Spatio-Temporal Consistency to Detect and Segment Carried Objects

Farnoosh Ghadiri[1]
farnoosh.ghadiri.1@ulaval.ca
Robert Bergevin[1]
bergevin@gel.ulaval.ca
Guillaume-Alexandre Bilodeau[2]
gabilodeau@polymtl.ca

[1] LVSN-REPARTI
Université Laval
Québec, Canada

[2] LITIV lab
Polytechnique Montréal
Montréal, Canada

Abstract

We present a new method to detect carried objects and to segment them accurately after detection. The proposed method includes several contributions: first, a new superpixel-based descriptor is proposed to identify carried object-like candidate regions using human shape modelling. Second, integrating spatio-temporal information of candidate regions to detect carried objects. We exploit the consistency of recurring carried object candidates viewed over time to detect the final carried object locations based on their motion and location priors. Last, the detected carried object regions are accurately segmented. Compared to existing methods, our approach is not only focusing on detecting carried objects. It takes a step forward and accurately segment them. Our method to carried object segmentation couples local appearance cues with location priors of the detected carried objects to produce accurate segmentation. Experimental evaluation on two datasets demonstrates that both our carried object detection and segmentation methods significantly outperform competing algorithms.

1 Introduction

In recent years, surveillance cameras have become ubiquitous. The rapid growth of the video surveillance market increased the need for automatically understanding information from videos. In the domain of video analysis, automatically analysing human activities involving carried objects attracts a lot of attention, and automatically detecting carried objects is at the heart of this field.

Detecting objects carried by people is a difficult task due to the significant variation in types of carried objects with regards to their shape, size and color. They also often blend with the person's clothing e.g. a handbag carried by a person can be easily confused with a part of a person's clothes.

There are three main schools of thought to detect carried object (CO), namely (i) analysing the human shape to detect anomalies [1, 2, 3, 4], (ii) learning models for each type of carried object [5, 6, 7] and (iii) analysing difference in human gait and motion patterns [1, 8, 9]. Learning a model for each object type is quite challenging due to the variety of object type carried by a person. This approach is usually application specific. Analysing human shape is

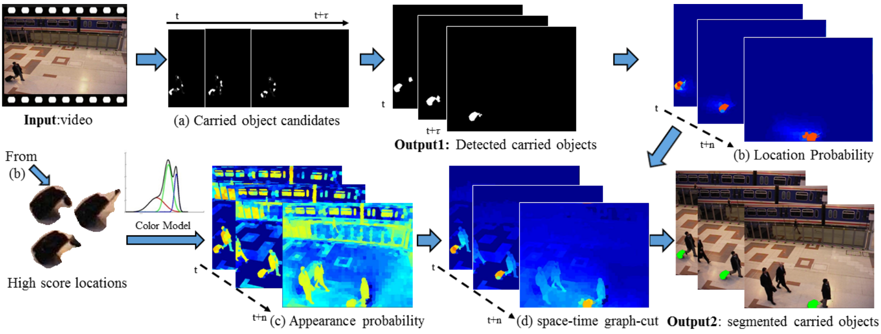
Input:video
t
t+τ
(a) Carried object candidates
t
t+τ
Output1: Detected carried objects
t
t+n
(b) Location Probability
From
(b)
High score locations
Color Model
t
t+n
(c) Appearance probability
t
t+n
(d) space-time graph-cut
Output2: segmented carried objects

Figure 1: Illustration of our spatio-temporal CO detection and segmentation method.

also not straightforward because of different clothing styles and physical differences between people. Most of the approaches in this area make the problem of detecting COs simplified by considering only unoccluded large objects for which their relative distance to the center of a person is not changing [2, 10]. The method called "ECE" by Ghadiri et al. [3] improved significantly on this aspect by diminishing the number of assumptions for CO detection. They detect COs by analysing person contours and explaining the anomalies in terms of COs in the still video frame. Their work outperforms state-of-the-art methods without any assumption on the protrusion and object size and shape. We build over this method by extending their approach to the spatio-temporal domain in order to better discriminate of CO regions from background and a person's regions.

In this paper, we explore how we can detect and segment carried objects in a video sequence. The main idea is to use both static and dynamic cues to detect persistent carried object-like regions and then leverage both the appearance and the location of these regions to segment COs. To implement this idea, we first introduce a superpixel- based descriptor to extract Shape Context (SC) features from every superpixel in the over-segmented frame. We then score each superpixel with regards to its similarity to a stored codebook of local human shapes. The complement of the score value of each superpixel within the foreground is considered as the likelihood of a superpixel to belong to a CO (Fig. 1(a)). To capture motion and persistence of each superpixel candidate, we integrate information over all frames (Fig. 1 (output 1)). Intuitively, a candidate region that appears frequently throughout the video will more likely be a CO. Conversely, one that only occasionally occurs is more likely to be an un-interesting, person or background region. Using this spatial information of COs in each frame, an appearance model is built and COs are segmented using Graph-Cuts on a space-time Markov Random Field (MRF) (See Fig. 1).

Our main contribution is an automatic approach to segment carried objects on every frame of a surveillance video. While the previous methods focused more on the detection of COs, we take a step forward to segment carried objects on every frame of the video. Toward this goal, important novel components of our method include (1) a new superpixel-based descriptor to generate a carried object candidate regions, (2) integrating spatio-temporal information of candidate COs to detect carried objects. In terms of using temporal information, our method differs from the work of [2, 10], since any assumption on spatial position of COs relative to the body and the walking direction of people is not necessary, and (3) a space-time graph segmentation that refines the detection results to segment COs. We applied our method on two publicity available datasets: PETS 2006 and i-Lids AVSS, and show state-of-the-arts results compared to two recent methods. The source code and data of our method is available at https://sites.google.com/site/cosegmentation/.

2 Background

Since CO types can vary with regards to their shapes and sizes, the most successful approaches in this area focused on analysing human silhouettes followed by analysing their motion patterns to differentiate between human regions and CO regions. Haritaoglu et al. [4] and Javed et al. [11] detected COs by analysing the periodicity of non-symmetric regions in a silhouette. Both works assumes that an unencumbered person silhouette is symmetric about the body axis. Therefore, asymmetric regions with motion patterns that are different from the limbs are considered as COs. Lee et al. [12] detected COs as outliers in the extracted foreground using a dynamic shape model of human motion. Chayanurak et al. [1] proposed a star skeleton model for a person silhouette and detected COs by analysing a time series of motions of extracted skeleton limbs. Damen and Hogg [2] subtracted the best matching temporal template exemplar from a generated temporal template of a tracked person. Remaining regions from the subtraction which are static and fit in a spatial prior map are labeled as COs. This previous work is further developed by Tzanidou et al. [13] by introducing a color temporal template. Tavanai et al. [10] proposed a convex shape model and an elongated shape model for COs. In their work, an estimation of a person region is obtained by the output of a person detector and the search for convex/elongated regions is limited to the area of the non-person regions.

The efficiency of the previous methods depends heavily on the precision of the foreground extractor, the CO location and the assumption that COs are significantly protruding from a person's body. Dondera et al. [14] integrated different descriptors to segment moving object into regions based on their color and motion. Then, they characterized each region by a set of features, and used these features as input to a classifier. The weakness of this work is a high similarity of CO shapes, and shapes of a person's body parts negatively affect the results. Ghadiri et al. [3] did not make any specific assumptions about the COs shapes and their location. They proposed a method based on analysing a person contours to alleviate the previously stated problems. In their work, an extracted edge map of a moving object is compared to an ensemble of contour exemplars. Contours that are less probable to be a member of the person class, are further analysed by assigning a region to each contour to differentiate between a person's region and a CO region.

3 Proposed Method

Our proposed method is inspired by [3] but extends it and outperforms it considerably since: 1) it can detect COs more precisely by integrating static and dynamic cues, and 2) it more accurately segments COs by minimizing an energy function based on a learned dynamic appearance model of detected regions and their location priors. The basic principles behind the design of the algorithm are the following:

- Superpixel as a segmentation unit: Contours convey key information about object shapes. Analysing local shape information of contours can benefit from superpixels by gathering contour information in perceptually meaningful atomic regions.

- Spatio-temporal consistency: Under the assumption that COs are not appearing just in a single frame of a video, we exploit the consistency of recurring CO candidates viewed over time to boost the detection results on each frame.

- Dynamic model of carried object and non-carried object: using initial detection results over time, we can model CO and non-CO regions in terms of their appearance and their location in the form of an energy function to segment carried objects.

Based on these principles, our algorithm consists of the three main stages; (1) initial CO location estimation, (2) CO/non-CO labeling refinement, and (3) accurate CO segmentation.

3.1 Initial carried object location estimation

Our approach starts by building a codebook of local features extracted from a training set. The goal of this stage is to produce an initial estimate of CO regions by comparing local features of a person possibly carrying an object against a codebook consisting of persons that do not carry objects. We follow the same steps as ECE [3] to produce CO candidates, but we exploit superpixel information to generate better CO candidates. To achieve this, similarly to ECE [3], we first model walking/standing people from different viewpoints by constructing a codebook of local features with their distance to the center of the person on the training images. Unlike ECE which samples each training image using a regular grid to extract Shape Context (SC) features, we over-segment the training images into superpixels and then extract SC descriptors on the center of every superpixels. This provides more meaningful local features than features extracted from a regular grid in case of deformable objects like a person. Therefore, each codeword $c_j = (sc^c_j, d^c_j, v^c_j)$ records three types of information of a superpixel s_j, that is the shape context feature sc^c_j, the relative distance d^c_j of the center of the superpixel s_j to the centroid of the person bounding box, and the discretized viewpoint of the person $v^c_j \in \{1,2,..,8\}$. We will now describe how we detect COs using this codebook.

In contrast to ECE that detects carried objects on a still frame, our method needs a short video sequence (SVS) of approximately 2 walking cycles (τ frames) of a person trajectory to detect COs. The temporal information of the short video sequence will be added to the system in the detection phase. At this stage the SVS is used to generate CO candidates in each frame independently.

Given a SVS, a region of interest (ROI) is tracked by the tracker proposed by Jia et al. [15] and its foreground extracted by Papazoglou et al. [16] method. The ROI is a bounding box that contains a person and potentially one or more COs. An estimation of the scale and the center of the person in the ROI is obtained by a window-based detector [17] in the first frame of the video, like in [3]. Then, the person scale and center in the subsequent frames are calculated by the changes in the size of the tracking bounding box and its displacement compared to the first frame.

Having this information, SC features are extracted from each superpixel s^t_i in the over-segmented tracking bounding box (Fig. 2 (c)). Each superpixel is then matched against the codebook and the matching score for s^t_i is computed as follow:

$$\begin{aligned} P(s^t_i|d^t_i) &= exp(-\|sc^c_j - sc^t_i\|)P(d^t_i|d^c_j,\Sigma) \\ \text{with,}\quad P(d^t_i|d^c_j,\Sigma) &= \frac{1}{2\pi\sqrt{|\Sigma|}}exp(-\frac{1}{2}(d^t_i - d^c_j)^T\Sigma^{-1}(d^t_i - d^c_j)) \end{aligned} \tag{1}$$

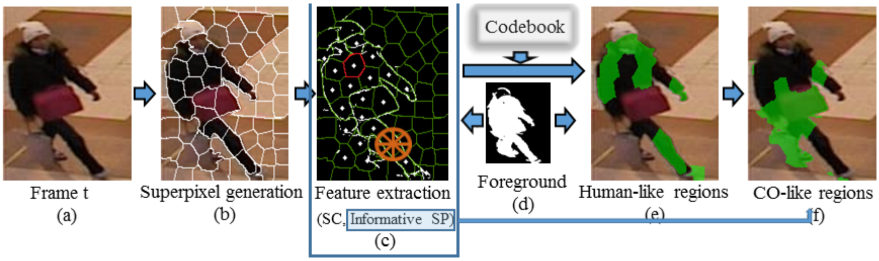

Figure 2: Overview of the initial carried object location estimation.

where d_i^t (d_j^c) is the relative distance of the s_i^t (s_j) to the center of a person bounding box in the test frame t (codeword). Σ is a 2×2 diagonal covariance matrix with $\Sigma_{11} < \Sigma_{22}$. Among the matching sc_j^c, their most frequent person viewpoint v_i in the codebook is chosen as the observation viewpoint. Then, the final probability for each superpixel is obtained by :

$$P(s_i^t|v_i) = \max_j exp(-\|sc_j^{Vc} - sc_i^t\|)P(d_i^t|d_j^{Vc}, \Sigma) \tag{2}$$

where Vc is a subset of the codebook with the specific viewpoint v_i. The probability of each superpixel reflects the likelihood of a superpixel to belong to a human region. Since we consider only two classes (person or CO) within the foreground region, the complement of the person probability map (Fig. 2(e)) is the CO probability map as shown in Figures 2(f) and 1(a)).

In the CO probability map, uninformative superpixels are excluded. We define a superpixel as informative if its boundaries are sufficiently supported by edges. A measure of the informativeness is obtained by calculating the number of overlaps between the boundary pixels of a superpixel and the edges in the edge map, divided by the perimeter of the superpixel. Fig. 2 (c) shows an uninformative superpixel with red boundary (it is not considered in the final decision, Fig. 2(f)).

3.2 Carried object/Non-carried object labeling refinement

Consider Fig. 3(b), the initial probability map for the COs based on one frame is not accurate enough because of ambiguities in human hypotheses and unreliable foreground. A superpixel can belong to a person or a shadow but still have a high value in the probability map. Nevertheless, this probability map provides initial hypotheses for the CO detection based on spatial and temporal information. Based on the assumption that carried objects are appearing in more than K percent of the frames in a SVS, we propose a method to accumulate the initial CO probability over the entire short video sequence to build a voting map for each superpixel at each frame.

Probability value V_j^t of each superpixel s_i^t is updated by a two-step algorithm. In the first step, starting from frame $t+1$ toward the last frame $t+\tau$ in SVS, each probability value V_j^t at frame t is refined by propagation with the information of the probability values of superpixels in previous frame $t-1$ and temporal connectivity using

$$V_j^t = V_j^t + \frac{\sum_i \phi(s_i^{t-1}, s_j^t).\psi(s_i^{t-1}).V_j^{t-1}}{\sum_i \phi(s_i^{t-1}, s_j^t)} \tag{3}$$

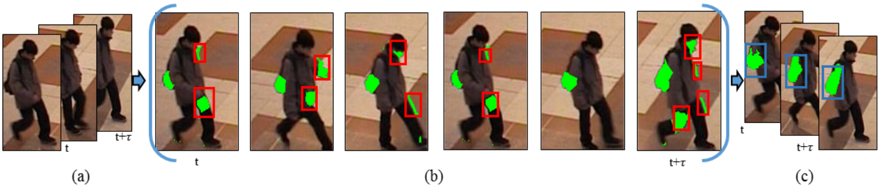

Figure 3: Carried object detection. (a) Input video (frame t to $t+\tau$). (b) CO candidate regions (red bounding boxes show the spurious candidates). For clarity only highly probable CO regions are shown in the figures (green segmented regions). (c) Detected COs delineated by blue bounding boxes.

where $\phi()$ is the percentage of pixels in superpixel s_i^{t-1} that connect to superpixel s_j^t by temporal connections as defined by the optical flow displacement in [18] ;$\psi()$ is a quality measure of optical flow transformation defined similarly to [16] by

$$\psi(s_i^{t-1}) = exp(-\lambda^{\psi} \sum_{p \in s_i^t} \| \bigtriangledown \overrightarrow{f_p} \|) \tag{4}$$

where $\bigtriangledown \overrightarrow{f_p}$ is the optical flow gradient in s_i^t and λ is a constant. In the next step, we redo the propagation from the last frame toward the first frame by equation 3. The final probability value for each superpixel is the sum of the two steps over the number of frames. In each frame, superpixels with a probability over $K/100 * max(V_j^t)$ are chosen as belonging to carried object (Fig. 3 (c)).

3.3 Accurate carried object segmentation

Once the CO detections are obtained in the SVS of walking pedestrian, the segmentation results are obtained by a graph-based method to get per-pixel CO segmentation results in the entire video. To this end, we propagate the detection results in the SVS (frame t to $t+\tau$) to the whole video from the first frame to the last frame $t+n$ and also in the reverse order using Equation 3. Then results are normalized and used for location priors for COs in each frame Fig. 1 (b).

Having the location priors for COs in the entire video sequence, we formulate accurate CO segmentation as a superpixel labeling problem with two labels (CO and non-CO). To evaluate a labeling, we define an energy function similar to [16, 19, 20]. It is given by

$$E(l) = \sum_{t,i} U_i^t(l_i^t) + \gamma \sum_{(i,j,t) \in N} W_{ij}^t(l_i^t, l_j^{t+1}) \tag{5}$$

where l is the labeling of a superpixel that can take values $l_i^t \in 0,1$ with 1 corresponding to a CO and 0 corresponding to a non-CO. N consists of four spatially and two temporally connected superpixels. The neighborhood term W^t encourages spatial and temporal smoothness and is a standard contrast-modulated Potts potential as defined in [16, 19, 21]. The data term U^t defines the cost of labeling superpixel s_i^t with label l_i:

$$U_i^t(l_i^t) = -log(A_i^t(l_i^t, d) + wV_i^t(l_i^t)) \tag{6}$$

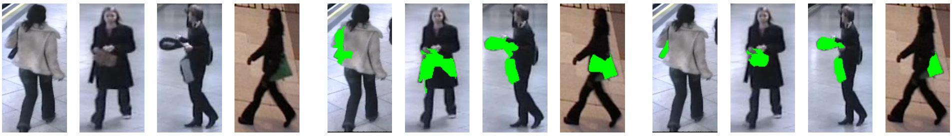

Figure 4: Comparison of the CO detection to our CO segmentation. (First column) original image. (Second column) initial CO detection. (Third column) refined CO segmentation.

V^t encourages CO labeling in the area obtained as CO detection (§ 3.2). The scalar w weights the unary potential V^t. The first term A^t is the color-induced cost, evaluating how likely a superpixel is to be CO or non-CO according to the appearance model of a CO, which is explained next.

Appearance model A^t. Using CO location priors obtained by propagating detected CO object to the entire video sequence (Fig. 1 (b)), we model CO and non-CO appearance by estimating two Gaussian Mixture Models (GMM) in the RGB colorspace: (1) a GMM for the superpixels belonging to the CO and (2) a GMM for the superpixels not belonging to COs. We define A^t as the probability of a superpixel RGB color computed from each GMM. A superpixel that has similar appearance in terms of its color to the CO (non-CO) will have high cost if labeled as non-CO (CO) (see Fig. 1 (c)).

Fig. 4 shows the refined segmentation results in comparison to the results obtained by our spatio-temporal detector in§ 3.2. As it can be seen from the second column in Fig. 4, some parts of the clothes are initially detected as COs. Building an appearance model of COs and non-COs regions combined with spatial information of COs in the form of an energy function, leads to more precise segmentation by separating better clothes from COs.

4 Experimental Evaluation

Implementation Details. Superpixel segmentation is carried out with Simple Linear Iterative Clustering (SLIC) [22] to independently segment each video frame into $N = 1000$ superpixels. SC features are extracted by the work of Wang et al. [23] and the size of the SC feature is chosen based on the superpixel scale. All parameters are constant for all experiments.

Experimental setup. We evaluate our method (named STCOD, for Spatio-Temporal Carried Object Detection) on two datasets: PETS2006 [24] and i-Lids AVSS [25]. These datasets were originally introduced to detect abandoned objects in a train station. For 75 individuals in the PETS 2006, Damen and Hogg [2] provides video sequences of the tracked individuals along with 83 CO ground-truth bounding box. Ghadiri et al. [3] provides ground-truth of 68 COs for 59 individuals in i-Lids AVSS in the first frame where a person appears in the video. To complete this ground-truth, we manually created short video sequences for each of the 59 individuals in i-Lids AVSS. For each tracked person in the video, CO ground-truths bounding box are obtained by tracking the ground-truth of the COs in the first frame as provided by [3].

To evaluate our method, we will compare it with Damen and Hogg [2] and ECE [3] and report their results as originally stated in [3]. It should be noted that these methods, including ours, all take a single decision per video sequence. However, ECE uses only the

	Prec.	Rec.	TP	FP	FN
STCOD	**67%**	**72%**	**60**	**29**	**23**
ECE [3]	57%	71%	59	44	24
Damen and Hogg [2]	50%	55%	46	45	37

Table 1: Comparison of our method with ECE [3], Damen and Hogg [2] over PETS 2006 with a $IoU = 0.15$ overlap threshold.

	Prec.	Rec.	TP	FP	FN
STCOD	**73%**	**69%**	**47**	**17**	**21**
ECE [3]	62%	60%	41	25	27
Damen and Hogg [2]	52%	47%	32	29	36

Table 2: Comparison of our method with ECE [3], Damen and Hogg [2] over i-Lids AVSS with a $IoU = 0.15$ overlap threshold.

first frame of the whole sequence, Damen and Hogg uses the whole video sequence and we only use a short video sequence of about 2 sec to detect COs. We quantify the performance with the percentage of CO ground-truth bounding boxes which are correctly localized as in [2] (intersection-over-union (IoU)> 0.15). Since the method of Damen and Hogg [2] only detects objects protruding from a person body, the threshold value 0.15 that they standardized for CO detection is much lower than what is typically used in object detection (50%).

We also evaluate our refined segmentation results. CO ground-truth bounding boxes are loosely delineating the COs. Therefore, they cannot efficiently be used to evaluate the segmentation results. To quantify segmentation accurately, we manually segmented the first frame of all tracked individuals in both datasets. For all 134 individuals, we evaluate the performance with the number of wrongly label pixels over the all frames in the segmentation datasets (Segmentation Error (SE) [18]). We also the evaluate accuracy of our segmentation with the proportion of correctly labelled pixels [26].

Results. As shown in Tables 1 and 2, our method substantially improves over the ECE and [2] on the precision. Our method significantly reduces the number of False Positive (FP) because it uses spatio-temporal information as described in § 3.2. In comparison to ECE, we also have a slight improvement in the number of True Positives (TP) because of a better CO candidate regions provided by our new superpixel-based descriptor and also using information of more than one frame. Fig. 5 depicts precision, recall and F1-score of our algorithm as the threshold of overlap (IoU) is varied. Results on both datasets show that our method achieves higher precision and recall as compared to other methods when the threshold is increasing.

As Table 3 shows, our method considerably outperforms the segmentation results of ECE method over the datasets as it tightly integrates appearance along with spatial consistency as segmentation cues. Since, the method of Damen and Hogg cannot segment carried objects, we excluded it from our comparison. Fig. 6 shows example frames from two datasets. Our method accurately segments all COs in Fig. 6 except Fig. 6 (k,d), as it misses two COs in both frames. The failure in the segmentation is primarily due to lack of spatial information provided by our CO detector. On the other hand, as explained earlier in § 3.3, in some cases that our CO detector fails to correctly label non-CO regions (i.e. provides inaccurate spatial

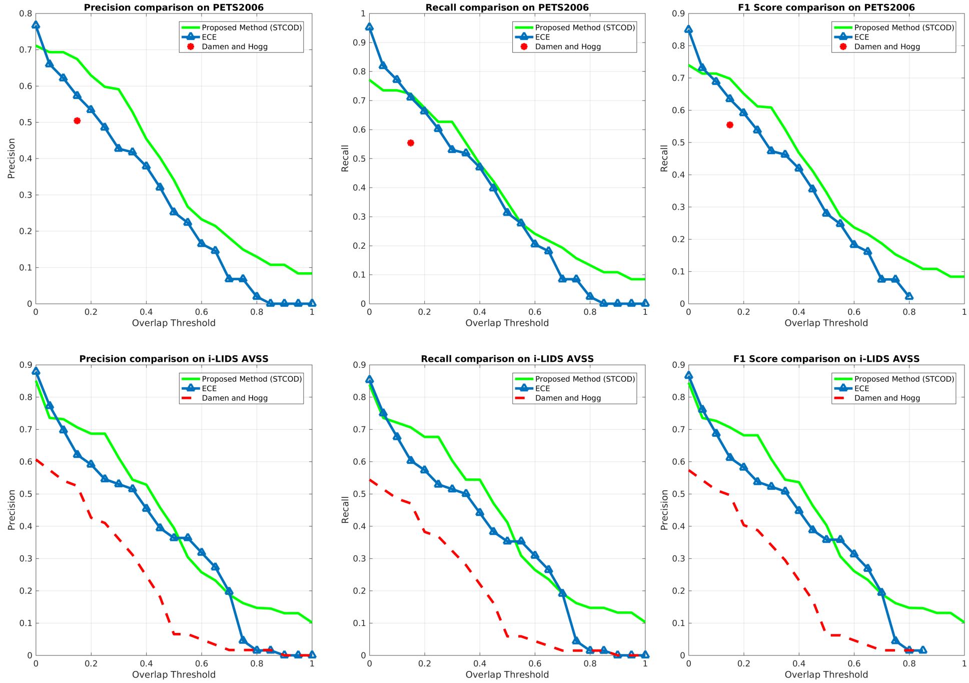

Figure 5: Precision, Recall and F1-score plots as function of the overlap threshold.

	Accuracy	Segmentation Errors (SE)
STCOD	**58%**	**1721**
ECE [3]	49 %	2061

Table 3: Comparison of our segmentation method with the ECE method.

information), our segmentation method is able to minimize those wrongly labeled regions to the regions which share less color similarity with non-CO regions (See Figures 6 (a,i) and 4).

5 Conclusion

In this paper, we make three contributions. First, we propose a new method to generate candidates for carried objects that exploit local cues obtained by the shape context descriptor and global information of a human shape. Second, we integrate the set of carried object candidates in the short video sequence to the spatio-temporal domain. Our experimental evaluation demonstrates that our spatio-temporal detector diminishes the limitations of previous methods, like ECE that often fail to discriminate clothes and shadows from COs. Last, we segment carried objects by optimizing an energy function, which includes appearance cues and location priors obtained by detected CO regions. Evaluation results on the segmentation method demonstrate a significant improvement of the segmentation.

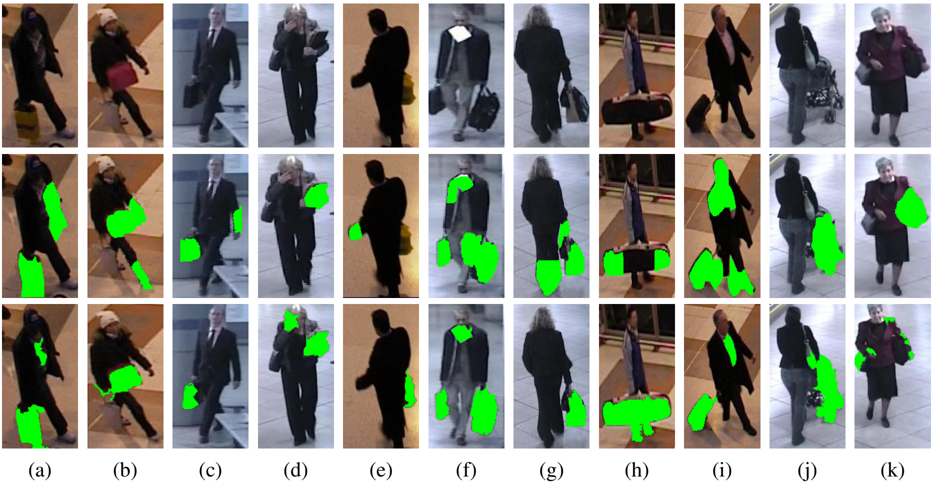

Figure 6: Comparison of our segmentation results to segmentation results of ECE [3]. First row: first frame of several tracked individuals from i-Lids AVSS and PETS 2006. Second row: segmentation results of ECE method. Third row: segmentation results of our method.

References

[1] R. Chayanurak, N. Cooharojananone, S. Satoh, and R. Lipikorn, "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, pp. 736–739.

[2] D. Damen and D. Hogg, "Detecting carried objects from sequences of walking pedestrians," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 6, pp. 1056–1067, June 2012.

[3] F. Ghadiri, R. Bergevin, and G.-A. Bilodeau, "Carried object detection based on an ensemble of contour exemplars," in *Proceedings of the 14th European Conference on Computer Vision-Part VII*. Springer International Publishing, 2016, pp. 852–866.

[4] I. Haritaoglu, R. Cutler, D. Harwood, and L. Davis, "Backpack: detection of people carrying objects using silhouettes," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, 1999, pp. 102–107 vol.1.

[5] A. Branca, M. Leo, G. Attolico, and A. Distante, "Detection of objects carried by people," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 3, 2002, pp. III–317–III–320 vol.3.

[6] T. W. Chua, K. Leman, H. L. Wang, N. T. Pham, R. Chang, D. D. Nguyen, and J. Zhang, "Sling bag and backpack detection for human appearance semantic in vision system," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 2130–2135.

[7] Y. Qi, G.-C. Huang, and Y.-H. Wang, "Carrying object detection and tracking based on body main axis," in *2007 International Conference on Wavelet Analysis and Pattern Recognition*, vol. 3, Nov 2007, pp. 1237–1240.

[8] T. Senst, R. Evangelio, and T. Sikora, "Detecting people carrying objects based on an optical flow motion model," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, Jan 2011, pp. 301–306.

[9] T. Senst, A. Kuhn, H. Theisel, and T. Sikora, "Detecting people carrying objects utilizing lagrangian dynamics," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, Sept 2012, pp. 398–403.

[10] A. Tavanai, M. Sridhar, F. Gu, A. Cohn, and D. Hogg, "Carried object detection and tracking using geometric shape models and spatio-temporal consistency," in *Computer Vision Systems*, ser. Lecture Notes in Computer Science, M. Chen, B. Leibe, and B. Neumann, Eds. Springer Berlin Heidelberg, 2013, vol. 7963, pp. 223–233.

[11] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ser. ECCV '02. London, UK, UK: Springer-Verlag, 2002, pp. 343–357.

[12] C.-S. Lee and A. Elgammal, *Carrying Object Detection Using Pose Preserving Dynamic Shape Models*. Springer Berlin Heidelberg, 2006, pp. 315–325.

[13] G. Tzanidou, I. Zafar, and E. Edirisinghe, “Carried object detection in videos using color information,” *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 10, pp. 1620–1631, Oct 2013.

[14] R. Dondera, V. Morariu, and L. Davis, “Learning to detect carried objects with minimal supervision,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 759–766.

[15] X. Jia, H. Lu, and M. H. Yang, “Visual tracking via adaptive structural local sparse appearance model,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1822–1829.

[16] A. Papazoglou and V. Ferrari, “Fast object segmentation in unconstrained video,” in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1777–1784.

[17] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[18] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, “Motion coherent tracking using multi-label mrf optimization,” *International Journal of Computer Vision*, vol. 100, no. 2, pp. 190–202, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11263-011-0512-5

[19] Y. J. Lee, J. Kim, and K. Grauman, “Key-segments for video object segmentation,” in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1995–2002.

[20] D. Zhang, O. Javed, and M. Shah, “Video object segmentation through spatially accurate and temporally dense extraction of primary object regions,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 628–635.

[21] C. Rother, V. Kolmogorov, and A. Blake, “"grabcut": Interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[23] L. Wang, J. Shi, G. Song, and I.-F. Shen, “Object detection combining recognition and segmentation,” in *Proceedings of the 8th Asian Conference on Computer Vision - Volume Part I*, ser. ACCV’07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 189–199.

[24] “Workshop on performance evaluation of tracking and surveillance (pets),” http://www.cvg.reading.ac.uk/PETS2006/data.html.

[25] “i-lids dataset for avss 2007,” http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html.

[26] F. P. X. X. G. Gabriela Csurka (Xerox Research Centre Europe), Diane Larlus, “What is a good evaluation measure for semantic segmentation?” in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2013.