Universities of Leeds, Sheffield and York
http://eprints.whiterose.ac.uk/

Detecting Carried Objects
from Sequences of Walking Pedestrians

Dima Damen, *Member, IEEE,*
Department of Computer Science, University of Bristol, UK.
David Hogg, *Member, IEEE*
School of Computing, University of Leeds, UK.

Abstract—This paper proposes a method for detecting objects carried by pedestrians, such as backpacks and suitcases, from video sequences. In common with earlier work [14], [16] on the same problem, the method produces a representation of motion and shape (known as a temporal template) that has some immunity to noise in foreground segmentations and phase of the walking cycle. Our key novelty is for carried objects to be revealed by comparing the temporal templates against view-specific exemplars generated offline for unencumbered pedestrians. A likelihood map of protrusions, obtained from this match, is combined in a Markov random field for spatial continuity, from which we obtain a segmentation of carried objects using the MAP solution. We also compare the previously used method of periodicity analysis to distinguish carried objects from other protrusions with using prior probabilities for carried-object locations relative to the silhouette. We have re-implemented the earlier state of the art method [14] and demonstrate a substantial improvement in performance for the new method on the PETS2006 dataset. The carried-object detector is also tested on another outdoor dataset. Although developed for a specific problem, the method could be applied to the detection of irregularities in appearance for other categories of object that move in a periodic fashion.

Index Terms—baggage detection, carried-objects detection, silhouette analysis, temporal templates, template matching, periodicity analysis.

❖

1 INTRODUCTION

The detection of carried objects is a potentially important objective for many security applications of computer vision. However, the task is inherently difficult due to the wide range of objects that can be carried by a person, and the different ways in which they can be carried. This makes it hard to build a detector for carried objects based on their appearance in isolation or jointly with the carrying individual. An alternative approach is to look for irregularities in the silhouette of a person, suggesting they could be carrying something. This is the approach that the method presented in this paper adopts, and whilst there are other factors that may give rise to irregularities, such as clothing and build, experiments on two datasets are promising.

The detector assumes a static background and starts by averaging aligned foreground segmentations of a walking pedestrian to produce a representation of motion and shape (known as a *temporal template*) that has some immunity to noise in foreground segmentations and phase of the walking cycle. This representation, introduced in [8], was used in [14], [16] for the same application. A ground-plane homography is used to estimate the walking direction of the person relative to the camera's viewing direction. The temporal template is then matched against a pre-compiled exemplar temporal template of an unencumbered pedestrian viewed from the same direction. Protrusions from the exemplar are detected as candidate pixels for carried objects. Finally,
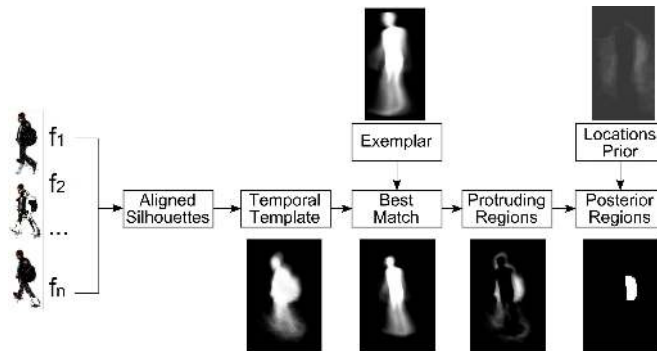

Fig. 1. Foreground blobs (obtained from background subtraction) from one trajectory, each centred around the median, are first aligned. The temporal template represents the frequency of each aligned pixel being part of the foreground. The exemplar temporal template from a similar viewing angle is transformed (translation, scaling and rotation) to best match the generated temporal template. By comparing the temporal template to the best match, protruding regions are identified. A Markov random field (MRF) with a trained spatial prior is used to segment carried objects.

prior information about the expected spatial position of carried objects relative to the body is incorporated, together with a spatial continuity assumption in order to improve the segmentation of pixels representing the carried objects. Figure 1 summarises, along with an

example, the process of detecting carried objects. It is worth mentioning that the detector does not require re-training when applied to a different scene, as will be shown in the experiments section.

Section 2 reviews previous work on the detection of carried objects. Section 3 presents our method, based on matching temporal templates. It studies the pros and cons of using periodicity analysis to classify protrusions, and then incorporates spatial priors and a continuity assumption for segmenting carried objects. Experiments comparing the performance with the earlier work from Haritaoglu *et al.* [16] on the PETS2006 dataset [11] are presented in Section 4. It shows an improved performance when using template matching and a spatial prior. The approach was applied, without any re-training, to a dataset recorded on the campus of the University of Leeds over a full day (LEEDS2009). Results show the ability to detect carried objects from longer video sequences where the viewpoint changes along the trajectory. Section 5 shows how this detector can be used to relate people entering and exiting a building by relating their carried objects, which is one of the applications of this detector.

2 PREVIOUS WORK

Several previous methods have been proposed for detecting whether an individual is carrying an object. The *Backpack* [14], [16] system detects the presence of carried objects from short video sequences of pedestrians (typically lasting a few seconds) by assuming the pedestrian's silhouette is symmetric, and that people exhibit periodic motion while moving unencumbered. Binary foreground segmentations are aligned using edge correlation. The aligned foreground segmentations are averaged to create the temporal template that records the proportion of frames in the video sequence in which each pixel was segmented within the foreground. Next, symmetry analysis is performed. The principal axis is computed using principal component analysis of two-dimensional locations, to find the dominant direction and the median. For each location, asymmetry is detected by reflecting the point in the principal axis (Figure 2). The proportion of frames in which each location was asymmetric is calculated. Temporally-consistent asymmetric locations are grouped into connected components representing candidate blobs.

Backpack then distinguishes between blobs representing carried objects and those being parts of limbs by analysing the periodicity of the horizontal projection histograms. The periodicity analysis calculates the periodic frequency of the full body, as well as the periodicity of each horizontal slice (perpendicular to the principal axis) containing an asymmetric connected component. *Backpack* assumes the frequency of a horizontal slice that contains a limb is numerically comparable to that of the full body. Otherwise, it is believed to contain a carried object. Figure 3 reviews the process using a

Fig. 2. For each foreground segmentation, the principal axis is found and is constrained to pass through the median coordinate of the foreground segmentation. Light grey represents the asymmetric regions.

re-implementation of *B*ackpack [14], [16]. From the re-implementation, errors in the *Backpack* method arise from two main sources. Firstly, the position of the principal axis is often displaced by the presence of the carried object. This might be slightly improved by forcing the principal axis to pass through the centroid of the head [15] or the ground position of the person's feet [17]. Secondly, accurate periodicity analysis requires a sufficient number of walking cycles to successfully retrieve the frequency of the gait. In Section 3.1 we review a method for periodicity that gives better results than computing over a horizontal slice.

Later work by Benabdelkader and Davis [1] expanded the work of Haritaoglu *et al.* by dividing the temporal template perpendicular to the principal axis into three equal-size slices. The periodicity and amplitude of the time series along each slice are studied to detect deviations from the 'natural' walking person and locate the vertical position of the carried object. They concluded that the main limitation in Haritaoglu *et al.*'s method is the sensitivity of the axis of symmetry to noise, as well as to the location and size of the carried object(s).

The work of Lee and Elgammal also uses silhouettes for detecting carried objects on a per-frame basis [18].
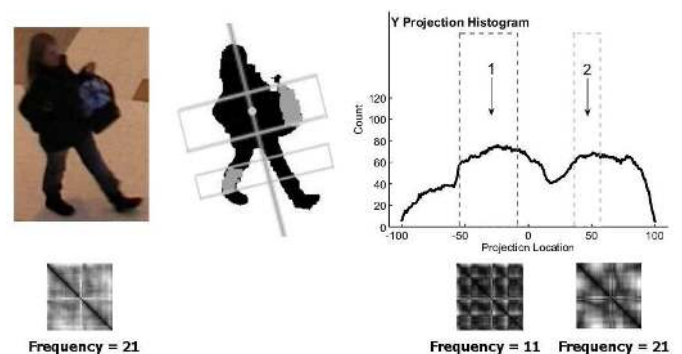

Fig. 3. Light grey represents the two asymmetric connected components, and the corresponding horizontal slices. The slices are projected onto the horizontal projection histogram (right). The periodicity is computed for the full histogram [freq = 21] and for slices 1 [freq = 11] and 2 [freq = 21]. As slice 2 has the same frequency as the full body, it is not considered a carried object.

The training process finds a low-dimensional representation of the kinematic manifold given the joint angles in three dimensional space over time. For the silhouette at each frame, the best match of the foreground shape with training shapes is found. An iterative process fills the holes in the foreground segmentation to find better matches, as matching relies on aligning the centres of gravity. Carried objects are then defined as the unmatched pixels in the foreground region. The approach was only qualitatively analysed. This work is similar to the method proposed in this paper as pixels outside an aligned shape are considered carried objects, yet detection is performed on single frames which is sensitive to noise and errors in foreground segmentation.

Branca *et al.* [4] try to identify intruders in archaeological sites. Intruders are defined as those carrying objects such as a probe or a tin. They assume a person is detected and segmented. Their approach thus tries to detect such carried objects within the segmented foreground region. Detection is based on wavelet decomposition, and the classification uses a supervised three-layer neural network, trained on examples of probes and tins in foreground segmentations.

Differentiating people carrying objects, without locating the carried object, has also been studied. One example is the work by Nanda *et al.* [21]. Supervised learning was accomplished based on examples of unencumbered pedestrians and outliers. Outliers are "unusual-looking pedestrians... caused by wearing a hat or carrying an object". A three-layer neural network was trained for classification. This work's performance depends on the presence of a similar object within the same viewpoint in the training data. Alternatively, the work of Tao *et al.* [25] tries to detect pedestrians carrying heavy objects by performing gait analysis. The task was performed using general tensor discriminant analysis, and was tested on the USF HumanID gait analysis dataset.

The work of Ghanem and Davis [13] detects abandoned baggage by comparing the person before approaching a region of interest and after leaving it. Carried objects are detected by comparing the temporal templates (the term 'occupancy map' is used in their work to reference the same concept) and colour histograms of the 'before' and 'after' sequences. The approach assumes the person is detected twice, and that the trajectory of the person before approaching the region of interest and after departing are always correctly connected. It also assumes all observed individuals follow the same path, and thus uses two static cameras to record similar viewpoints.

Similarly, Chuang *et al.* assume the person is seen with and without the carried object [5]. The ratio of the colour histograms between consecutive frames is used to detect the change in colour components and thus the removal of an object. By observing people approaching one another, the work detects the exchange of carried baggage indicating suspicious events like thefts. The assumption of observing the person before and after

the change in carrying status is application-specific and cannot be used as a general carried-object detector.

The novel method in this paper (Section 3) also uses the temporal template but differs from earlier work [13], [14] in matching the generated temporal template against an exemplar temporal template generated offline from a 3D model of a walking person. Several exemplars, corresponding to different views of a walking person, are generated from binary silhouettes. The temporal templates provide better immunity to noise in foreground segmentations, and enable matching for video sequences as opposed to individual frames. The approach does not require the pedestrian to be detected with and without the carried object, and can handle different viewpoints. It detects any type of carried object (not merely bags and backpacks), and can be considered a general approach for detecting protrusions from deformable tracked objects for which there is a 3D model.

3 THE METHOD

Our method starts by creating the temporal template from a sequence of tracked pedestrians as proposed by Haritaoglu *et al.* [16]. Background subtraction tracker based on a Kalman filter is used to retrieve trajectories of moving connected components. For each trajectory, foreground segmentations are extracted for each frame by cropping around the trajectory's centre point using a window of a fixed size $h \times w$. The foreground segmentations at each frame are often noisy due to shadows and camouflage. The temporal template is created by aligning and then averaging these foreground segmentations. Figure 4 shows a set of foreground segmentations for a single trajectory and their corresponding temporal template. To align the segmentations, Haritaoglu *et al.* used

Fig. 4. Foreground segmentations along with the created temporal template.

an edge correlation over a 5×3 search window. To avoid a predefined displacement window, we use Iterative Closest Point (ICP) [26] to align successive boundaries. Experimentally, it gives a more accurate alignment in the presence of shape variations between consecutive frames (Figure 5). While the method in [16] averages all aligned silhouettes, we introduce an additional step to further decrease the noise in the temporal templates.

All the frames are ranked in ascending order by their L_1 difference from the generated temporal template. The top $p\%$ of the frames are used to re-calculate a more stable template. p is set to 80 in the results shown below. The more expensive Least Median of Squares (LMedS) estimator [23] gave similar results. Section 4 compares the LMedS estimator to the L_1 ranking method.

Fig. 5. Edge correlation temporal template over 15×15 (left) and 30×30 (middle) search windows. ICP temporal template (right) does not require any parameters.

Fig. 7. Eight exemplar temporal templates, created to represent eight viewpoints. The model number corresponds to the camera number in the EPFL dataset (Figure 6)

In aligning foreground segmentations to create the temporal template, the method assumes that the people do not change their walking direction, otherwise the averaged temporal template would be ill-formed. Therefore, each trajectory is partitioned into clips of 2 seconds (the last clip will typically be shorter). The temporal template is created separately for each clip. The trajectories are partitioned uniformly regardless of whether the viewpoint, the direction of motion or the scale have changed. An alternative that we did not explore would be to use a sliding window.

Having derived a temporal template from a tracked clip, it is compared to an exemplar temporal template viewed from a similar direction to identify protrusions by matching. The exemplar temporal templates are derived from a 3D model of an unencumbered walking person. We have used a dataset, gathered at the Swiss Federal Institute of Technology (EPFL) [9]. This dataset is collected from 8 people walking at different speeds (3 to 7 km/h) on a treadmill. Their motion was mapped onto a 3D Maya model and the dataset consists of binary silhouettes from eight simulated views (Figure 6). The EPFL dataset of silhouettes has previously been used for pose detection, 3D reconstruction and gait recognition [9], [12].

The exemplar temporal templates are created by aligning and averaging the silhouettes as explained above for each camera view. We have used 8 directions to correspond to the 8 cameras available from the EPFL dataset (Figure 7). This seems sufficient to cover varying viewpoints, but we have not tested using a different number of models. Nothing restricts our method from extending to more exemplars covering finer sampling of the viewing hemisphere (including different elevations). We did not have access to the Maya model from the EPFL dataset [9], but to silhouettes from 8 directions
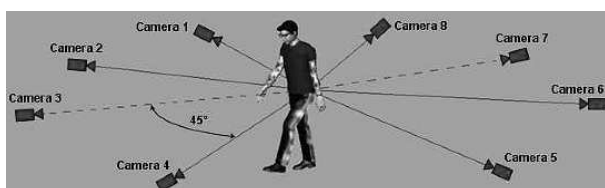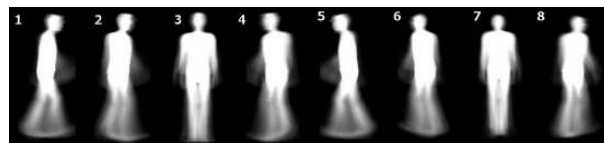
only. With access to the Maya model (or similar), the temporal template for any given view could in principle be generated on the fly.

To decide on which exemplar to use, a homography is estimated from the image plane to a coordinate frame on the ground-plane. This is estimated manually, but automatic calibration methods could be used instead. These methods vary between utilising trajectories of walking pedestrians [19] or moving cars [27], [2], to estimating calibration parameters from line segments in the static scene [22]. The point on the ground-plane directly below the camera is estimated from the vertical vanishing point. The angle between the line connecting this point to the pedestrian and the direction of the pedestrian's motion gives the viewing direction, assuming the pedestrian is facing the direction of motion (Figure 8). This ignores the elevation of the camera above the ground. The mean of the computed viewing directions over the clip's frames is used to select the corresponding exemplar.

The chosen exemplar is first scaled so that its height is the same as that of the generated temporal template. The median coordinate of the temporal template is aligned with that of the corresponding exemplar. An exhaustive search is then performed for the best match over a range of transformations. In the results, the search ranges for scales, rotations and translations are [0.75:0.05:1.25], [-15:5:15] degrees and [-30:3:30] pixels respectively. The match between two temporal templates is an L_1-like
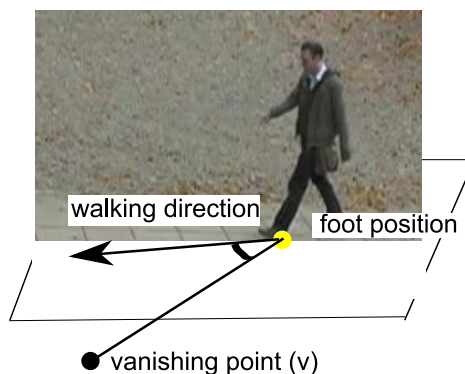

Fig. 8. For each frame, the foot projection is mapped from the image plane to the ground plane using an estimated homography. The viewpoint angle is that between the walking direction and the line connecting the vertical projection of the camera to the foot position.

Fig. 6. Eight cameras for capturing the Silhouettes at EPFL. Diagram from [9]

metric, weighted by the vertical coordinate of each pixel, giving higher weight to the head and shoulder region. The cost of matching a transformed model (M_T) to the person's temporal template (P) is shown in (1), where y ranges from 1 at the bottom to h at the top of the temporal template.

$$d(P, M_T) = \sum_{(x,y)=(1,1)}^{(w,h)} |P(x,y) - M_T(x,y)|y \qquad (1)$$

The best transformation $\widehat{T}$ is the one that minimises the matching cost

$$\widehat{T} = \operatorname*{argmin}_{T} \; d(P, M_T) \qquad (2)$$

Figure 9 shows an example of such a match and the located global minimum. The best transformation $\widehat{T}$ is then used to identify areas protruding from the temporal template:

$$v(x,y) = \max(0, P(x,y) - M_{\widehat{T}}(x,y)) \qquad (3)$$

Pixels where $P(x,y) < (M)_{\widehat{T}}(x,y)$ are assumed to have been caused by noise, or poor foreground segmentation. Carried objects can then be detected from the temporal protrusion of the pixels represented by v (Figure 9c) using one of two methods. The first method uses the periodicity of these pixels to differentiate the limbs from other protrusions. Section 3.1 explains how the periodicity analysis is performed. A simplified version of this method has been used in [16] to classify asymmetric regions into detected objects and other protrusions. Alternatively, Section 3.2 segments detected objects using a binary-labeled MRF formulation, combining prior information and spatial continuity.
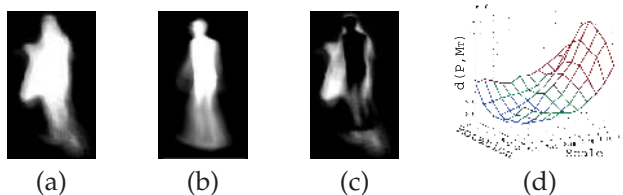

Fig. 9. The temporal template of the person (a) is matched to the corresponding exemplar (b), the global minimum (d) results in a map of temporal protrusion v of the pixels (c). In (d), the matching cost of the best translation for different scales and rotations is shown.

3.1 Classification by periodicity analysis

Periodicity analysis was proposed by Haritaoglo *et al.* to distinguish carried objects from other asymmetric regions. The algorithm for periodicity analysis described here is based on the original work by Cutler and Davis [6], [7].

After aligning foreground segmentations using ICP, L_1 is used to compare two aligned foreground segmentations. For a sequence of n frames, an $n \times n$ matrix records
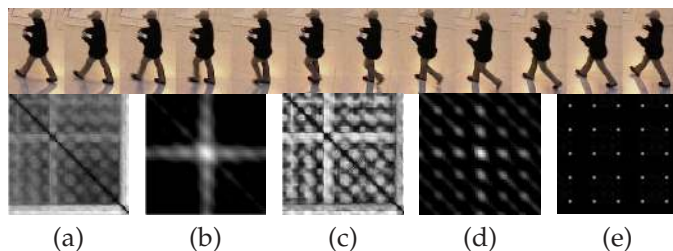

Fig. 10. The sequence on top shows 12 frames representing half a walking cycle. The frequency (f=12) is found using periodicity analysis. First, the similarity matrix is calculated (a). When (a) is used to compute the autocorrelation matrix (b), the periodicity is not obvious. Adaptive histogram equalisation is applied to (a) to generate a contrast enhanced image (c). The resultant autocorrelation matrix (d) shows clear periodicity, and the chosen square lattice (e) represents the correct frequency (f=12).

the L_1 difference between each pair of frames in the sequence, and is referred to as the 'similarity matrix'. Figure 10a illustrates the similarity matrix (S) where darker cells indicate higher similarity. The contrast in the similarity image is sometimes not so clear. Thus adaptive histogram equalisation is used to enhance the contrast. This contrast-enhancement step is added to the original Cutler and Davis algorithm as it improves the performance for noisy foreground segmentations. Next, the similarity matrix (S) is converted to an autocorrelation matrix (A) using (10) from [7]. The size of the autocorrelation matrix depends on the autocorrelation region R around each point in the similarity matrix.

$$A(dx, dy) = \frac{\sum\limits_{(x,y)\in R} (V(x,y)V(x+dx, y+dy))}{\sqrt{\sum\limits_{(x,y)\in R} V(x,y)^2 \sum\limits_{(x,y)\in R} V(x+dx, y+dy)^2}}$$

$$(4)$$

In (4), $V(x,y) = S(x,y) - \overline{S_R(x,y)}$ where S_R is the region of size R centred around (x,y). The function V subtracts the mean of the values in region R centred at (x,y) from the similarity value $S(x,y)$.

After obtaining the autocorrelation matrix (Figure 10d), $45°$ square lattices are used to find the dominant frequency. For a range of possible frequencies $d \in [minFreq, maxFreq]$, an $R \times R$ lattice is created with the frequency d. A normalized L_1 measure (i.e. divided by the number of points in the lattice) between the autocorrelation matrix and the binary lattice is used compare different frequencies. The frequency of the lattice with the minimum normalised L_1 measure is selected as the dominant frequency. If multiple minima are found, the frequency with the smallest value is chosen as the sequence's frequency. Figure 10 presents an example of how the dominant frequency is found. The confidence of the periodicity is defined as the percentage of points in the lattice that correspond to a local maximum with an error below a specified threshold.

In addition to the periodicity analysis performed for

the full body, a similar analysis is performed for each protruding region. A threshold for $v(x, y)$ from (3) is selected, and the pixels are grouped into connected components. Small connected components, composed of less than a minimum number of pixels, are removed. For each remaining connected component, the foreground images are masked by the connected component, and the masked foreground images are re-analysed for periodicity. If f_b is the dominant frequency of the full body, and f_c is the frequency of the connected component then the difference is used to classify protrusions using a threshold. Each connected component c is labeled as belonging to a carried object ($m_c = 1$) or not ($m_c = 0$).

$$m_c = \begin{cases} 1 & |f_b - f_c| > \alpha \\ 0 & otherwise \end{cases} \quad (5)$$

This periodicity analysis requires a sufficient number of cycles to produce useful autocorrelation matrices. The carried-object detector presented in this paper relies on short video sequences, to ensure the person does not change walking direction during the sequence. Short sequences often fail to show any detectable periodicity. Section 3.2 introduces a different approach to segment carried objects.

3.2 Classification using a spatial prior and continuity

Protrusions from the exemplar temporal template can be at locations where carried objects are not expected, due, for example, to the wearing of hats on top of heads or undetected shadow under the feet. We construct a spatial prior for carried objects in each of the eight views ($d = 1..8$). Thus $\theta_d(x, y)$ is the probability that the pixel (x, y) in view d belongs to a carried object.

These probabilities are estimated from the occurrence of carried objects at each pixel in ground-truth temporal templates. Training values are also used to estimate the distribution of protrusion values conditioned on their labelling. Finally, this information is combined into a binary Markov Random Field (MRF), determining an energy function which is minimised. The spatial prior for carried objects is estimated from training data by mapping the protrusion map (Figure 9c), using the inverse of the best transformation, to align to its corresponding exemplar. Each spatial position $\mathrm{x} = (x, y)$ has to be labeled as belonging to a carried object ($m_\mathrm{x} = 1$) or not ($m_\mathrm{x} = 0$). Using the raw protrusion values $v(\mathrm{x})$ calculated in (3), the class-conditional densities $p(v|m_\mathrm{x} = 1)$ and $p(v|m_\mathrm{x} = 0)$ are modeled as Gaussians based on training data.

The energy function to be minimised $\mathrm{E}(m)$ over Image I is given by (6).

$$\mathrm{E}(m) = \sum_{\mathrm{x} \in I} \Big(\phi(v|m_\mathrm{x}) + \omega(m_\mathrm{x}|\theta_d) \Big) + \sum_{(\mathrm{x},\mathrm{z}) \in \mathcal{C}} \psi(m_\mathrm{x}, m_\mathrm{z}) \quad (6)$$

$\phi(v|m_\mathrm{x})$ represents the cost of assigning a label to x based on its protrusion value $v(\mathrm{x})$ in the image:

$$\phi(v|m_\mathrm{x}) = \begin{cases} -\log(p(v|m_\mathrm{x} = 1)) & \text{if } m_\mathrm{x} = 1 \\ -\log(p(v|m_\mathrm{x} = 0)) & \text{if } m_\mathrm{x} = 0 \end{cases} \quad (7)$$

$\omega(m_\mathrm{x}|\theta_d)$ is based on the map of prior probabilities θ_d given a specified walking direction d:

$$\omega(m_\mathrm{x}|\theta_d) = \begin{cases} -\log(\theta_d(\mathrm{x})) & \text{if } m_\mathrm{x} = 1 \\ -\log(1 - \theta_d(\mathrm{x})) & \text{if } m_\mathrm{x} = 0 \end{cases} \quad (8)$$

The interaction potential ψ follows the Ising model over the cliques, where $\mathcal{C}$ represents all the pairs of 4-connected neighboring pixels in the image I:

$$\psi(m_\mathrm{x}, m_\mathrm{z}) = \begin{cases} \lambda & \text{if } m_\mathrm{x} \neq m_\mathrm{z} \\ 0 & \text{if } m_\mathrm{x} = m_\mathrm{z} \end{cases} \quad (9)$$

The interaction potential ψ is fixed regardless of the difference in protrusion values v at x and z, because the protrusion values represent the temporal continuity, and not the colour or texture information. The smoothness cost term λ in (9) affects the size of connected components in the resulting segmentation. Smaller λ segments smaller protrusions and detects smaller carried objects, yet increases the chance of segmenting other small protrusions resulting from clothing or body build. In the experiments, the term λ was optimised over a training set then used in testing.

3.3 Relative height for combining detections

In most cases, the same carried object will be detected in several clips along the trajectory. Detections from all clips are combined, so the carried object is detected throughout the trajectory's frames when visible. This is done exhaustively, starting from the first clip along the trajectory. For each carried object, Figure 12 shows the vertical extent of the temporal template (h_1, h_2) and that of the detected bag (b_1, b_2). The relative height for a carried object $\mathrm{vbounds}(c)$ is a normalized tuple as follows:

$$\mathrm{vbounds}(c) = \left[\frac{b_2 - h_2}{h_1 - h_2}, \frac{b_1 - h_2}{h_1 - h_2} \right] \quad (10)$$
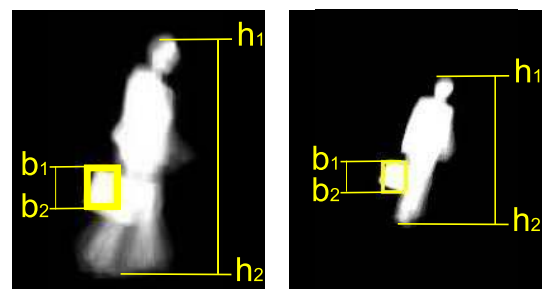

Fig. 12. The height of the baggage relative to the temporal template is used to relate the detections of the same carried object across the clips of a single trajectory.

Fig. 11. Frame detections are compared to ground-truth (yellow). When more than 50% of frames result with an overlap above 15% (green), a detection is labelled as true-positive. Two true-positives are marked (cyan).

For the same trajectory, the relative heights of two carried objects in different clips are matched using the interval overlap in (11) where relative height tuples are treated as closed intervals. Two detections c_1, c_2 are considered of the same carried object if (11) is satisfied.

$$\frac{|\text{vbounds}(c_1) \cap \text{vbounds}(c_2)|}{|\text{vbounds}(c_1) \cup \text{vbounds}(c_2)|} > 0.5 \qquad (11)$$

When multiple detections are recorded and a carried object from one clip matches more than one carried object at another clip, the one with the highest overlap is chosen.

4 EXPERIMENTS AND RESULTS

This section presents results on two datasets. First the method is evaluated on the publicly available PETS2006 dataset. Next, the estimated priors from PETS2006 are used to detect carried objects in an outdoor dataset recorded on the campus of the University of Leeds (LEEDS2009). The ground truth of carried objects was manually obtained for both datasets. For each carried object, a ground-truth bounding box b_{gt} is recorded for each frame in which the object is present. For a detected carried object, the bounding box b_p is calculated by projecting the detection onto the foreground segmentation in every frame in the clips where the object was detected. The detection is then labeled as true positive if (12) is satisfied in more than 50% of the frames in the sequence.

$$\frac{\text{area}(b_p \cap b_{gt})}{\text{area}(b_p \cup b_{gt})} > 0.15 \qquad (12)$$

This measure of overlap is the same as that defined in [10]. A low overlap threshold is chosen because the ground truth bounding boxes enclose the whole carried object, while the methods compared in this paper (including [16]) only detect the parts of the object that do not overlap the body. Bounding boxes were chosen instead of pixel masks for simplicity. Figure 11 shows frames from a single trajectory that result in two true positive detections. Multiple detections of the same object are counted as false positives. As the method cannot deal with groups of people moving together, such trajectories were manually removed from both datasets after tracking.

4.1 PETS2006

The seven sequences of the PETS2006 dataset are used, and the third camera is selected, as there is a greater number of people seen from the side. The ground-plane homography was established using the calibration measurements provided as part of the dataset. Moving objects were tracked using an off-the-shelf tracker [20] to retrieve foreground segmentations. Trajectories shorter than 10 frames in length were discarded. The number of individually tracked people was 106.

The carried objects in the dataset varied between boxes, hand bags, briefcases and suitcases. Unusual objects are also present, like a guitar in one example. The PETS2006 dataset is recorded at a train station. Travel bags and dragged suitcases are thus frequent in the sequences. In some cases, people were carrying more than one object. Ground truth for carried objects was obtained manually for all 106 individuals. 83 carried objects were tracked, and the bounding box of each was recorded for each frame (Figure 14). The results

Fig. 13. Ground-truth bags are shown in yellow, true-positive in green and false-positive in red.
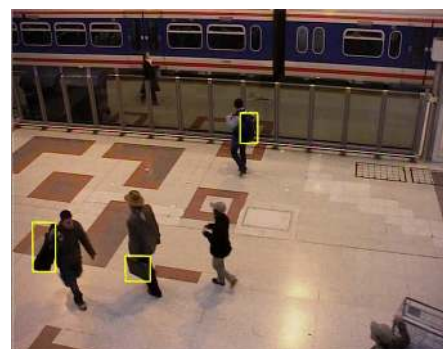

Fig. 14. PETS2006 third camera viewpoint showing ground truth bounding boxes representing carried objects.

compare the re-implementation of *Backpack* as specified in [14], [16] with the proposed method (Section 3). The same temporal templates are used as the input for both methods.
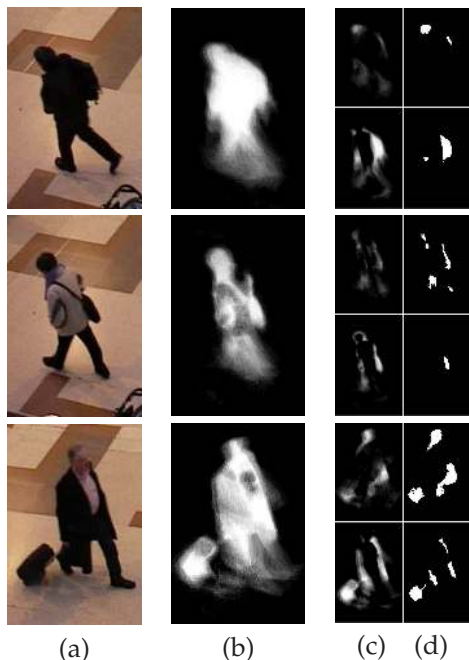

(a)　　　　　(b)　　　(c)　(d)

Fig. 15. Three examples (a), along with their temporal templates (b) are assessed using both techniques. The asymmetric regions (c-top) thresholded (d-top) and the protruding regions (c-bottom) thresholded (d-bottom) show some examples of how template matching retrieves better estimate of the carried objects.
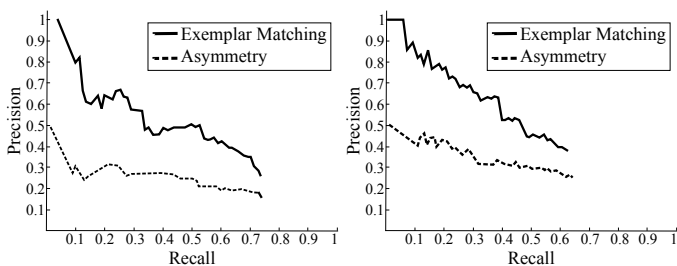

Fig. 16. PR curves for detecting carried objects by the method proposed in this paper (Exemplar Matching) compared to [16] (Asymmetry) without (left) and with (right) periodicity analysis to filter the retrieved blobs.

First, we compare asymmetry used in Haritaoglu's work [16] with detecting protrusions by matching to a view-specific exemplar. The pixels resulting from both methods are thresholded and grouped into connected components. Figure 15 shows some examples comparing asymmetry analysis with matching to an exemplar temporal template. Two parameters are varied to draw the Precision-Recall (PR) curves in Figure 16 (left), one for the threshold Figure 15(c-d) and one for the minimum size of the accepted connected component. The curves are generated by linearly interpolating the points
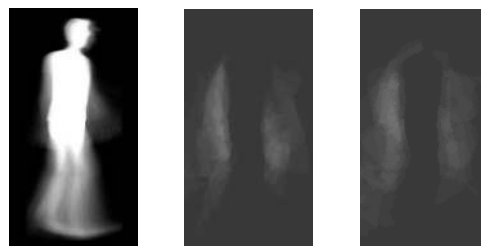

Fig. 17. For the second exemplar (left), $\theta_{2,6}$ was generated using the first (middle) and the second (right) training sets. The spatial prior θ has high values where stronger evidence of carried objects had been seen in training. A prior of 0.2 was used when no bags were seen.

representing the maximum precision for each recall. Maximum precision on a recall of 0.5, for example, was improved from 0.25 to 0.51. Maximum recall was 0.74 for both techniques, as noisy temporal templates and non-protruding carried objects affect both techniques.

The results are then compared with periodicity analysis. To achieve that, all optimal setting points along the curves in Figure 16 (left) are used to detect protrusions. For each pair of parameters, the frequencies of connected components are compared to that of the full body. Two thresholds for periodicity analysis are varied. These are for the minimum confidence and the threshold α in (5). Figure 16 (right) shows PR curves analogous to those in Figure 16 (left) but now combined with periodicity analysis, again taking the maximum precision for each recall. For periodicity analysis, the region R is set to 30 in all experiments, and the frequency range to $d \in [8{:}1{:}20]$. By implementing the Cutler-Davis periodicity analysis (Section 3.1), only 35% of the retrieved protrusions showed any detectable periodic motion. The improved performance of the matching method is still apparent. In addition, comparing the corresponding curves shows that periodicity analysis helps improve the performance for both methods.

Next, spatial continuity is assumed along with trained priors (Section 3.2). The pedestrians in the dataset were divided into two sets, the first containing 56 pedestrians (Sets 1-4 in PETS2006) and the second containing 50
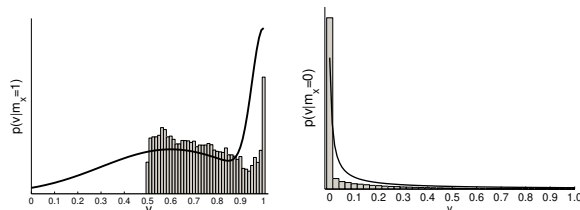

Fig. 18. Pixel values distribution for objects (left) and non-objects (right) $v(\mathbf{x})$. Thresholded pixels ($>$0.5) that match true detections are grouped and compared to ground truth, then are used to train $p(v|m_{\mathbf{x}} = 1)$. The rest are used to train $p(v|m_{\mathbf{x}} = 0)$.

pedestrians (Sets 5-7). The results are presented using two-fold cross validation. At each fold, one set is used to estimate the spatial prior θ_d, and carried objects are detected in the other test set. To estimate the priors, connected components are obtained from the map of protrusions v using a threshold of 0.5. Correct detections, by comparing to bounding boxes from the ground truth, are used to train for spatial priors of carried objects separately for each directionally-specific exemplar. To make use of the small training set, maps of opposite exemplars are combined. For example, the first and the fifth exemplars are separated by $180°$. θ_1 and θ_5 are thus combined by horizontally flipping one and calculating the weighted average $\theta_{1,5}$ (by the number of blobs). The same applies for $\theta_{2,6}$, $\theta_{3,7}$ and $\theta_{4,8}$. Figure 17 shows $\theta_{2,6}$ using the two disjoint training sets. A sufficient number of carried objects from each viewing angle is required for good spatial priors. We have shown that the same spatial priors can be applied over a range of viewing angles without a substantial reduction in performance (Section 4.2). Alternatively, simulated training for carried object locations can be used. In principle, view-specific priors could be generated automatically given examples of simulated carried bags associated with the 3D Maya model.

Figure 18 presents the distribution of $v(\mathrm{x})$ for carried objects ($m_{\mathrm{x}} = 1$) and other protrusions ($m_{\mathrm{x}} = 0$). By studying these density distributions, $p(v|m_{\mathrm{x}} = 1)$ was approximated by two Gaussian distributions, one for stable carried objects, and another for swinging objects. The parameters of the two Gaussians were manually chosen to approximately fit the training density distributions.

$$p(v|m_{\mathrm{x}} = 1) = \gamma\mathcal{N}(v; 0.6, 0.3) + (1-\gamma)\mathcal{N}(v; 1.0, 0.05) \quad (13)$$

The density distribution $p(v|m_{\mathrm{x}} = 0)$ resembles a reciprocal function. It was thus modeled as:

$$p(v|m_{\mathrm{x}} = 0) = \frac{1/(v + 0.01)}{\log(1 + 0.01) - \log(0.01)} \quad (14)$$

The denominator represents the area under the curve for normalisation. The max-flow algorithm, proposed in [3], and its publically available implementation, minimises the energy function (6) retrieving regions representing carried objects. Both γ in (13) and λ in (9) were optimised on each training set (Table 1).

	Set1	Set2
γ in 13	0.64	0.66
λ in 9	2.2	2.5

TABLE 1
Parameter optimisation for two-fold cross validation on the PETS2006 dataset

Table 2 presents the precision and recall results along with the actual counts for the complete dataset, showing that classification using spatial prior and continuity - implemented as a MRF - produces higher precision

and recall results. The results show that the periodicity analysis approach used in [16] did not achieve more than 27% recall rate as the horizontal projection histogram often results in a different periodicity than that of the carried object, and similar to the periodicity of the full body.

	Precision	Recall	TP	FP	FN
Thresholding	39.8%	49.4%	41	62	42
Periodicity [16]	27.7%	27.4%	23	61	60
Periodicity [7] (Sec 3.1)	45.2%	50.6%	42	52	41
Spatial prior - MRF (Sec 3.2)	50.5%	55.4%	46	45	37

TABLE 2
Better performance was achieved by introducing spatial prior and continuity. Results are compared to simple thresholding and periodicity analysis. The Cutler and Davis [7] periodicity algorithm outperforms the horizontal slicing approach from [16].

When analysing the periodicity results using the method from [7] on the 106 trajectories from the PETS dataset, 15 trajectories failed to show any periodic pattern. All the 15 trajectories were less than 26 frames in length. For the remaining 91 trajectories, the estimated periodicity varied between 8 and 14 frames with an average of 10.7 frames ($\sigma = 1.22$). The number of frames required to estimate a periodic motion (with a confidence above 80%) varied between 23 and 30 frames, with an average of 24.27 frames ($\sigma = 1.15$).

The results above used the ranking technique proposed in Section 3 for generating the temporal template. When the temporal template is generated using the more expensive LMedS estimator, the precision and recall values are exactly the same as those in Table 2. The time required to compute the temporal template increases from an average of 2μs when ranking is used to an average of 64μs when LMedS estimator is used (computed using Matlab on a 2.53GHz desktop). Figure 19 shows an example of visual differences between the temporal templates generated using the two methods.

To evaluate the effect of the spatial prior, we experiment with removing the term $\omega(m_{\mathrm{x}}|\theta_d)$ from the
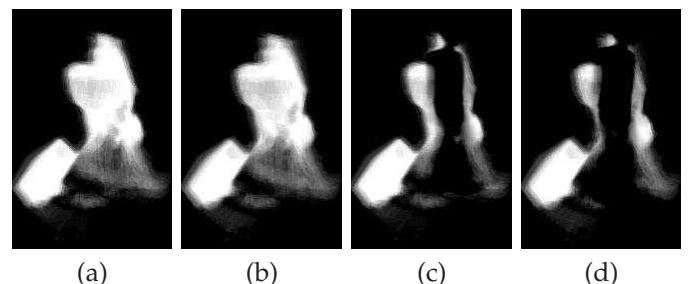

(a)　　　(b)　　　(c)　　　(d)

Fig. 19. An example of the visual differences between the temporal template and difference images generated using the ranking (a,c) and the LMedS estimator (b,d). The three carried bags (Figure 22(c)) are correctly detected in both cases.

energy function. The PR curves in Figure 20 demonstrate the advantage of introducing spatial priors, and the examples in Figure 21 show how prior models affect estimating carried objects.
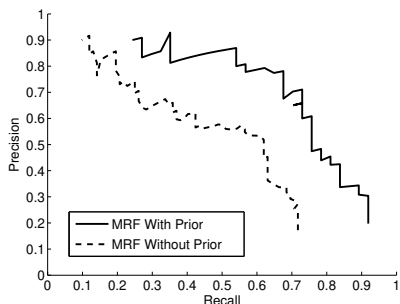

Fig. 20. PR Curves for detecting carried objects using MRF. Introducing spatial priors for carried objects produces better performance. λ in (9) was varied between [0.1:0.1:6] to produce the PR curves.

Reasons behind FP detections	
Protruding parts of clothing	15
Protruding body parts	10
Extreme body proportions	6
Incorrect template matching	5
Noisy temporal template	5
Duplicate matches	4
Total	45
Reasons behind FN detections	
Bag with little or no protrusion	9
Dragged bag tracked separately by tracker	6
Carried object between legs	5
Carried object not segmented from background	4
Little evidence of spatial prior in training	3
Swinging small object	3
Noisy template	3
Incorrect template matching	2
Merging two protruding regions into one	2
Total	37

TABLE 3
Reasons behind False Positive (FP) and False Negative (FN) detections.

Quantitatively, for the 45 false positive, and 37 false negative cases in the last row of Table 2, Table 3 dissects these results according to the reason of their occurrence. Figure 22 presents a collection of results highlighting reasons for success and the main sources of failure.

4.2 LEEDS2009

The second dataset consists of a continuous 12 hour period recorded during a working day (7am - 7pm). The tracker retrieved the set of trajectories that passed through the zone surrounding a building entrance (marked with a grey rectangle in Figure 23). Trajectories thus correspond to staff and students entering and exiting a university building entrance. After manually
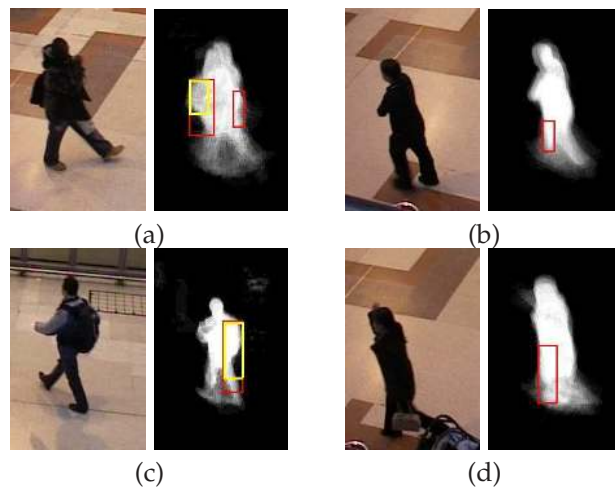

Fig. 21. Carried object detections with spatial prior (yellow) and without (red). Prior information drops candidate blobs at improbable locations (a,b), and better segments the object (a,c). It nevertheless decreases support for carried objects in unusual locations (d).
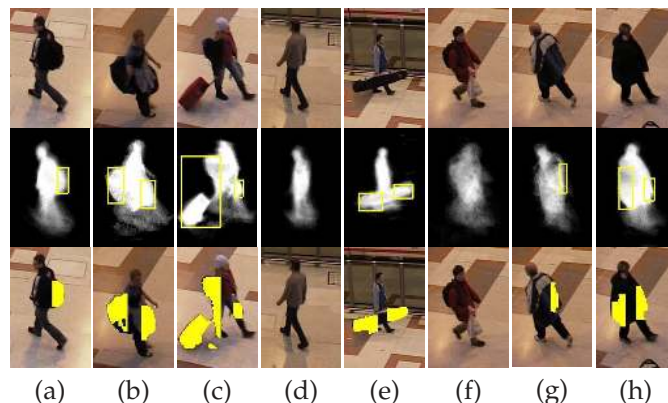

Fig. 22. The proposed method can identify single (a) or multiple (b,c) carried objects. (d) shows its ability to classify true negative cases. Objects extending over the body are split into two (e). Failure cases may result from poor temporal templates due to poor foreground segmentation (f). The spatial prior could favor some false positive objects (g). This method is not expected to cope with extreme body proportions (h). The second row shows the detections projected into the temporal templates, and the third row shows detections projected into a single frame of the sequence.

removing groups of people walking together, 326 trajectories remained for baggage detection. The ground truth consists of 233 carried objects that vary between backpacks, handbags, suitcases, shopping bags, boxes and carried coats.

Figure 23 shows the dataset's viewpoint along with three frames for the same person at different times. This dataset differs from PETS2006 in that a person is tracked with a significant change in the viewing angle (relative to the camera) along the trajectory. The depth of view also introduces a change in scale along the trajectory

Fig. 23. A collage showing the viewpoint for the LEEDS2009 dataset, with the entrance zone shaded in gray, along with three frames from one trajectory showing the different viewpoints.

for people walking toward or away from the camera. The ground-plane homography was manually obtained, along with finding the vanishing point. The baggage detections for the complete dataset were based on the same spatial priors and smoothness cost trained using the PETS2006 dataset. The results demonstrate the ability of spatial priors to be transformed between different camera viewpoints and elevations as they are mapped to the same exemplars.

Figure 24 shows the baggage detection results for the same trajectory from three clips (showing one frame from each clip). As the baggage detector assumes the bag is protruding from the normal silhouette, different viewpoints give rise to different detections. While no protrusions are apparent from the first viewpoint, the second viewpoint enabled detecting the carried bag, while the third viewpoint showed both the carried bag and carried jacket as protrusions.

Table 4 presents the results for the LEEDS2009 dataset when compared to the ground truth. Notice that no training was performed on the current scene. All parameters were estimated from the PETS2006 dataset. On

	Precision	Recall	TP	FP	FN
MRF - Prior	34.51%	76.52%	176	334	54

TABLE 4
Results for the LEEDS2009 full sequence detections

this dataset, for all true-positive carried objects, the mean percentage of frames along the trajectory with an overlap that exceeded 15% with the ground-truth bounding box was 83% ($\sigma = 16\%$). For false-positive carried objects, the frames with an overlap above 15% had a mean of 5% ($\sigma = 12\%$).

Fig. 25. LEEDS2009 - a collection of correctly detected carried objects.

A selected collection of success and failure carried-object detections are shown in Figures 25 and 26. Figure 25 shows 8 trajectories with successful detections. The detections are shown on the temporal template and projected on a single frame in each case. Figure 26 shows 7 incorrect detections. They cover a range of cases in which the detector fails. Case (a) results from poor foreground segmentation. The tracked individual is wearing a jacket which is very similar to the background's colour. Camouflaging results in a noisy temporal template and incorrect detections. Failure case (b) results from the baggage not being segmented as part of the foreground.

Fig. 24. Three clips from the same trajectory. Different viewpoints result in different carried object detections.

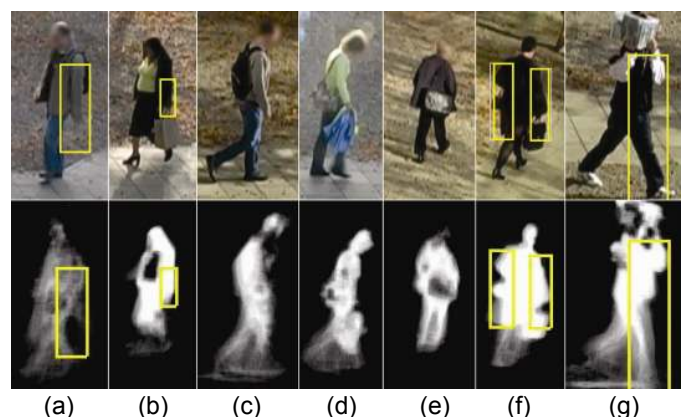(a)   (b)   (c)   (d)   (e)   (f)   (g)

Fig. 26. LEEDS2009 - a collection of incorrect baggage detections.

The stationary arm holding the bag is detected as the carried object instead. Cases (c), (d) and (e) are false negative cases where the carried object is not sufficiently protruding to be detected. Case (f) detects two carried objects but the bounding box extends to include the stationary arms carrying the objects as well as a protruding coat. The last case (g) fails in matching the temporal template to the unencumbered model. By reviewing the matching cost in (2), the match gives higher weight to matching the head and the shoulders of the model. In this example, the head and the shoulders are occluded by the carried object, which resulted in an incorrect match. Table 5 gives assumed reasons for all the false positive and false negative detections in Table 4.

Reasons for FP detections	
Protruding parts of clothing	48
Protruding body part	31
Static limbs	28
Extreme body proportions	11
Incorrect template matching	116
Noisy temporal template	48
Duplicate matches	4
Undetected shadow	9
Insufficient overlap with GT	39
Total	334
Reasons for FN detections	
Bag with little or no protrusion	12
Carried object between legs	9
Carried object not segmented from background	12
Little evidence of spatial prior in training	2
Noisy template	8
Incorrect template matching	11
Total	54

TABLE 5
LEEDS2009 - Reasons behind FP and FN detections.

5 APPLICATION

The carried-object detector presented in Section 3 is implemented in MATLAB and can be downloaded from: http://www.comp.leeds.ac.uk/vision/BaggageDetection. The system operates at around 5fps and assumes the foreground segmentation is available. This section presents one application in which the detector was utilised.

Applied to the LEEDS2009 dataset (Section 4.2), the baggage detector is used to assist linking people as they enter and exit a building entrance. Trajectories are classified into those entering or exiting the building based on their walking direction. A trajectory can be connected to one earlier trajectory of a different type, thus allowing to connect people entering a building to them exiting later, or a person exiting the building and returning to it later. People are matched by their clothing colour and projected height. The application tests whether the presence of carried objects can better connect the different appearances of the same individual. The height overlap criterion in (11) as well as an RGB histogram intersection [24] of the segmented pixels match
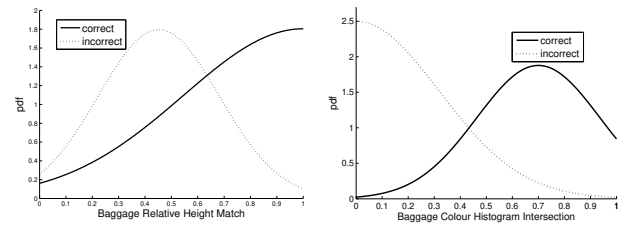

Fig. 27. Trained Gaussians for carried-object matches. Correct and incorrect relative height (left) and colour histogram intersection (right) are estimated as Gaussians.

the carried objects between trajectories. A separate 2 hour training sequence is used to fit a Gaussian to correct and incorrect values of the histogram intersection and height overlap. Figure 27 shows the trained Gaussians for the sequence.

The Hungarian algorithm (the Munkres algorithm) is used to find one-to-one optimal person matches. When carried-object detection is incorporated in the matching, the number of correctly connected trajectories increases from 7 to 15 while the number of falsely connected trajectories decreases from 175 to 131 trajectories. Figure 28 shows a collection of carried objects matched between different appearances of the same individual separate by a few hours each. The detector might be used in other applications like detecting abandoned baggage in hidden areas, exchanging objects and thefts.
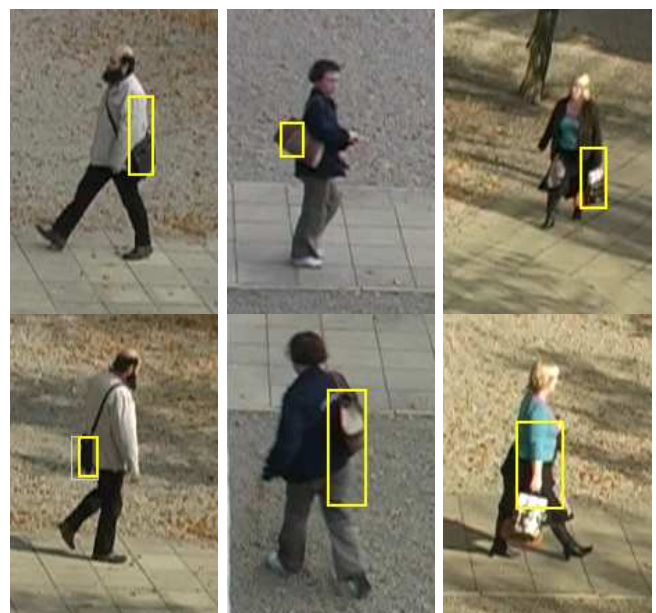

Fig. 28. Three examples of people correctly linked only when carried objects are considered in matching. Notice that in the last example, the two appearances are correctly linked despite some change in clothing.

6 CONCLUSION

We have proposed a novel method for detecting carried objects, aiming at higher robustness than noisy single

frame segmentations. Carried objects are assumed to protrude from the normal silhouette. We adopt the use of a temporal template, introduced for the same application in [16], but match against exemplars rather than assume that unencumbered pedestrians are symmetric. Evaluated on two dataset, one indoors and one outdoors, the method achieves a substantial improvement in performance over the state of the art. Training for spatial priors of carried objects and using an MRF to encode spatial constraints results in further improved performance. Results demonstrate that the spatial prior for carried object need not be trained for each new scene, and can be transformed to different datasets.

The method assumes that parts of the carried objects are protruding from the body's silhouettes. Due to its dependence on protrusion, the method cannot detect non-protruding carried objects. It may not be able to distinguish carried objects from protruding clothing or non-average build. Future improvements to this method might be achieved using texture templates to assist segmentation based on appearance. The detector can be used in different applications. The paper presents the results of one application that matches carried objects between different temporally-separate appearances of the same individual.

REFERENCES

[1] C. Benabdelkader and L. Davis. Detection of people carrying objects : a motion-based recognition approach. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition (FGR)*, pages 378–384, 2002.

[2] B. Bose and E. Grimson. Ground plane rectification by tracking moving objects. In *Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2004.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[4] A. Branca, M. Leo, G. Attolico, and A. Distante. Detection of objects carried by people. In *Proc. Int. Conf on Image Processing (ICIP)*, volume 3, pages 317–320, 2002.

[5] C. H. Chuang, J. W. Hsieh, L. W. Tsai, S. Y. Chen, and K. C. Fan. Carried object detection using ratio histogram and its application to suspicious event analysis. *IEEE Trans. on Circuits and Systems for Video Technology*, 19(6):911–916, 2009.

[6] R. Cutler and L. Davis. View-based detection and analysis of periodic motion. In *Proc. Int. Conf. on Pattern Recognition*, volume 1, pages 495–500, 1998.

[7] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):781–796, 2000.

[8] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 1997.

[9] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose detection using Bayesian spatio-temporal templates. *Computer Vision and Image Understanding*, 104(2):127–139, 2006.

[10] M. Everingham and J. Winn. The PASCAL visual object classes challenge (VOC2007) development kit. Technical report, 2007.

[11] J. Ferryman, editor. *Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*. IEEE, New York, 2006.

[12] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[13] N. M. Ghanem and L. S. Davis. Human appearance change detection. In *Proc. Image Analysis and Processing (ICIAP)*, pages 536–541, 2007.

[14] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis. Backpack: detection of people carrying objects using silhouettes. In *Proc. Int. Conf. on Computer Vision (ICCV)*, volume 1, pages 102–107, 1999.

[15] I. Haritaoglu, D. Harwood, and L. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *Proc. IEEE Workshop on Visual Surveillance*, 1999.

[16] I. Haritaoglu, D. Harwood, and L. S. Davis. W^4: real-time surveillance of people and their activities. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):809–830, 2000.

[17] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(4):663–671, 2006.

[18] C.-S. Lee and A. Elgammal. Carrying object detection using pose preserving dynamic shape models. In *Proc. Conf. of Articulated Motion and Deformable Objects (AMDO)*, pages 315–325, 2006.

[19] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In *Proc. Int. Conf. on Pattern Recognition (ICPR)*, volume 1, pages 562–567, 2002.

[20] D. Magee. Tracking multiple vehicles using foreground, background and motion models. In *Proc. Workshop on Statistical Methods in Video Processing*, pages 7–12, 2002.

[21] H. Nanda, C. Benabdelkedar, and L. Davis. Modelling pedestrian shapes for outlier detection: a neural net based approach. In *Proc. Intelligent Vehicles Symposium*, pages 428–433, 2003.

[22] R. Pflugfelder and H. Bischof. Online auto-calibration in man-made worlds. In *Proc. Digital Image Computing, Techniques and Applications (DICTA)*, 2005.

[23] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[24] M. Swain and D. Ballard. Color indexing. *Int. Journal of Computer Vision (IJCV)*, 7(1):11–32, 1991.

[25] D. Tao, X. Li, S. J. Maybank, and W. Xindong. Human carrying status in visual surveillance. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2006.

[26] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *Int. Journal of Computer Vision (IJCV)*, 13(2):119–152, 1994.

[27] Z. Zhang, M. Li, and T. Tan. Robust automated ground plane rectification based on moving vehicles for traffic scene surveillance. In *Proc. Int. Conf. on Image Processing (ICIP)*, pages 1364–1367, 2008.

Dima Damen (M'09) received BSc (2002) in computer science from Birzeit Universtiy, MSc (2003) in distributed multimedia systems and PhD (2009) in Computer Vision from University of Leeds, UK. Her thesis entitled 'Finding explanations for sets of events' focused on modeling the domain's knowledge and constraints using attributed grammars, and searching for the best global explanation in a Bayesian framework.

Dr. Damen is currently a research associate at the University of Bristol, working on cognitive workflow analysis with on-body visual sensors. She worked previously as a lecturer at both the University of Petra and Jordan University, Amman, Jordan. She is a member of IEEE and BMVA. Her research was published in BMVC, CVPR and ECCV publications.

David Hogg received the BSc degree in applied mathematics from the University of Warwick, the MSc degree in computer science from the University of Western Ontario, and the PhD degree from the University of Sussex. He was on the faculty of the School of Cognitive and Computing Sciences at the University of Sussex from 1984 until 1990 when he was appointed as full Professor of Artificial Intelligence at the University of Leeds, where he is now Pro-Vice-Chancellor for Research and Innovation. During 1999-2000 he was a visiting professor at the MIT Media Lab in Cambridge. He has been associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence (2006-2010) and program co-chair of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. His current research is on activity analysis, dealing especially with learning and the integration of qualitative and quantitative representations. He is a Fellow of ECCAI and a member of the IEEE Computer Society.