Carried Object Detection based on an Ensemble of Contour Exemplars

Farnoosh Ghadiri[1], Robert Bergevin[1], Guillaume-Alexandre Bilodeau[2]

[1] LVSN-REPARTI, Université Laval
farnoosh.ghadiri.1@ulaval.ca, bergevin@gel.ulaval.ca
[2] LITIV lab., Polytechnique Montréal
gabilodeau@polymtl.ca

Abstract. We study the challenging problem of detecting carried objects (CO) in surveillance videos. For this purpose, we formulate CO detection in terms of determining a person's contour hypothesis and detecting CO by exploiting the remaining contours. A hypothesis mask for a person's contours is generated based on an ensemble of contour exemplars of humans with different standing and walking poses. Contours that are not falling in a person's contour hypothesis mask are considered as candidates for CO contours. Then, a region is assigned to each CO candidate contour using biased normalized cut and is scored by a weighted function of its overlap with the person's contour hypothesis mask and segmented foreground. To detect COs from obtained candidate regions, a non-maximum suppression method is applied to eliminate the low score candidates. We detect COs without protrusion assumption from a normal silhouette as well as without any prior information about the COs. Experimental results show that our method outperforms state-of-the-art methods even if we are using fewer assumptions.

Keywords: Carried object detection, codebook, biased normalized cut

1 Introduction

Detecting objects carried by people provides a basis for smart camera surveillance systems that aim to detect suspicious events such as exchanging bags, abandoning objects, or theft. However, the problem of detecting carried objects (CO) has not yet received the attention it deserves, mainly because of the inherent complexity of the task. This is a challenging problem because people can carry a variety of objects such as a handbag, a musical instrument, or even an unusual/dangerous item like an improvised explosive device. The difficulty is particularly pronounced when objects are small or partially visible.

Despite a lot of efforts in object detection, not much work has been done to detect COs. A successful approach such as deformable part model (DPM) [1] for object detection is not directly applicable to CO detection since COs may

This research was supported by FRQ-NT team grant No. 2014-PR-172083

Fig. 1: Three examples of persons with COs and their maximal response for a contour detector and corresponding segmentation by [5].

not be easily represented as a single deformable model or a collection of deformable parts. In addition, COs do not usually appear as regions enclosed by the contours or as compact regions with distinct gray-level or colour. This makes them difficult to segment (Fig. 1 illustrates this problem). There are few works that exploit appearance-based object detection approaches to detect COs. These approaches are mostly limited to the recognition of specific objects. Other approaches [2],[3],[4] use motion information of human gait to detect CO. To detect COs, motion of an average walking unencumbered person is modeled and those motion detections not fitting in the model are selected as COs. These approaches are usually based on the assumption that COs are sufficiently large to distort the spatio-temporal structure. To develop a more generic CO detector, prior information of human silhouette is used to help better discriminate between a person's region and a CO. To detect irregular parts in a human silhouette, some researchers [6],[7],[8] generate a generic model of a normal human silhouette and then subtracts it from a segmented foreground. The main assumption in these approaches is that COs alter a normal silhouette. This assumption limits these approaches to detect COs that are significantly protruding from normal silhouette and to miss those that are located inside it. Moreover, these approaches are highly dependent on the precise segmentation of foreground. Therefore, they usually cannot distinguish between COs and different types of clothes or imperfections of the segmented foreground if they all cause protrusions.

In this paper, we present a framework (sketched in Fig. 2) named Ensemble of Contour Exemplars (ECE) that combine high-level information from an exemplar-based person identification method with the low-level information of segmented regions to detect COs. A person's contour hypothesis that is learned from an ensemble of exemplars of humans is used to discriminate a person's contours from other contours in the image. We then use low-level cues such as color and texture to assign a region to each contour that does not belong to the person's contours. Each region is considered a candidate CO and is scored based on high-level information of foreground and person's contours hypothesis. Then, a non maximum suppression method is applied to each region to suppress any region that is not the maximum response in its neighborhood.

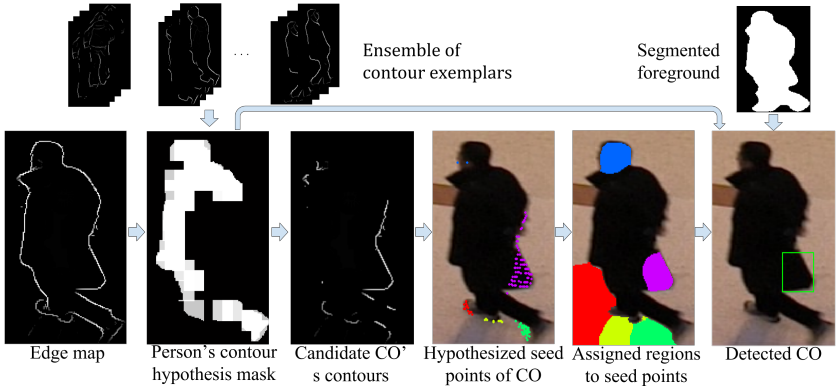


Fig. 2: An overview of our system (ECE).

Contributions: Our two main contributions are: (1) generating a person's contour hypothesis combined with low-level information cues to detect COs. Analyzing irregularity of a person's contours instead of human silhouettes enables our method to detect COs that are too small to alter normal human silhouette and those that are contained inside it; and (2) no prior knowledge of CO shape, location and motion is assumed. Having no assumption on the motion of the person enables our method to be applied on any single frame where a person appears instead of relying on short video sequences of a tracked person.

2 Related Work

Detecting COs can be formulated as an object detection problem. Object detection is often conducted by object proposal generation and then by classification. Zheng et al. [9] detected COs using contextual information extracted from a polar geometric structure. Extracted features are fed into a Support Vector Machine (SVM) classifier to detect two types of luggages (suitcases and bags). Considering only the appearance of COs leads to numerous false detections corresponding to the head, hands, feet, or just noise. Therefore, more works have focused on incorporating prior information about humans to facilitate the detection of the COs.

Branca et al. [10] detected pedestrians as well as two types of COs using a SVM classifier and wavelet features. When a pedestrian is localized in a frame, a sliding window with different sizes is applied around the pedestrian to find the CO. Instead of a pre-trained model for CO, Tavanai et al. [11] utilized geometric criteria (convexity and elongation) among contours to find COs in non-person region. A person's region is obtained by applying a person detector to obtain a bounding box followed by a color-based segmentation method. By assuming that COs are protruding from a window where a person is likely to occur, the two

largest segments that are obtained from the color-based segmentation method are considered as regions belonging to a person. Then, under the assumption that only a carry event is occurring, a set of detections by geometric shape models is refined by incorporating spatial relationships of probable COs with respect to the walking person.

Pedestrian motion can be modeled as made of two components: a periodic motion for a person's limbs and a uniform motion corresponding to the head and torso. Under the assumption that COs are held steadily, their motion can also be formulated as a uniform motion. Having this information helps the CO detector to search only regions with uniform motion. The main idea of [4] is that uniform motion of people carrying objects does not fit the average motion profile of unencumbered people. Pixels of moving objects with motion that do not fit the pre-trained motion model of people without CO are grouped as carried objects. In Dondera et al. [12] method, CO candidates are generated from protrusion, color contrast and occlusion boundary cues. Protruding regions from a person's body are obtained by a method similar to [4] to remove limbs and then generate a template of unencumbered pedestrian (urn-shaped model) with the aim of removing the head and torso. A segmentation-based color contrast detector and an occlusion boundary based moving blob detector are applied to detect other candidate COs. Each candidate region is characterized by its shape and its relation to a human silhouette (e.g. relative distance of centroid of person's silhouette to the object center) and classified using a SVM classifier as a CO or a non-CO.

The majority of works on CO detection have combined human motion cues with prior information about the silhouette of human to detect irregular parts in the human body such as the existence of COs. Chayanurak et al. [3] detected a CO using the time series of limbs motion. In their work, a star skeleton represents the human shape. Each limb of the star is analyzed through the time series of normalized limb positions. The limbs which are motionless or which are moving with the overall human body motion are detected as limbs related to the COs. Haritaoglu et al. [13] detected COs from a short video sequence of a pedestrian (typically lasting a few seconds) by assuming that unencumbered human shape is symmetric about its body axis. Asymmetric parts are grouped into connected components as candidate CO regions. Asymmetric regions that belong to the body parts are discriminated by periodicity analysis. The work of Damen et al. [7] is based on creating a temporal template of a moving person and subtracting an exemplar temporal template (ETT) from it. The ETT is generated offline from a 3D model of a walking person and is matched against the tracked person temporal template. Protruding regions from ETT are considered as likely to be COs if they are at expected locations of COs. Prior information about CO location is learned from the occurrence of COs in the ground truth temporal template. These information and protrusion cues are combined into a Markov Random Field (MRF) framework to segment COs. Tzanidou et al. [6] follow the steps of [7] method to detect COs but utilize instead color temporal templates.

In this work, we use prior information about the human body to build a normal human model. However, the main difference is that our method relies on the person's contours instead of his silhouette to detect irregularities with respect to the normal human model. We show that our human model can efficiently be used to find the regions that belong to COs.

3 Our Approach

The goal of our approach is to have a fully automatic system to detect COs on any frame where a person appears in the camera field of view. Using only one frame to detect COs makes the algorithm robust to events such as handling over a luggage, or a change in the person's direction. To detect COs, we build on two sources of information. The first is the output of the person's contours hypothesis generator. The second source of information is the output of a bottom-up object segmentation. Our contribution is to combine this information to discriminate between COs and other objects (person, background).

3.1 Building Human Models

To build human contour models and to detect COs, we first need to detect the moving regions corresponding to a person and the COs in a video. To accomplish this task, the DPM person detector [1] is applied on each frame. The intuition behind this is to find a person's location as well as obtaining a rough estimation of his height and width for further scale analysis. Since COs can protrude from the obtained person's bounding box, extracted foreground by a foreground extractor is used to find a second bounding box that bounds the person and the COs. The largest connected component of the extracted foreground that significantly overlap with the obtained person's bounding box is selected as our moving object target. In the rest of the paper, we will use the term moving object to refer to the person and the CO.

Learning an Ensemble of Contour Exemplars The output of a person detector is a window where a person is likely to occur. Thus this information is very coarse to discriminate the person's contours from other object contours. To have a class-specific contour detection, we follow [14] to generate a hypothesis mask for the person's contours. Our aim is to learn contours of humans dressed with various clothes with different standing or walking poses by building a codebook of local shapes. We label our training images into 8 classes corresponding to 8 possible walking directions of a person. Each class includes persons with different types of clothes and different walking poses. A foreground mask for each image is extracted by a foreground extractor. COs are removed manually if the foreground contains COs. Fig. 3 shows an example of exemplars in 8 categories.

Given a training image, a person is detected as described in the previous paragraph and is scaled so that his height and width are the same as a pre-defined

Fig. 3: Exemplars in different directions and poses. From top to bottom: exemplars, foreground mask and contour exemplars are shown respectively.

size. A foreground mask corresponding to a detected person is extracted and contours inside the mask are extracted by the method of [5]. The obtained contours are highly localized since the method uses multiple cues such as brightness, color and texture. However, this information is not adequate to discriminate among contours of a person, COs and background. Using information of the contours, the foreground and the person's bounding box, we build a codebook of local shapes. The foreground mask is sampled uniformly with sampling interval sm, and for each sample, the shape context (SC) feature is extracted from contours inside the foreground mask.

Each codebook entry $ce_i = (s_i^{ce}, d_i^{ce}, k_i, m_i)$ records four types of information of a sample i on the segmented foreground, where, s_i^{ce} is a Shape Context (SC) [15] feature, d_i^{ce} is a relative distance of each sample to the center of the person's bounding box, k_i is a class identification of an exemplar that i-th sample belongs to, and m_i is a patch of foreground mask with the center of sample i.

Using information of relative distance of each sample to the centroid of the person, redundant codebook entries can be removed. To this end, codebook entries with similar SC features are removed if their relative distances to the centroid of a person d_i^{ce} and d_j^{ce} are close enough to each other. The closeness of d_i^{ce} and d_j^{ce} is calculated as :

$$D_{ij} = exp(-\|d_i^{ce} - d_j^{ce}\|) \tag{1}$$

3.2 Carried Object Detection

Given a test video frame, a moving object mv corresponding to a person and his CO is detected as explained previously. The extracted moving object is scaled based on its size, as obtained from the person detector. Then, the foreground is sampled uniformly with sampling interval sm and SC feature is extracted

for each sample. Having a rough estimation of the person location by the person detector, a relative distance of samples to the centroid of person is obtained. Therefore, each sample t_i of the foreground in the test frame can be expressed by its SC feature and its relative position to the person's center as $t_i = (s_i^{mv}, d_i^{mv})$. Using this information, a hypothesis mask for the person is generated by classifying the person into one of eight classes and then generating a hypothesis based on the obtained class as described below. Our intuition behind the person's view classification is that a person's contours in one view can show similar characteristics to a CO contours in another view. Therefore, to detect COs, each person's contour should be compared with the contour exemplars with the same viewing direction.

Person's View Classification Each t_i is compared with a codebook entry, only if their relative distances to the centroid of a person d_i^{ce} are close enough to each other. The probability of matching sample t_i at location d_i^{mv} to a set of codebook entries ce_j is defined by Equation 2.

$$P(t_i|d_i^{mv}) = \sum_j exp(-\|s_j^{ce} - s_i^{mv}\|)P(d_i^{mv}|d_j^{ce}, \Sigma),$$
$$\text{Where:} \quad P(d_i^{mv}|d_j^{ce}, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} exp(-\frac{1}{2}(d_i^{mv} - d_j^{ce})^T \Sigma^{-1}(d_i^{mv} - d_j^{ce})) \tag{2}$$

where the 2×2 covariance matrix Σ is diagonal and $\Sigma_{11} < \Sigma_{22}$ to diminish the effect of error in the calculation of person's height with DPM. Note that all moving objects in test and training images are scaled so that the person's height and width are the same as a pre-defined size. Therefore, each s_i^{mv} compared with the one in the training set that is located in the same area as d_i^{mv}. If a match is found, the corresponding codebook entry will cast a vote to the class, which it belongs to. The class with the maximum number of votes is selected as the person's view class.

Hypothesis Generation Now, we can build a hypothesis mask of the person's contours by backtracking the matching results of the person's view class. From all codebook entries in the specific view ce_k that are matched to a t_i, we choose the one with maximum matching score and select its foreground patch m_j as hypothesis mask for t_i. Probability of the assigned patch $Patch_i$ to the sample t_i is calculated by Equation 3.

$$P(Patch_i|t_i) = \max_j exp(-\|s_j^{ce_k} - s_i^{mv}\|)P(d_i^{mv}|d_j^{ce_k}, \Sigma)m_j \tag{3}$$

We only keep the patches with probability higher than 0.8 to build a hypothesis mask. Fig. 4 shows two examples of hypothesis mask for a person's contours which the probability of each patch is between 0.8 and 1. With the information of the hypothesis mask H, we can now analyze the contours that do not fall inside this hypothesis mask H as candidate CO contours. To determine which candidate CO contours belong to each of the three categories (CO, person, background), the three following steps are applied to the candidate contours.

Fig. 4: Two examples of hypothesis mask for a person's contours (Hypothesis mask is shown by white blocks). Gray value expresses the probability.

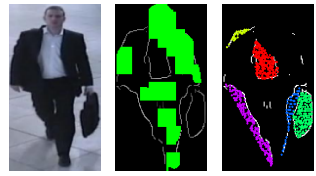


Fig. 5: An example of generating seed points. From left to right, original image, hypothesis mask and generated seed points.

Step 1: Seed Points Generation In this step, geometric information of a contour is used to obtain a rough estimation of the local shape of the object the contour belongs to. To accomplish this task, probable contours of COs are splitted at junction points. Each obtained contour is characterized by its curvature and the distance between its endpoints. We compute the curvature of a contour line by dividing its arc length to the the distance between its endpoints. Only high curvature contours are kept as more informative contours for further analysis. We use points located between a contour and the line joining its endpoints as seeds of a region to which the contour can be assigned to. To this end, each open contour is closed by connecting its two endpoints. Then, the enclosed area is uniformly sampled to generate the seeds. Fig. 5 shows the remaining contours obtained by subtracting hypothesis mask H^T and the associated seed points.

Step 2: Assigning a Region to a Set of Seed Points We formulate the problem of assigning a region R_j to a *i-th* contour of candidate CO contours, as an image segmentation problem. Here, we are looking for an image segment that has sufficient overlap with our pre-computed seed points. To this end, we apply biased normalized cut (BNC) by [16] to each object candidate. BNC starts by computing the K smallest eigenvectors of the normalized graph Laplacian $\mathscr{L}_G$ where the edge weight w_{ij} of the graph are obtained by the contour cue of section 3.1. Eigenvectors that are well correlated with our obtained seed points s_P are up-weighted using the following Equation:

$$w_i \leftarrow \frac{u_i^T D_G se}{\lambda_i - \gamma}, \text{ for } i = 2, ..., K \tag{4}$$

where $u_1, u_2, .., u_K$ is the eigenvectors graph laplacian $\mathscr{L}_G$ of corresponding to the K smallest eigenvalues $\lambda_1, \lambda_2, ..., \lambda_K$. D_G denotes the diagonal degree matrix of graph G. se is a seed vector and γ controls the amount of correlation. The BNC for each set of seed points se_j is the weighted combination of eigenvectors by the pre-computed weight w_i. Fig. 6 shows the result of applying BNC with different

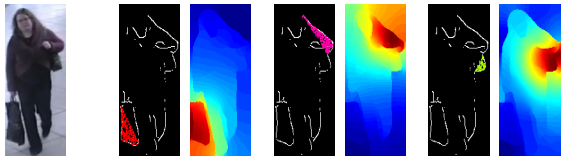


Fig. 6: Output of Biased Normalized cut for three sets of seed points with correlation parameter $\gamma = 0$.

seed points. The results of BNC for each set of seed points se_j is thresholded to segment region R_j.

Step 3: Non Maximal Suppression (NMS) For each region R_i, a score value V_i (calculated in Equation 5) is obtained based on overlapping ratio of the region with both complement of hypothesis mask H and foreground mask M.

$$V_i = (1+w)\frac{R_i \cap (1-H)}{R_i} + \frac{R_i \cap M}{R_i}, \quad \text{where:} \quad w = \sum_{k \in (R_i \cap H)}^{n} (1 - P(Patch_k|t_k))/n \tag{5}$$

We weight the overlapping ratio of complement of hypothesis mask and the region by multiplying it to the average probability of all samples $P(Patch_i|t_i)$ (calculated in Equation 3) in the intersection area. If region value V_i is lower than pre-defined threshold T then the region is rejected. Then a NMS method is applied to each region. In case of overlapping regions, only the one with the highest score V_i is accepted as a CO. The procedure to detect COs from the regions is formulated as follows:

Algorithm 1 Non-Maximal Suppression (NMS)

```
for each region R_i , i∈ N do
    Obtain V_i using Equation 5
    If V_j > T, and n is number of samples in the (R_i ∩ H)
    for each region R_j which is in the neighbor of R_i do
        Obtain V_j using Equation 5
        If V_j > T & V_j > V_i then Remove R_i
         Else Remove R_j end if
    end for
    Else Remove R_i end if
end for
```

4 Experimental Evaluation

The images for the training set are manually gathered from three different sources: PETS 2006 [3], i-Lids AVSS [4] and INRIA pedestrian [17] datasets. INRIA dataset is composed of still images and is only used in the training to complement frames from PETS and i-Lids. This way, we are able to keep more sequences of PETS and i-Lids for testing. In each image, a person is detected by DPM and its foreground is extracted automatically for PETS and i-Lids datasets, and manually for the INRIA dataset. Since our method is not too sensitive to the extracted foreground, we can use any foreground extractor in both testing and training steps. Here, we use a foreground extractor named PAWCS [18] for both PETS and i-Lids datasets. COs are removed manually from the obtained foreground. Then each person is labeled as one of 8 classes regarding the 8 possible viewpoints. For each class, an average of 15 persons (exemplars) are selected. Around 15 additional exemplars are obtained by horizontally flipping the previously selected ones.

We evaluate our algorithm on two publicly available datasets: PETS 2006 and i-Lids AVSS. For each dataset, COs are annotated with a ground truth bounding box. A detection is evaluated as true using the intersection over union criteria (IOU). That is, if the overlap between the bounding box of the detected object (b_d) and that of the groundtruth b_{gt} exceeds $k\%$ by the equation 6, the detection is considered a true positive (TP). Otherwise, it is considered a false positive (FP). Source code for CO detection and annotations for i-Lids dataset are available at https://sites.google.com/site/cosdetector/home.

$$overlap(b_d, b_{gt}) = \frac{b_d \cap b_{gt}}{b_d \cup b_{gt}} \tag{6}$$

4.1 PETS 2006

PETS 2006 contains 7 scenarios of varying difficulty filmed from multiple cameras. We selected 7 sequences of PETS 2006 that use the third camera. Eighty-three ground-truth bounding boxes of COs are provided online by Damen et al. [7] for 75 individuals among 106 pedestrians. Individuals that are not in the set provided by [7] are used in the training set. Since [7] relies on a short sequences of tracked person to detect COs, a tracked clip for each person is also provided. We detect moving objects on the first frame of each short video sequences of 75 pedestrians as described in section 3.1, and our CO detector is applied on the obtained moving object. Fig 7 shows the result of our method on PETS dataset. Our algorithm can detect a variety CO successfully. However, some body parts are detected, since they are not modeled by the exemplar.

To compare with [7] and [11] methods, we use the results presented in their papers with overlap threshold $k = 0.15$ as in [7]. This threshold value is much

[3] http://www.cvg.reading.ac.uk/PETS2006/data.html
[4] http://www.eecs.qmul.ac.uk/ andrea/avss2007_d.html

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

Fig. 7: Successes and failures of our approach on PETS 2006. First row: results after applying NMS on the segmented regions. Second row: bounding boxes (BB). Green BBs are TP detections, red and yellow BBs are FP detections. Yellow BBs are multiple detections of same CO. (a, e, h) failures because of poor person model for body's part, (b) failure because a clothe pattern is detected as an irregularity, (f) object is splitted since its edges are wrongly classified as the person's contours, (g) object is splitted in two regions because the small bag on the larger luggage is not detected.

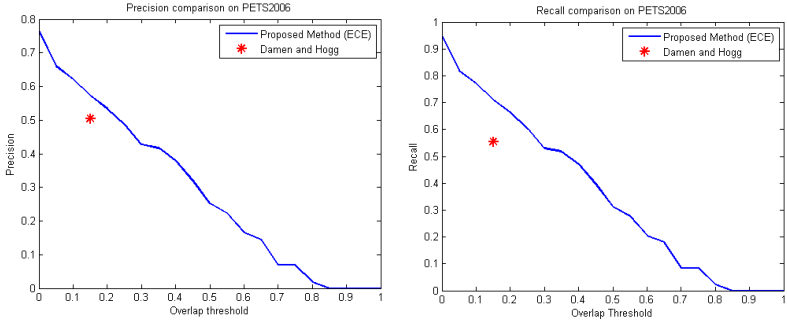
lower than typically used in object detection (0.5), since [7] only detects the parts of the object that protrude from the person's body. The comparison shows (see Table 1) that we achieve a higher detection rate and a slightly better FP rate compared to [7]. Comparing our method to [7] and [11] in terms of F1 score, we can see that that our method outperforms them by about 10%. It should be noted that both [7] and [11] use the whole sequence to detect COs while we only use the first frame of the whole sequence and still obtain better results.

	Prec.	Rec.	TP	FP	FN	F1 Score
Proposed Method (ECE)	**57**%	**71**%	**59**	**44**	**24**	**63**%
Damen et al. [7]	50%	55%	46	45	37	52%
Tavanai et al.*[11]	-	-	-	-	-	≈ 53%

* Estimated from the F1 score plot in their paper.

Table 1: Comparison using PETS 2006 with a 0.15 overlap threshold.

Using an overlap threshold of 0.15 may not show the real performance of a CO detector, since it can detect large parts of a person's body as a CO and still have a high score because the required overlap for good detection is too small. For thoroughness and to give a better idea of the performance of our method, we



(a) (b)

Fig. 8: Precision and recall plots as function of the overlap threshold on PETS 2006.

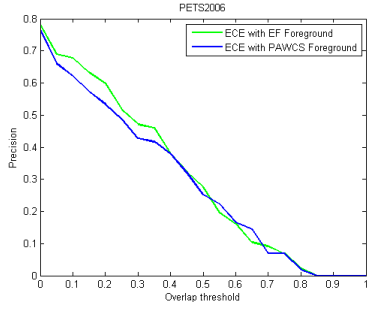


Fig. 9: Comparison of precision with two different foreground extractors.

depict precision and recall of our algorithm as the threshold of overlap is varied in Fig. 8.

We also explore the effects of foreground extraction in terms of detection performance. Fig. 9 shows the results of our method with two different foreground extractors: Based on a simple thresholding on a results of optical flow by [19] and based on background subtraction with PAWCS [18]. The results show that our algorithm is not too sensitive to the extracted foreground. This robustness to the extracted foreground comes from the fact that we assign a region to each contour and analyze the region by the amount of overlap with the extracted foreground. Although, extracting the foreground with [19] has slightly improved our results on PETS, it does not occur in general cases. In this case, some parts of the foreground where abrupt movement exist such as a person's limbs are missing. These errors are surprisingly beneficial in some scenarios by reducing the number of false positives, which however increases the number of false negatives in other cases.

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

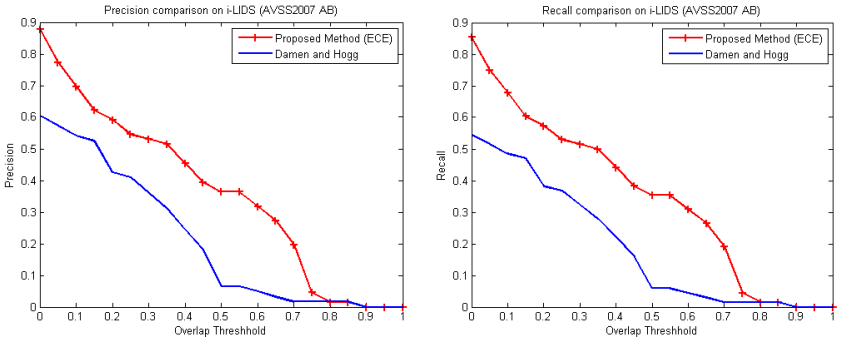
Fig. 10: Successes and failures of our approach (Right) compared with [7] (Left) on i-Lids AVSS. Bottom rows: detected bounding boxes (BB), Top rows segmented objects. Green BBs are TP detections and red BBs are FP detections.

4.2 i-Lids AVSS

Since all parameters (SC size, *sm*, T) are only dependent on the person's scale and all detected pedestrians are scaled to a pre-defined window size (as described in section 3.1) our algorithm can be testetrolld on other datasets with the same parameters used for PETS. i-Lids AVSS 2007 consists of both indoor and outdoor surveillance videos. We use three videos recorded at a train station. Fifty-nine individuals among 88 are selected for the test, and their 68 COs are manually annotated. Individuals that are not in the test set are used for the traning set. COs in this dataset are varied and include document holders, handbags, briefcases, and trolleys. Again, we compared our method with the state of the art method of Damen et al. [7], who are providing their code online. To apply [7] on i-Lids dataset, we prepared short video sequences of our selected individuals to create spatio-temporal template. Furthermore, in each frame, the person is detected manually and its foreground is obtained using PAWCS method. Since, [7] is sensitive to the extracted foreground, we only apply PAWCS to detect more accurate foreground mask. Viewing direction of a person is selected manually, as calibration data are not provided with this dataset. Detected COs on the temporal templates are projected onto the first frame of the sequence.

Fig. 10 shows the results of our method (ECE) compared with [7]. It can be seen that our method can detect COs more successfully, and the boundaries of the COs are better delimited. Fig. 10 (a-b) shows the ability of our algorithm to detect objects with less protrusion or contained inside the person's body area. Fig. 10 (c,g) shows failure cases as result of poor person model for the person's clothes and body parts respectively. Fig. 10 (d,e) shows two false negative (FN) cases as they are both identified as part of the person's clothes.

Table 2 shows the results of our method and [7] on i-Lids Dataset with overlap threshold ($k = 0.15$). Although, we achieved better results compare to



(a) (b)

Fig. 11: Precision and recall plots as function of the overlap threshold on i-Lids.

[7], as discussed previously, $k = 0.15$ is very low to show the real performance of the system. As shown in Fig. 10, a large detected part of a person's body that contains a CO is counted as TP with $k = 0.15$. To view the complete picture, we plot the precision and recall of our algorithm and [7] with different overlap thresholds (Fig. 11). Fig. 11 justifies the results of Fig. 10 as it shows that our algorithm achieves better performance with all overlap thresholds.

	Prec.	Rec.	TP	FP	FN
Proposed Method(ECE)	**62%**	**60%**	**41**	**25**	**27**
Damen et al. [7]	52%	47%	32	29	36

Table 2: Comparison of [7] with the proposed method over i-Lids AVSS.

5 Conclusion

We presented a framework for detecting COs in surveillance videos that integrates both local and global shape cues. Several models of a normal person's contours are learned to build an ensemble of contour exemplars of humans. Irregularity in a normal human model is detected as COs. Our experiments indicate that learning human model from human's contours makes the system more robust to the factors that may give rise to irregularities such as clothing, than methods that model humans based on silhouettes [7]. Using biased normalized cut to segment object combined with the high-level information of human model, provides us with a rough estimation of the CO shape. Our method can have a better estimation of CO shape than [7], and it can be used for future analysis such as recognition of the object type.

References

1. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1627–1645
2. Senst, T., Kuhn, A., Theisel, H., Sikora, T.: Detecting people carrying objects utilizing lagrangian dynamics. In: Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. (Sept 2012) 398–403
3. Chayanurak, R., Cooharojananone, N., Satoh, S., Lipikorn, R.: Carried object detection using star skeleton with adaptive centroid and time series graph. In: Signal Processing (ICSP), 2010 IEEE 10th International Conference on. (Oct 2010) 736–739
4. Senst, T., Evangelio, R., Sikora, T.: Detecting people carrying objects based on an optical flow motion model. In: Applications of Computer Vision (WACV), 2011 IEEE Workshop on. (Jan 2011) 301–306
5. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(5) (May 2011) 898–916
6. Tzanidou, G., Zafar, I., Edirisinghe, E.: Carried object detection in videos using color information. Information Forensics and Security, IEEE Transactions on **8**(10) (Oct 2013) 1620–1631
7. Damen, D., Hogg, D.: Detecting carried objects from sequences of walking pedestrians. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(6) (June 2012) 1056–1067
8. Mitzel, D., Leibe, B.: Taking mobile multi-object tracking to the next level: People, unknown objects, and carried items. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part V. ECCV'12, Berlin, Heidelberg, Springer-Verlag (2012) 566–579
9. Zheng, W.S., Gong, S., Xiang, T.: Quantifying contextual information for object detection. In: Computer Vision, 2009 IEEE 12th International Conference on. (Sept 2009) 932–939
10. Branca, A., Leo, M., Attolico, G., Distante, A.: Detection of objects carried by people. In: Image Processing. 2002. Proceedings. 2002 International Conference on. Volume 3. (2002) III–317–III–320 vol.3
11. Tavanai, A., Sridhar, M., Gu, F., Cohn, A., Hogg, D.: Carried object detection and tracking using geometric shape models and spatio-temporal consistency. In Chen, M., Leibe, B., Neumann, B., eds.: Computer Vision Systems. Volume 7963 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 223–233
12. Dondera, R., Morariu, V., Davis, L.: Learning to detect carried objects with minimal supervision. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on. (June 2013) 759–766
13. Haritaoglu, I., Cutler, R., Harwood, D., Davis, L.: Backpack: detection of people carrying objects using silhouettes. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. Volume 1. (1999) 102–107 vol.1
14. Wang, L., Shi, J., Song, G., Shen, I.F.: Object detection combining recognition and segmentation. In: Proceedings of the 8th Asian Conference on Computer Vision - Volume Part I. ACCV'07, Berlin, Heidelberg, Springer-Verlag (2007) 189–199
15. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(4) (Apr 2002) 509–522

16. Maji, S., Vishnoi, N., Malik, J.: Biased normalized cuts. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. (June 2011) 2057–2064
17. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1. (June 2005) 886–893 vol. 1
18. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: A self-adjusting approach to change detection based on background word consensus. In: Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. (Jan 2015) 990–997
19. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In: Computer Vision and Pattern Recognition. (2015)