, Figure 3 illustrates the data association procedure at the feature layer for the CAVIAR dataset. The comparison results for the features area, perimeter, gradient, and speed in the ideal case should be as low as possible. However, the comparison result for color histogram similarity and overlapping pixels between the candidate and the model, in the ideal case, should be as high as possible. In the example demonstrated in Figure 3, the head, right arm and right leg are considered as *a match*, whereas the left arm and the left leg are considered as *not-a-match* due to occlusion. Consequently, features that are not matched are not updated in the human model. Please note that these comparisons are obtained between the model and the observation for the same person (i.e. target and candidate are the same person). This is clearly demonstrated with the speed feature where the comparison always returns *a match*. If the candidate is a different person, then the feature comparison will vary significantly.

We also used a sub-set of the AVSS 2007 dataset for the i-Lids bag challenge. This is a dataset for event detection in CCTV footage and the event of interest is abandoned baggage. For our experiment we only used 489 frames with 125 frames of background from a sub-sequence of the video labeled as "AVSS AB Easy_Divx.avi". The length of occlusions within this sub-set is around 20 frames. Representative frames are shown in Figure 4 (Please see [1] for further details on the dataset). In this sequence the target person is occluding another person in the FOV. Again, the rectangle around the object corresponds to the target and its parts, while the bright green rectangle represents the search range. The segmented blob and the head candidates are displayed from left in the lower left corner of each image. The first image shows the target person in full view with the body model by parts fitted onto the segmented blob. The second and third images display the cases where the target person is occluding another person in the FOV. The fourth image displays how the target survives the occlusion situation.

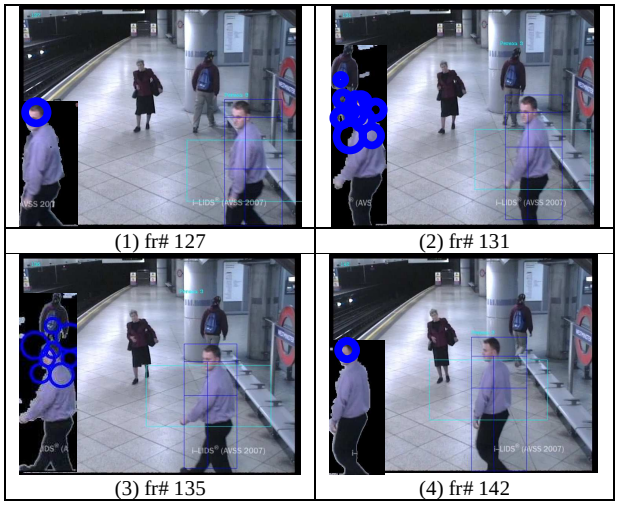


Figure 4. The proposed algorithm handling a case with occlusion from the AVSS 2007 dataset.

It is clear from Figures 2-4 that our method can successfully handle repeated, partial and/or challenging long-term occlusions.

## 4. Conclusions

In this paper, we presented a method for tracking of humans in crowded environments with occlusions that tend to be, a) repeated and frequent; b) partial, with different parts of the individual occluded at different times, and c) provided by different occluding elements at different times. Under these assumptions, we defined a part-based human tracking approach suitably supporting the various stages of data association and tracking by mirroring the occlusion distribution of crowded environments. We firstly adopted a simplified articulated human model consisting of five body parts with six degrees of freedom, and each part represented by a rich set of features. We then used a layered data association approach, where direct comparison between features (feature layer) and subsequently matching between parts of the same bodies (part layer) led to a final decision for the global match (global layer).

Overall, we demonstrated with experiments that the simple part-based model used supports effective tracking through repeated, partial and/or challenging long-term

occlusion typically encountered in surveillance videos. This model allows us correct target association even when only a few parts of a target are visible. Moreover, its updating procedure updates the model's parts independently of one another, guaranteeing that the overall model can be kept up to date even in the absence of completely unoccluded views even for sustained periods of time. However, this algorithm is expected to break down wherever medium-size occlusions affect most parts (e.g., 4 out of 5) of the model for a sustained period of time. Incorrect data association might then occur if the features of the occluding object are similar to those of the tracked target.

As future work we intend to extend the system presented in this paper in various ways. Firstly, we plan to change the semi-automated initialization to be a fully-automated initialization procedure. Secondly, the tracking of a single target can be extended to tracking multiple targets simultaneously. The method presented only handles challenging spatial occlusions; it can be extended to handle challenging temporal occlusions where the target person will be occluded by a single object extensively for a long period of time. Additionally, learning the weights and the thresholds automatically from labeled training data will also be explored.

## 5. Acknowledgements

This research is supported by the Australian Research Council and iOmniscient Pty Ltd under the ARC Linkage Project Grant Scheme 2006 - LP0668325.

## 6. References

[1] AVSS 2007 data set: http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html

[2] CAVIAR 2003 data set: http://homepages.inf.ed.ac.uk/ rbf/ CAVIARDATA1/

[3] P. Correia and F. Pereira, Video object relevance metrics for overall segmentation quality evaluation, EURASIP Journal on Applied Signal Processing, 2006(15):1-11, 2006.

[4] I. J. Cox, Review of statistical data association techniques for motion correspondence. Int. Journal on Computer Vision, 10(1): 53-66, 1993.

[5] A. M. Elgammal and L. S. Davis, Probabilistic framework for segmenting people under occlusion, Proc. of the Int. Conf. on Computer Vision, 2:145-152, 2001.

[6] C.E. Erdem, B. Sankur and A.M. Tekalp, Performance measures for video object segmentation and tracking, IEEE Tran. on Image Processing, 13(7): 937-951, 2004.

[7] P. F. Gabriel, J. G. Verly, J. H. Piater, A. Genon, The state of the art in multiple object tracking under occlusion in video sequences, Proc. of ACIVS, 2003.

[8] M. Isard and A. Blake, Condensation - Conditional density propagation for video tracking, Int. Journal of Computer Vision, 29(1):5-28, 1998.

[9] P. KaewTraKulPong and R. Bowden, An improved adaptive background mixture model for real-time tracking and shadow detection, Proc. 2nd European Workshop on Advanced Video-Based Surveillance Systems, 1-5, 2001.

[10] T.B Moeslund, A. Hilton and V. Krüger. A survey of advances in vision-based human motion capture and analysis. Int. Journal of Computer Vision and Image Understanding, 104(2-3): 90-126, 2006

[11] J. Pan and B. Hu, Robust occlusion handling in object tracking, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 1– 8, 2007.

[12] S. Park and J.K. Aggarwal, Simultaneous tracking of multiple body parts of interacting persons, Int. Journal on Computer Vision and Image Understanding, 102(1):1-21, 2006

[13] A. Prati and et al., Detecting moving shadows: Algorithms and evaluation," IEEE Trans. Pattern Analysis and Machine Intelligence, 25(7): 918-923, 2003.

[14] D. Roth and et al., Event-based tracking evaluation metric, IEEE Workshop on Motion and Video Computing, Copper Mountain, 2008

[15] A. Senior and et al., Appearance models for occlusion handling, Image and Vision Computing, 24(11): 1233-1243, 2006

[16] H. Tao, H. S. Sawhney and R. Kumar, Object tracking with Bayesian estimation of dynamic layer representations, IEEE Tran. on Pattern Analysis and Machine Intelligence, 24(1): 75-89, 2002.

[17] G. Welch and G. Bishop, An introduction to the kalman filter, Tech. Report (TR 95-041), University of North Carolina at Chapel Hill, Department of Computer Science.

[18] B. Wu and R. Nevatia, Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. Proc. of IEEE Int. Conf. on Computer Vision, 1: 90-97, 2005.

[19] B. Wu and R. Nevatia, Tracking of multiple, partially occluded humans based on static body part detection, Proc. of Computer Vision and Pattern Recognition, 1:951-958, 2006.

[20] A. Yilmaz, O. Javed and M. Shah, Object tracking: A survey, ACM Computing Surveys, 38(4):13.1-13.45, 2006.

[21] Z. Zhang and M. Piccardi, A review of tracking methods under occlusions, Proc. of the IAPR Conf. on Machine Vision Applications, 146-149, 2007.

[22] Z. Zhang, H. Gunes and M. Piccardi, An accurate algorithm for head detection based on XYZ and HSV hair and skin color models, Proc. of the IEEE Int. Conf. on Image Processing, 2008 (accepted and in press).

[23] T. Zhao and R. Nevatia, Bayesian human segmentation in crowded situations. Proc. of Computer Vision and Pattern Recognition, 2: 459-466, 2003.

[24] T. Zhao and R. Nevatia, Tracking multiple humans in complex situations. IEEE Tran. on Pattern Analysis and Machine Intelligence, 26(9): 1208-1221, 2004.

[25] S. K. Zhou and et al., Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Tran. on Image Processing, 13(11):1491-1506, 2004.

[26] Y. Zhou and H. Tao, A background layer model for object tracking through occlusion. Proc. of the Ninth IEEE Int. Conf. on Computer Vision, 2: 1079-1085, 2003.