Detecting Carried Objects in Short Video Sequences

Dima Damen and David Hogg

School of Computing, University of Leeds
{dima,dch}@comp.leeds.ac.uk

Abstract. We propose a new method for detecting objects such as bags carried by pedestrians depicted in short video sequences. In common with earlier work [1,2] on the same problem, the method starts by averaging aligned foreground regions 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. 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 obtained from this match is combined in a Markov random field with a map of prior probabilities for carried objects and a spatial continuity assumption, from which we obtain a segmentation of carried objects using the MAP solution. We have re-implemented the earlier state of the art method [1] and demonstrate a substantial improvement in performance for the new method on the challenging PETS2006 dataset [3]. 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.

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 we adopt, and whilst there are other factors that may give rise to irregularities, such as clothing and build, experiments on a standard dataset are promising.

Although the method has been developed for the detection of objects carried by people, there could be applications of the approach in other domains where irregularities in the outline of known deformable objects are of potential interest.

We assume a static background and address errors in foreground segmentations due to noise and partial occlusions, by aligning and averaging segmentations to generate a so-called 'temporal-template' - this representation was originally proposed in [1] for the same application. The temporal template is then matched

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part III, LNCS 5304, pp. 154–167, 2008.

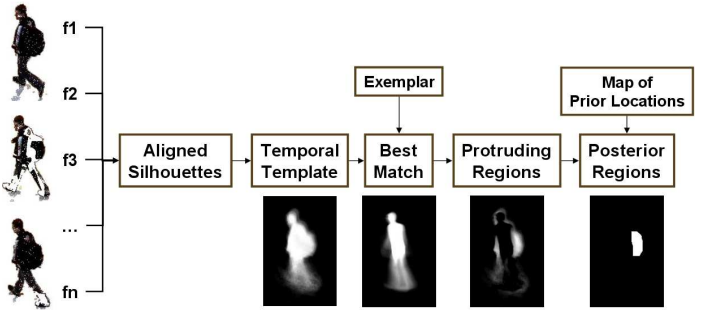

Fig. 1. All the frames in the the sequence are first aligned. The temporal template represents the frequency of each aligned pixel (relative to the median) being part of the foreground. The exemplar temporal template from a similar viewing angle is transformed (translation, scaling and rotation) to best match the generated temporal template. By comparing the temporal template to the best match, protruding regions are identified. MRF with a map of prior locations is used to decide on those pixels representing carried objects.

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, we incorporate prior information about the expected locations of carried objects together with a spatial continuity assumption in order to improve the segmentation of pixels representing the carried objects. Figure 1 summarizes, with the use of an example, the process of detecting carried objects.

Section 2 reviews previous work on the detection of carried objects. Section 3 presents our new method, based on matching temporal templates. Experiments comparing the performance of the earlier work from Haritaoglu *et al.* and our new method on the PETS2006 dataset are presented in Section 4. An extension of the method to incorporate locational priors and a spatial continuity assumption for detecting carried objects is presented in Section 5. The paper concludes with an overall discussion.

2 Previous Work

Several previous methods have been proposed for detecting whether an individual is carrying an object. The *Backpack* [1,2] 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 when a bag is not being carried, and that people exhibit periodic motion. Foreground segmentations are aligned using edge correlation. The aligned foreground masks for the complete video segment are combined into the temporal template that records the proportion of frames in the video sequence in which each (aligned) pixel was classified as foreground. Next, symmetry analysis is performed. The principal axis is computed using principal component analysis of 2-D locations, and is

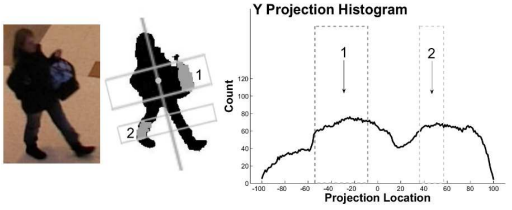

Fig. 2. For each foreground segmentation, the principal axis is found and is constrained to pass through the median coordinate of the foreground segmentation. Light gray represents the two detected asymmetric regions. Asymmetric regions are projected onto the horizontal projection histogram. Periodicity analysis is performed for the full histogram [Freq = 21] and for regions 1 [Freq = 11] and 2 [Freq = 21]. As region 2 has the same frequency as the full body, it is not considered a carried object.

constrained to pass through the median coordinate in the vertical and horizontal directions. For each location x, relative to the median of the blob, asymmetry is detected by reflecting the point in the principal axis. The proportion of frames in which each location was classified as asymmetric is calculated. 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 analyzing the periodicity of the horizontal projection histograms (See [1] for details). This estimates the periodic frequency of the full body, and that of each asymmetric region. *Backpack* assumes the frequency of an asymmetric blob that represents a limb is numerically comparable to that of the full body. Otherwise, it is believed to be a carried object. Figure 2 illustrates the process from our re-implementation.

From our own evaluation, errors in the *Backpack* method arise from four sources. Firstly, the asymmetric assumption is frequently violated. Secondly, the position of the principal axis is often displaced by the presence of the carried object. It may be possible to reduce this source of error by positioning the major axis in other ways, for example forcing it to pass through the centroid of the head [4] or the ground point of the person walking [5]. Thirdly, accurate periodicity analysis requires a sufficient number of walking cycles to successfully retrieve the frequency of the gait. Fourthly, the periodicity of the horizontal projection histogram does not necessarily reflect the gait's periodicity.

Later work by Benabdelkader and Davis [6] expanded the work of Haritaoglu *et al.* by dividing the person's body horizontally into three slices. The periodicity and amplitude of the time series along each slice is studied to detect deviations from the 'natural' walking person and locate the vertical position of the carried object. They verified 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).

Branca *et al.* [7] try to identify intruders in archaeological sites. Intruders are defined as those carrying objects such as a probe or a tin. The work assumes

a person is detected and segmented. Their approach thus tries to detect these 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. Nanda *et al.* [8] detect pedestrians carrying objects as outliers of a model for an unencumbered pedestrian obtained in a supervised learning procedure based on a three layer neural network. Alternatively, the work of Tao *et al.* [9] tries to detect pedestrians carrying heavy objects by performing gait analysis using General Tensor Discriminant Analysis (GTDA), and was tested on the USF HumanID gait analysis dataset.

Recent work by Ghanem and Davis [10] tackles detecting abandoned baggage by comparing the temporal template of the person before approaching a Region of Interest (ROI) and after leaving it. Carried objects are detected by comparing the temporal templates (the term 'occupancy map' was 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 ROI 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.

Our method uses the temporal template but differs from earlier work [1,10] by 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, were generated from reusable silhouettes used successfully for pose detection [11]. The use of temporal templates provides better immunity to noise in foreground segmentations. Our new approach does not require the pedestrian to be detected with and without the carried object, and can handle all normal viewpoints. It also generalizes to any type of carried object (not merely backpacks), and can be considered a general approach to protrusions from other deformable tracked objects.

This work provides the first real test of this task on the challenging PETS2006 dataset, which we have annotated with ground-truth for all carried objects. It is worth mentioning that this dataset does not depend on actors and thus records typical carried objects in a busy station. This enables us to demonstrate the generality of the approach, and clarify the real challenges of this task.

3 Description of the Method

Our method starts by creating the temporal template from a sequence of tracked pedestrians as proposed by Haritaoglu *et al.* [2]. We, though, introduced two changes to the procedure for creating the temporal template. Firstly, we apply Iterative Closest Point (ICP), instead of edge correlation, to align successive boundaries. ICP is performed on the edge points of the traced boundary around the foreground segmentation. Unlike edge correlation, this does not require a predefined search window, and in our experiments it gives a more accurate alignment

in the presence of shape variations between consecutive frames. Secondly, L_1 is used to rank the frames by their similarity to the generated temporal template. The highest ranked $p\%$ of the frames are used to re-calculate a more stable template. p was set to 80 in our experiments. The more computationally expensive Least Median of Squares (LMedS) estimator [12] gave similar results.

Having derived a temporal template from a tracked pedestrian, one of eight exemplars are used to identify protrusions by matching. These exemplar temporal templates represent a walking unencumbered pedestrian viewed from different directions. A set of exemplars for eight viewing directions was created using the dataset of silhouettes gathered at the Swiss Federal Institute of Technology (EPFL) [11]. The dataset is collected from 8 people (5 men and 3 women) walking at different speeds on a treadmill. Their motion was captured using 8 cameras and mapped onto a 3D Maya model. This dataset is comprised of all the silhouettes of the mapped Maya model, and has previously been used for pose detection, 3D reconstruction and gait recognition [11,13]. We average the temporal templates of different individuals in this dataset to create the exemplar for each camera view. The eight exemplars (Figure 3) are used for detecting the areas representing the pedestrian. The unmatched regions are expected to correspond to carried object(s).

To decide on which exemplar to use, we estimate a homography from the image plane to a coordinate frame on the ground-plane. We then use this to estimate the position and direction of motion of each pedestrian on the ground. 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 their direction of motion. We ignore the elevation of the camera above the ground in order to avoid having to generate new exemplars for different elevations, although this approximation may be unnecessary since generating the prototypes is fast and need only be done once. The mean of the computed viewing directions over the short video sequence is used to select the corresponding exemplar. Diagonal views (2,4,6,8) are used to match a wider range of angles (60°) in comparison to frontal views. This is because the silhouettes change more radically near frontal views.

The chosen exemplar is first scaled so that its height is the same as that of the generated temporal template. We align the median coordinate of the temporal template with that of the corresponding exemplar. An exhaustive search is then performed for the best match over a range of transformations.

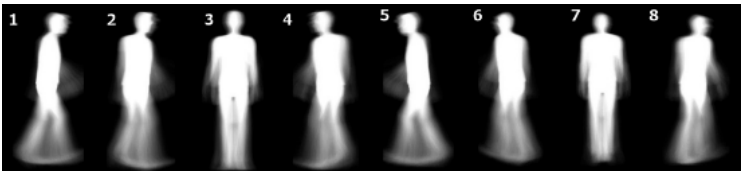

Fig. 3. The eight exemplar temporal templates, created to represent 8 viewpoints

(a) (b) (c) (d)

Fig. 4. The temporal template of the person (a) is matched to the corresponding exemplar (b), the global minimum (d) results in the protruding regions (c)

In our experiments, the chosen ranges for scales, rotations and translations were [0.75:0.05:1.25], [-15:5:15] and [-30:3:30] respectively. The cost of matching two templates is an L_1 measure, linearly weighted by the y coordinate of each pixel (plus a constant offset), giving higher weight to the head and shoulder region. Equation 1 represents the cost of matching a transformed model (M_T) to the Person (P), where h represents the height of the matched matrices.

$$d(M_T, P) = \sum_{x,y} |M_T(x,y) - P(x,y)|(2h - y) \tag{1}$$

The best match $\widehat{M_T}$ is the one that minimizes the matching cost

$$\widehat{M_T} = \underset{T}{\operatorname{argmin}}\ d(M_T, P) \tag{2}$$

Figure 4 shows an example of such a match and the located global minimum. The best match $\widehat{M_T}$ is then used to identify areas protruding from the temporal template:

$$\text{protruding}(x,y) = \max(0, P(x,y) - \widehat{M_T}(x,y)) \tag{3}$$

We are only concerned with areas in the person template that do not match body parts in the corresponding best match. Pixels where $P(x,y) < \widehat{M_T}(x,y)$ are assumed to have been caused by noise, or poor foreground segmentation. For the initial results in Section 4, the protruding values are thresholded and grouped into connected components representing candidate segmentations of carried objects. Another threshold limits the minimum area of accepted connected components to remove very small blobs. An enhanced approach is presented in Section 5 where segmentation is achieved using a binary-labeled MRF formulation, combining prior information and spatial continuity.

4 Experiments and Results

We used the PETS2006 dataset and selected the viewpoint from the third camera for which there is a greater number of people seen from the side. The ground-plane homography was established using the ground truth measurements provided as part of the dataset. Moving objects were detected and tracked using

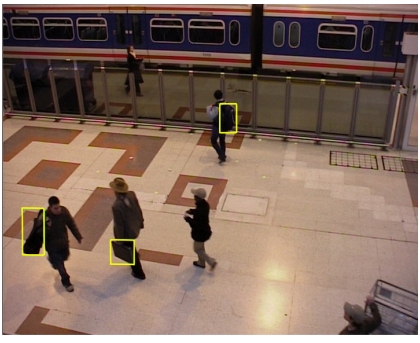

Fig. 5. PETS2006 Third camera viewpoint showing ground truth bounding boxes representing carried objects

a generic tracker [14] to retrieve foreground segmentations. The tracker has an automatic shadow remover that worked efficiently on the dataset. Trajectories shorter than 10 frames in length were discarded. As this method can not deal with groups of people tracked together, such trajectories were also manually removed. 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. In some cases, people were carrying more than one object. The number of individually tracked people was 106. 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 5). We chose bounding boxes instead of pixel masks for simplicity.

We compare our re-implementation of *Backpack* as specified in their papers [1,2] with our proposed method (Section 3). To ensure fair comparison, we use the same temporal templates as the input for both methods. A detection is labeled as true if the overlap between the bounding box of the predicted carried object (B_p) and that of the ground truth (B_{gt}) exceeds 15% in more than 50% of the frames in the sequence. The measure of overlap is defined by Equation 4 [15]:

$$\text{overlap}(B_p, B_{gt}) = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \tag{4}$$

A low overlap threshold is chosen because the ground truth bounding boxes enclose the whole carried object, while both methods only detect the parts of the object that do not overlap the body. Multiple detections of the same object are counted as false positives.

We first compare the blobs retrieved from both techniques without periodicity analysis. Each of the two algorithms has two parameters to tune, one for thresholding and one for the minimum size of the accepted connected component. Precision-Recall (PR) curves for the two methods are shown in Fig. 6 (left). These were generated by linearly interpolating the points representing the maximum precision for each recall. They show a substantial improvement in performance for the proposed method. Maximum precision on a recall of 0.5, for

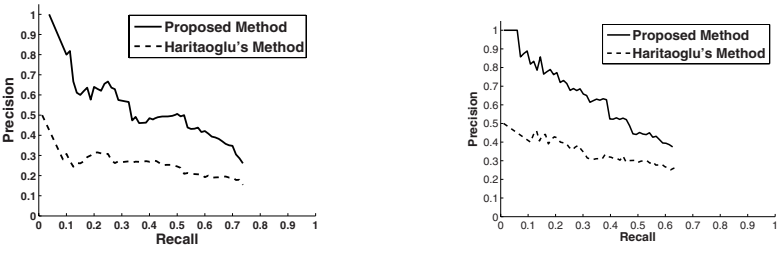

Fig. 6. PR curves for the proposed method compared to Haritaoglu *et al.*'s method without (left) and with (right) periodicity analysis to classify the retrieved blobs

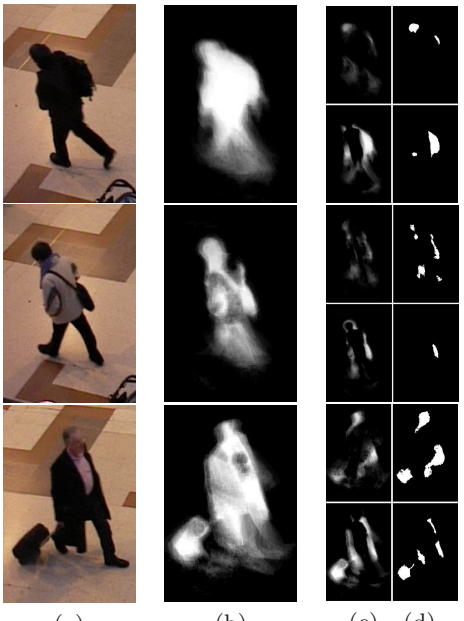

Fig. 7. Three examples (a), along with their temporal templates (b) are assessed using both techniques. Haritaoglu's method (c-top) thresholded (d-top) and our proposed method (c-bottom) thresholded (d-bottom) show some examples of how matching retrieves better estimate of the carried objects than symmetry.

example, was improved from 0.25 using asymmetry to 0.51 using matching. Maximum recall was 0.74 for both techniques, as noisy temporal templates and non-protruding carried objects affect both techniques. Figure 7 shows examples comparing asymmetry analysis with matching temporal templates.

To further compare the methods, we present the results after performing periodicity analysis. We thus take all optimal setting points represented by the curves in Fig. 6 (left), and vary the two thresholds for periodicity analysis. Figure 6 (right) shows PR curves analogous to those in Fig. 6 (left) but now

including periodicity analysis, again taking the maximum precision for each recall. The improved performance of our method is still apparent. In addition, comparing the corresponding curves shows that periodicity analysis improves the performance for both methods.

5 Using Prior Information and Assuming Spatial Continuity

The protruding connected components can be at locations where carried objects are not expected like hats on top of heads. We propose training for carried object locations relative to the person's silhouette to better differentiate carried objects from other protruding regions, and at the same time impose a spatial continuity assumption on the pixels corresponding to carried objects.

In this section, we show how training was used to generate a map of prior locations. Training values were also used to estimate the distribution of protrusion values conditioned on their labeling. Finally, this information is combined into a Markov random field, determining an energy function which is minimized. Results are presented along with a discussion of the advantages of training for prior locations. We divided the pedestrians into two sets, the first containing 56 pedestrians (Sets 1-4 in PETS2006) and the second containing 50 pedestrians (Sets 5-7). Two-fold cross validation was used to detect carried objects.

Training for carried object locations is accomplished by mapping the temporal template, using the inverse of the best transformation, to align its corresponding exemplar. During training, we obtain connected components using a threshold of 0.5. Correct detections, by comparing to bounding boxes from the ground truth, are used to train for locations of carried objects separately for each directionally-specific exemplar. A map of prior probabilities Θ_d is produced for each viewpoint d. Prior information for each location is calculated by the frequency of its occurrence within a correctly-detected carried object across the training set. To make use of our small training set, we combine the maps of opposite exemplars. 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 8 shows $\Theta_{2,6}$ using the two disjoint training sets.

We aim to label each location x within the person's temporal template as belonging to a carried object ($m_x = 1$) or not ($m_x = 0$). Using the raw protrusion values $v = \text{protruding}(x)$ calculated in Equation 3, we model the class-conditional densities $p(v|m_x = 1)$ and $p(v|m_x = 0)$ based on training data (Figure 9). By studying these density distributions, $p(v|m_x = 1)$ 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.

$$p(v|m_x = 1) = \gamma\mathcal{N}(v; 0.6, 0.3) + (1 - \gamma)\mathcal{N}(v; 1.0, 0.05) \tag{5}$$

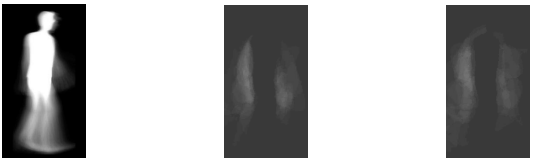

Fig. 8. For the second exemplar (left), $\Theta_{2,6}$(middle) was generated using sets 1-4, and $\Theta_{2,6}$(right) was generated using sets 5-7. The location model Θ has high values where stronger evidence of carried objects had been seen in training. A prior of 0.2 was used when no bags were seen. By symmetry, Θ_6 is a horizontal flip.

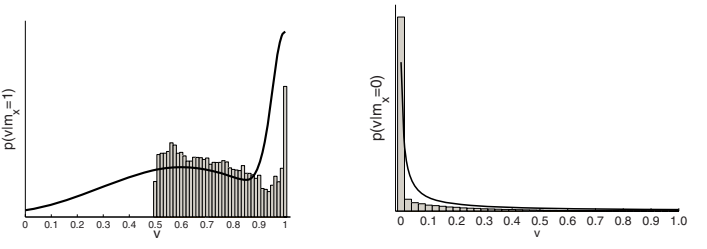

Fig. 9. Pixel values distribution for objects (left) and non-objects (right) protruding pixels. Thresholded pixels (>0.5) that match true detections when compared to ground truth, are used to train $p(v|m_x = 1)$. The rest are used to train $p(v|m_x = 0)$.

γ is the relative weight of the first Gaussian in the training set. Its value resulted to be 0.64 for the first set, and 0.66 for the second disjoint set. The density distribution $p(v|m_x = 0)$ resembles a reciprocal function. It was thus modeled as:

$$p(v|m_x = 0) = \frac{1/(v + \beta)}{\log(1 + \beta) - \log(\beta)} \tag{6}$$

β was set to 0.01. The denominator represents the area under the curve for normalization.

As we believe neighboring locations to have the same label, spatial continuity can be constrained using a Markov Random Field (MRF). The energy function to be minimized E(m) over Image I is given by Equation 7.

$$\mathrm{E}(m) = \sum_{x \in I} \Big(\phi(v|m_x) + \omega(m_x|\Theta) \Big) + \sum_{(x,y) \in \mathcal{C}} \psi(m_x, m_y) \tag{7}$$

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

$$\phi(v|m_x) = \begin{cases} -\log(p(v|m_x = 1)) & \text{if } m_x = 1 \\ -\log(p(v|m_x = 0)) & \text{if } m_x = 0 \end{cases} \tag{8}$$

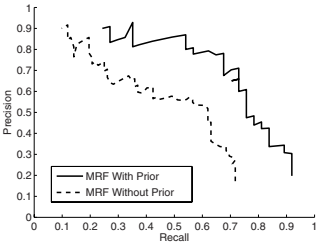

Fig. 10. PR Curves for detecting carried objects using MRF. Introducing location maps to encode prior information about carried object locations produces better performance.

(a) (b) (c) (d)

Fig. 11. The yellow rectangles show the choice of carried objects using MRF with location models. Red rectangles refer to MRF without location models. 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).

$\omega(m_x|\Theta)$ is based on the map of prior probabilities Θ given a specified walking direction:

$$\omega(m_x|\Theta) = \begin{cases} -\log(p(x|\Theta)) & \text{if } m_x = 1 \\ -\log(1 - p(x|\Theta)) & \text{if } m_x = 0 \end{cases} \tag{9}$$

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

$$\psi(m_x, m_y) = \begin{cases} \lambda & \text{if } m_x \neq m_y \\ 0 & \text{if } m_x = m_y \end{cases} \tag{10}$$

The interaction potential ψ is fixed regardless of the difference in protrusion values v at locations x and y. We did not choose a data-dependent term because the protrusion values represent the temporal continuity, and not the texture information at the neighboring pixels.

We use the max-flow algorithm, proposed in [16], and its publically available implementation, to minimize the energy function (Equation 7). Regions representing carried objects were thus retrieved. The smoothness cost term λ was optimized based on the used training set.

To evaluate the effect of introducing location models, the term $\omega(m_x|\Theta)$ was removed from the energy function and the results were re-calculated. λ was varied between [0.1:0.1:6] to produce the PR curves in Fig. 10 that demonstrate

Table 1. Better performance was achieved by introducing the MRF representation

	Precision	Recall	TP	FP	FN
Thresholding	39.8%	49.4%	41	62	42
MRF - Prior	50.5%	55.4%	46	45	37

the advantage of introducing location prior models. Examples in Fig. 11 show how prior models affect estimating carried objects.

In order to compare the MRF formulation with simple thresholding, we optimize the parameters on each training dataset and test on the other. For MRF, λ was optimized on the training datasets resulting in 2.2 and 2.5 respectively. Table 1 presents the precision and recall results along with the actual counts combined for the two test datasets, showing that MRF produces higher precision and recall results.

Quantitatively, for the 45 false positive, and 37 false negative cases, Fig. 12 dissects these results according to the reason for their occurrence. Figure 13 presents a collection of results highlighting reasons for success and the main sources of failure. We also demonstrate the video results at http://www.comp.leeds.ac.uk/dima/ECCVDemo.avi.

6 Conclusion

We have proposed a novel method to detect carried objects, aiming at higher robustness than noisy single frame segmentations. Carried objects are assumed to cause protruding regions from the normal silhouette. Like an earlier method we use a temporal template but match against exemplars rather than assuming that unencumbered pedestrians are symmetric. Evaluated on the PETS2006 dataset, the method achieves a substantial improvement in performance over the previously published method. Finally, we train on possible locations of carried objects and use an MRF to encode spatial constraints resulting in a further improvement in performance.

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 prior location in training	3
Swinging small object	3
Noisy template	3
Incorrect template matching	2
Merging two protruding regions into one	2
Total	37

Fig. 12. Reasons behind False Positive (FP) and False Negative (FN) detections

Fig. 13. The proposed method can identify single (1) or multiple (2,3) carried objects. (4) shows its ability to classify true negative cases. Objects extending over the body are split into two (5). Failure cases may result from poor temporal templates due to poor foreground segmentation (6). The map of prior locations could favor some false positive objects (7). The method is not expected to cope with extreme body proportions (8). The second row shows the detections projected into the temporal templates, and the third row shows detections projected into the images.

Due to its dependence on protrusion, the method 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 color information. In addition, the independence assumption in learning prior bag locations could be studied to utilize shapes of previously seen bags in producing better segmentations. When matured, this technique can be embedded into surveillance and security systems that aim at tracking carried objects or detecting abandoned objects in public places.

Acknowledgement

We would like to thank Miodrag Dimitrijevic at the CVLAB, EPFL and his colleagues for providing the dataset of silhouettes used in our research.

References

1. Haritaoglu, I., Cutler, R., Harwood, D., Davis, L.S.: Backpack: detection of people carrying objects using silhouettes. In: Proc. Int. Conf. on Computer Vision (ICCV), vol. 1, pp. 102–107 (1999)

2. Haritaoglu, I., Harwood, D., Davis, L.: W^4: real-time surveillance of people and their activities. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(8), 809–830 (2000)
3. Ferryman, J. (ed.): IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS). IEEE, New York (2006)
4. Haritaoglu, I., Harwood, D., Davis, L.: Hydra: Multiple people detection and tracking using silhouettes. In: Proc. IEEE Workshop on Visual Surveillance (1999)
5. Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., Maybank, S.: Principal axis-based correspondence between multiple cameras for people tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(4), 663–671 (2006)
6. Benabdelkader, C., Davis, L.: Detection of people carrying objects: a motion-based recognition approach. In: Proc. Int. Conf. on Automatic Face and Gesture Recognition (FGR), pp. 378–384 (2002)
7. Branca, A., Leo, M., Attolico, G., Distante, A.: Detection of objects carried by people. In: Proc. Int. Conf on Image Processing (ICIP), vol. 3, pp. 317–320 (2002)
8. Nanda, H., Benabdelkedar, C., Davis, L.: Modelling pedestrian shapes for outlier detection: a neural net based approach. In: Proc. Intelligent Vehicles Symposium, pp. 428–433 (2003)
9. Tao, D., Li, X., Maybank, S.J., Xindong, W.: Human carrying status in visual surveillance. In: Proc. Computer Vision and Pattern Recognition (CVPR) (2006)
10. Ghanem, N.M., Davis, L.S.: Human appearance change detection. In: Image Analysis and Processing (ICIAP), pp. 536–541 (2007)
11. Dimitrijevic, M., Lepetit, V., Fua, P.: Human body pose detection using Bayesian spatio-temporal templates. Computer Vision and Image Understanding 104(2), 127–139 (2006)
12. Rousseeuw, P.J.: Least median of squares regression. Journal of the American Statistical Association 79(388), 871–880 (1984)
13. Fossati, A., Dimitrijevic, M., Lepetit, V., Fua, P.: Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In: Proc. Computer Vision and Pattern Recognition (CVPR) (2007)
14. Magee, D.: Tracking multiple vehicles using foreground, background and motion models. In: Proc. ECCV Workshop on Statistical Methods in Video Processing, pp. 7–12 (2002)
15. Everingham, M., Winn, J.: The PASCAL visual object classes challenge (VOC 2007) development kit. Technical report (2007)
16. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(11), 1222–1239 (2001)