2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Sling Bag and Backpack Detection for Human Appearance Semantic in Vision System

Teck Wee Chua, Karianto Leman, Hee Lin Wang, Nam Trung Pham,
Richard Chang, Dinh Duy Nguyen and Jie Zhang

Abstract— **In many intelligent surveillance systems there is a requirement to search for people of interest through archived semantic labels. Other than searching through typical appearance attributes such as clothing color and body height, information such as whether a person carries a bag or not is valuable to provide more relevant targeted search. We propose two novel and fast algorithms for sling bag and backpack detection based on the geometrical properties of bags. The advantage of the proposed algorithms is that it does not require shape information from human silhouettes therefore it can work under crowded condition. In addition, the absence of background subtraction makes the algorithms suitable for mobile platforms such as robots. The system was tested with a low resolution surveillance video dataset. Experimental results demonstrate that our method is promising.**

Fig. 1. A combination of static pictures extracted from surveillance footage shows the suspected suicide bomber with backpack and sling bag at Bulgaria's Burgas airport, on July 18, 2012.

I. Introduction

With increasing threats from criminals and terrorists, industries and government agencies around the world are now focused on security. Video surveillance is playing a central role in security efforts. Conventional systems like Closed Circuit TV (CCTV) could only provide passive recording capabilities. On the other hand, intelligent video analytics can provide proactive security solutions. The core of such systems is the ability to perform automatic event monitoring or interpretation of the video contents such as extraction of semantic descriptions from a tracked person. Those semantic attributes will be stored together with the historical walking path of the person. The stored information can be used for potential forensic support. For example, Fig. 1 shows a man that the Bulgarian Interior Ministry says is the suicide bomber that killed seven people including himself at the airport. Some associated semantic attributes could be 'blue shirt', 'long hair', 'wearing hat', 'carrying backpack' and 'carrying sling bag'.

In this work, we are interested in detecting sling bags and backpacks in real-time. One of the earliest works was "Backpack" algorithm [1] that detects people carrying objects by computing aligned silhouette periodicity and shape symmetry analysis. The algorithm is based on the observations that human body shape is symmetric and people exhibit motion periodicity when they are moving unencumbered. Similar to [1], the method proposed in [2] also aligns silhouette to produce temporal template. Instead of assuming the silhouette of unencumbered people is symmetric, the template is compared against view-specific exemplars of unencumbered people generated using 3D software. A wavelet approach is used in [3] to extract features from silhouettes and neural network is trained on a set of positive and negative samples. In [4], the silhouette of a person is divided into four horizontal segments. The temporal variation of the horizontal bounding box width is represented as time series. Bag is detected when the time series does not satisfy some periodicity constraints. Likewise, [5] computes star skeletonization on silhouettes and extracts the normalized x-y coordinates of the star limbs. The temporal coordinates are represented as time series and they are compared against several thresholds to determine the existence of bag. Method proposed in [6] uses the distances between the silhouette boundary points and body main axis as the features. The feature dimension is later reduced using principal component analysis and support vector machine is used for classification. In [7] and [8], the authors use foreground density features with spatial granularity and homographic calibrated object size features to classify human and luggage from silhouettes. Another branch of technique relies on human gait analysis to detect the presence of bag. In [9], a set of Gabor based human gait appearance models is used to extract features from averaged silhouettes. The higher order feature is classified using general tensor discriminant analysis. Note that all methods stated above require an accurate background subtraction technique to segment foreground objects. This is often difficult to achieve in video surveillance due to the challenges such as poor lighting condition, shadows and moving background. Another drawback is that most systems above require profile view that reveal the protruding part of the bag. In addition, those methods are unable to handle crowded scene when foreground regions overlap or merge. Silhouette alignment may not be feasible due to large variations of human postures

The authors are with Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632.
tewchua@i2r.a-star.edu.sg

and camera views.

A statistical optical flow based motion model was proposed in [10] to describe the motion of people that can be used to detect people carrying objects. However, the computation of optical flow is expensive and motion information is not available when people are static.

In this paper, we present two novel and efficient algorithms for sling bag and backpack detection that do not require foreground segmentation and are capable of running in real-time. The proposed system attempts to detect the presence of bags from multiple directions including frontal and rear views in addition to profile view. This enables the detection of bags in open spaces such as concourse in a subway station or hotel lobby as people move in different directions. The algorithm was implemented as part of high level human appearance description module in a real-time multi-camera surveillance system. Therefore, the proposed method is designed with practicality and robustness in mind. It should be noted that the detection of non-wearable bag such as handbag and trolley luggage is beyond the scope of this work. We also assume there exists good contrast between bags and clothings.

This paper is organized as follows. Section II describes the framework of the proposed method including head localization and upper body estimation, segmentation of potential bag region, and followed by the computation of distinctive shape features for sling bag and backpack detection. Experiment results and discussions are given in Section III. Finally, Section IV concludes the paper.

II. Proposed Algorithm for Sling Bag and Backpack Detection

Fig. 2 presents the overall architecture of our proposed system which includes an analytic engine, a database, and a search interface. The semantic attributes and the person's unique identifier (tracked ID) detected by the analytic module are stored in the database which allow user to perform search queries. The detail description of each component in the analytic engine is explained next.

A. Head localization and upper body estimation

Since human head is the least occluded part of the body during crowded condition, we use head tracking algorithm to track the individuals rather than full body counterpart. For head detection, we extract the head feature using Local Binary Pattern (LBP) [11] and Histogram-of-Oriented-Gradient (HOG) [12]. The detector was learned from large amount of training data from different angles using Adaboost [13]. Here, we do not emphasize on any specific detector. We believe that any properly trained head detector would be sufficient. Once the head has been localized, we employ the tracker from [14] to track the individuals. With tracking information, the move direction of the individual can be obtained. If a person is static, we record its last moving direction. Next, the upper body bounding box $[x_{tl\,\mathrm{B}}\ y_{tl\,\mathrm{B}}\ w_{\mathrm{B}}\ h_{\mathrm{B}}]$ is estimated from the head bounding box as the followings:

$$x_{tl\,\mathrm{B}} = x_{tl\,\mathrm{H}} - w_{\mathrm{H}} \times 0.2 \tag{1}$$

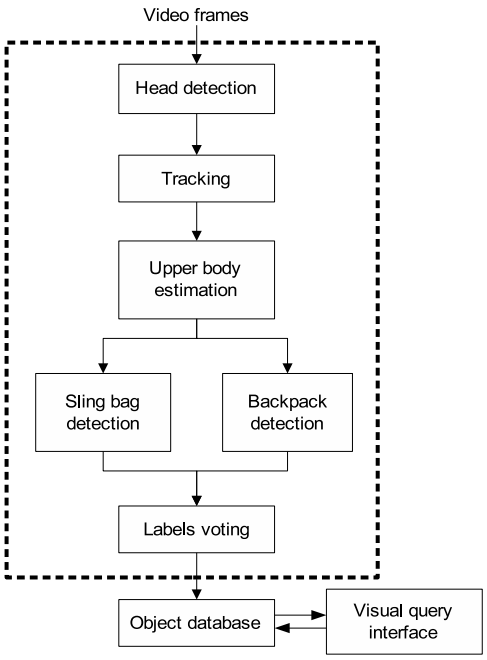


Fig. 2. Overall system architecture. Analytic module is located within the dashed line box.

$$y_{tl\,\mathrm{B}} = y_{bl\,\mathrm{H}} \tag{2}$$

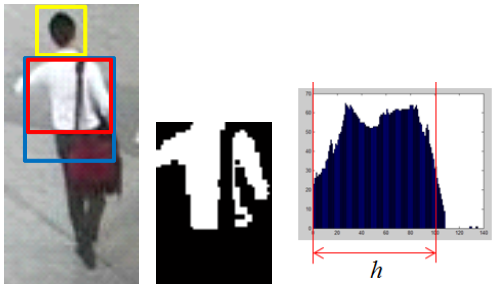
$$w_{\mathrm{B}} = w_{\mathrm{H}} \times 1.2 \tag{3}$$

$$h_{\mathrm{B}} = h_{\mathrm{H}} \times 2.0 \tag{4}$$

where subscripts 'B' and 'H' denote body and head respectively while 'tl', 'bl', 'w', 'h' refer to top-left, bottom-left, width and height of the bounding box. Since most surveillance cameras are placed at non-lateral angle, the upper body height will vary w.r.t. the object distance from the camera. Besides, we also noticed that the shirt length also depends on whether the person tucks in or not. As such, we set the initial upper body height as (4). Note that the initial height may cover part of the lower body, the exact upper body will be redefined as follows. Firstly, we apply spatial averaging on the region-of-interest (ROI). Secondly, color histograms are computed to search for the most dominant color. The dominant color value is set as the target value. Next, K-means clustering is computed to split the ROI into two regions. The region that has mean color value closest to the target value is segmented as upper body as shown in Fig. 3(b). The binary pixels are projected onto Y-axis, the redefined height is computed as sum of the histogram bin that has its value above 50% of w_B as shown in Fig. 3(c).

B. Segmentation of Potential Bag Regions

As mentioned earlier, we made the assumption that the contrast between the bag and clothing regions is sufficient for detection. In fact, based on our observation bags are usually darker than the clothes. We use adaptive thresholding technique to segment the darker region. Moreover, we apply an oval-shaped mask to remove the dark background regions



(a) (b) (c)

Fig. 3. (a) Yellow box: head bounding box, blue box: initial upper body bounding box, red box: refined upper body bounding box (b) grayscale K-Means clustering result, and (c) Y-axis projection histogram of (b) to determine the refined upper body height, h.

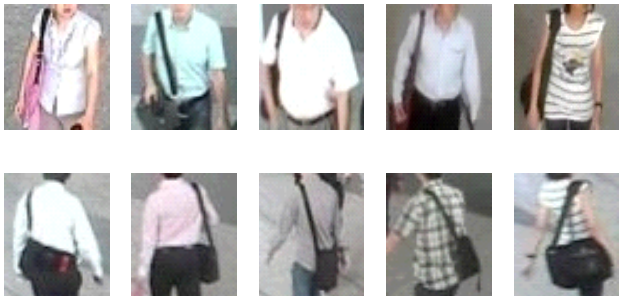


Fig. 4. Examples of sling bag from frontal (top row) and rear (bottow row) views.

at the boundaries. The ROI is normalized to 75×100 pixels to eliminates the variations of human sizes.

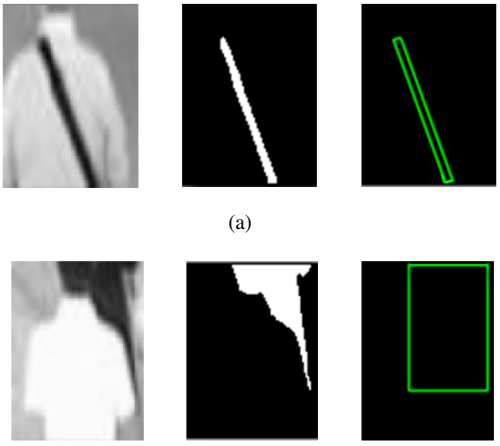
C. Sling Bag Detector

As seen in Fig. 4, the most obvious cue of a sling bag is its straight strap across the upper body. However, due to personal preference of the individual or the variations in the camera angles, the strap may appear in different directions. This makes the detection through model-based learning approaches difficult. Here, we propose an efficient two-pass framework that utilizes two different detectors to detect the presence of sling bags. The main idea is to detect the existence of narrow near-parallel lines across the upper body.

The first detector attempts to detect near-parallel lines using geometrical property of the strap. Given a blob, we compute the contour and get its perimeter, L_c. The contour is then approximated with a polygon with accuracy proportional to the contour perimeter. Next, we compute the minimal area bounding rotated rectangle of the approximated contour. We define a metric to measure the parallelism of the rectangle:

$$P = \frac{2 \times \max(b_w, b_h)}{L_c} \tag{5}$$

where b_w and b_h denote the width and height of the rotated rectangle. The maximum theoretical value of P is 1. If the blob is long rectangle shaped then P should be close to 1. We set a threshold of $P \geq 0.8$ to detect the presence of sling bag's strap. For example, the shape in Fig. 5(a) returns the parallel metric P of 0.91 while (b) returns 0.496, thus only (a) will be classified as carrying sling bag.



(a)

(b)

Fig. 5. (Left) original image, (middle) segmented possible bag region, and (right) minimal area bounding rotated rectangle.

In the second detector, the edges of the potential bag regions are extracted using canny edge detection. Probabilistic Hough transform [15] is computed to find the pairs of near-parallel lines with width less than 15 pixels, minimum length of 20 pixels, and angle $35° \leq \theta \leq 90°$. Note that the threshold values are determined based on the normalized ROI size stated in Section II-B.

As far as the robustness is concerned, the geometrical-based detector is less sensitive against the clothing with rich texture. However, it is unable to handle the case where the blob of the strap merges with the dark background blob, which violates the rectangular blob rule. On the other hand, the Hough transform-based detector may still detect partial near-parallel lines even though the blob is not in rectangular shape. To optimize the detection result, we propose a two-pass framework which uses the geometrical-based detector in the first pass and Hough transform-based detector in the second pass. Once a sling bag has been detected in the first pass, the second pass can be skipped. Otherwise, the second pass will be carried out to detect the sling bag.

D. Backpack Detector

As shown in Fig. 6, the detection of backpacks could be challenging due to appearance variations caused by changing body postures and different camera views. We propose a systematic detection approach based on the geometrical shapes of backpacks.

After the potential bag regions has been segmented, morphological operations are performed to clean up the noises and re-connect the loosely disconnected blobs. We assume that the largest blob is the most probable bag region. We exploit the geometrical properties of the backpack shapes to determine the presence of backpack. The advantage of using geometrical properties is that the detection is more robust against moderate rotation (about 30 degrees tolerance).

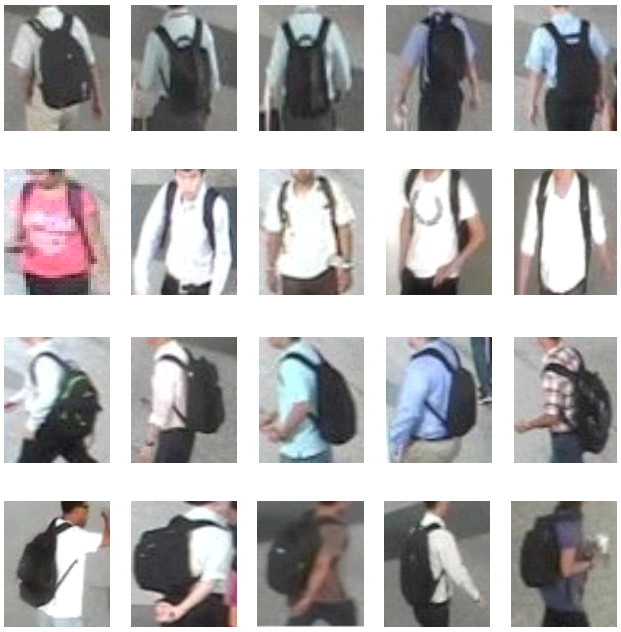


Fig. 6. Examples of backpack from rear (first row), frontal (second row), left profile (third row), and right profile (fourth row) views.

Contour analysis is used to extract the external boundary of the blob. Next, we find the convex hull of the boundary and locate all the convex defects. Fig. 7 shows an example of segmented backpack blob with its convex hull and defects. Each edge (consists of two segments) is associated with a convex point. Due to the jagged edge effect of the enlarged image ROI, many defects with shallow depth will be detected. However, those defects can be safely removed as it is highly unlikely that they belong to the real edges of the bag. We set a depth threshold of 3 pixels. Defects with too small or too large angle between two segments are also discarded. Besides, to further eliminate the false candidates we only keep the defects with the ratio between two segments, $\frac{L_s}{L_l} \geq 0.5$, where L_s and L_l denote the shorter and longer segments respectively. Here we define three types of convex defects (see Fig. 8):

- Top defect: this upright 'V'-shaped defect is formed by the straps around the shoulder region.

$$\begin{aligned} 0^\circ \leq \theta_{T1} \leq 85^\circ \\ 95^\circ \leq \theta_{T2} \leq 180^\circ \\ 30^\circ \leq |\theta_{T1} - \theta_{T2}| \leq 135^\circ \end{aligned}$$

- Left defect: this left-opening defect is formed by the left edge of the backpack.

$$\begin{aligned} 65^\circ \leq \theta_{L1} \leq 170^\circ \\ 195^\circ \leq \theta_{L2} \leq 270^\circ \\ 30^\circ \leq |\theta_{L1} - \theta_{L2}| \leq 165^\circ \end{aligned}$$

- Right defect: this right-opening defect is formed by the right edge of the backpack.

$$\begin{aligned} 0^\circ \leq \theta_{R1} \leq 105^\circ \\ 270^\circ \leq \theta_{R2} \leq 345^\circ \\ 30^\circ \leq |\theta_{R1} - \theta_{R2}| \leq 165^\circ \end{aligned}$$

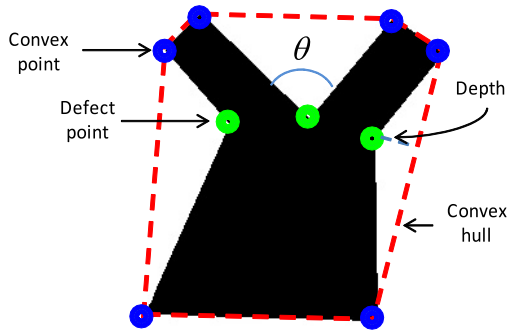


Fig. 7. Backpack detection: convex hull is computed to obtain the convex points and defect points. The angle θ formed by the points is used to characterize the backpack boundaries.

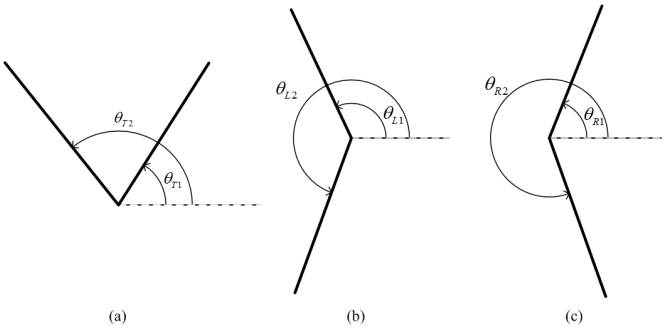


Fig. 8. Defect types for backpack detection: (a) top defect (b) left defect and (c) right defect.

Since the appearance of bag looks different from various views as shown in Fig. 6, we use different criteria to classify the backpack edges to detect the presence of backpack. From the tracking information, we are able to estimate the moving direction of a person. Therefore, it is possible to know the view of the upper body:

1) Rear View
   - A backpack is detected if the top defect and at least one side defect are detected.
2) Front View
   - A backpack is detected if there are two near-parallel straps detected using the two-pass framework described in Sec. II-C.
3) Left Profile View
   - A backpack is detected if left defect is detected.
4) Right Profile View
   - A backpack is detected if right defect is detected.

E. Labels Voting

For each tracked person, we accumulate all the detection class labels ('no bag', 'backpack', 'sling bag') during the

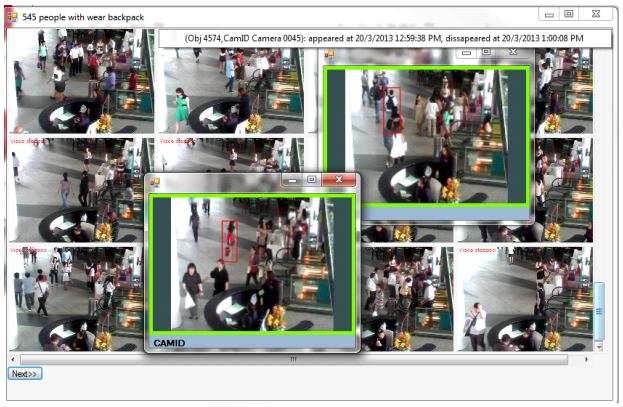


Fig. 9. Target search user interface which allows high level semantics to be used for query.

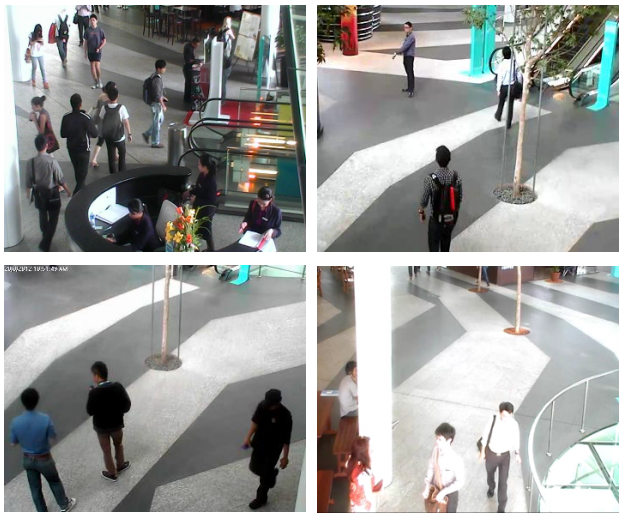


Fig. 10. Example of video frames.

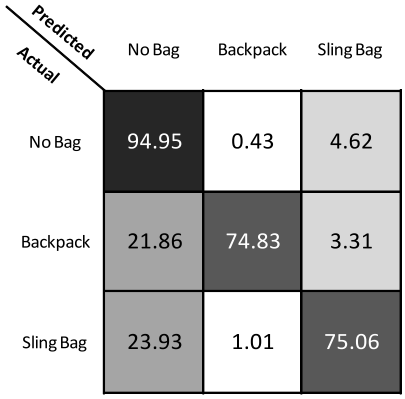


Fig. 11. Confusion matrix.

tracking duration and perform majority voting to determine the classification result. Note that for frontal view backpack detection (case 2 in the previous section), both straps may not be simultaneously visible at all the time due to the varying walking directions of the person. As such, for this particular view we weighted the vote for backpack class two times higher than the other classes. For example, suppose the total number of frames is 20 frames, in which 6 frames are classified as 'no bag', 8 frames are classified as 'sling bag', and 6 frames are classified as 'backpack'. The normalized votes for 'no bag', 'sling bag' and 'backpack' considering the weight are 0.23, 0.31, and 0.46 respectively. Therefore, the person is considered carrying a backpack. The final decision along with the person ID will be stored in the database for forensic search or multi-camera tracking system. As mentioned earlier, the backpack and sling bag detection algorithms are implemented as part of the analytic modules in our multi-camera tracking system, along with hair length and clothing color detection modules, to generate high level human descriptions. Fig. 9 shows our graphical user interface for target search with semantic labels. The user can compose query such as people with red shirt, short hair and carrying backpack during a certain time duration.

III. Experiments

We have evaluated the performance of the algorithms using 13 real surveillance video footages covering different time of the day and camera views. Each video of 30 minutes duration was recorded at 12.5 fps at CIF (352x288 pixels) resolution. Fig. 10 shows some examples of image frames. We neglect the upper body with less than 20 pixel width as the details of the bags tend to vanish at this resolution. In the case of inter-person occlusion, we ignore the upper body with more than 50% overlap. In order to achieve a more reliable result, we only consider the individuals with at least 20 tracked frames (approximately 1 second) in the scene. The total number of people that are successfully tracked in all the videos is 1241 in which 693 people do not carry bag, 151 people carry backpacks, and 397 people carry sling bags.

The framework has been implemented in C++ and runs on an Intel 2.4 GHz quad-core computer. The computation of sling bag detection on one 75×100 pixels ROI requires 3ms on average while backpack detection requires 4ms.

Fig. 11 shows the confusion matrix of the classification results. The average accuracy for all three classes is 81.61%. The accuracy of detecting a person without bag correctly is much higher which is at 94.95%. A small number of people without any bag is misclassified as carrying backpack (0.43%) or sling bag (4.62%). The reason why there are more misclassification cases for the later is that the detection of sling bag straps may not be as robust as the detection of the backpack shapes which is more distinctive. About $20 - 25\%$ of people who carry bags are falsely classified as without bags. Most of the false detections are due to the segmentation error that causes the bag regions merge with the clothings or dark background. As a consequence, the outline of the segmented region does not reflect the characteristic of a backpack or a sling bag. Upon investigation, there are also some interesting false detection cases as shown in Fig. 12 and Fig. 13. For example, some people carry their backpacks only

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

Fig. 12. Some examples of sling bag detection errors due to: (a) the strap is too short, (b) the strap is too thin, (c) the segmentation errors with texture clothing, (d) the strap is covered by hand, (e) backpack is wore as sling bag but the strap is too short to be detected.

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

Fig. 13. Some examples of backpack detection errors due to: (a–b) the hairs and jacket at both shoulder area are mistaken as the backpack straps, (c) the hair region is merged with the clothing, (d) the segmented shirt fulfils the convex defect criteria, (e) one side of the straps is much thinner/shorter than the other side.

on one shoulder[1]. From the rear view, there is no formation of 'V' shape from both straps thus it is not classified as backpack. Moreover, the short strap segment on one shoulder is also too short to be classified as sling bag. There are also cases where the hairs are much longer than both shoulders, thus the hairs portions are misclassified as backpack straps from the frontal view.

There are 3.31% of the backpack cases being misclassified as sling bag. Most of them occur during the near frontal view where the both strap sizes may not be the same or one strap is occasionally not visible (see Fig. 13(e)). A relatively small number of people carrying sling bags (1.01%) or without bag (0.43%) are misclassified as carrying backpacks. This is because from the frontal view long hair can confuse the system such that the hair was detected as backpack straps.

Since the problem we are dealing with is considerably novel, there are no directly similar methods to compare our results with. For instance, methods in [1]–[9] require computation of silhouette which cannot handle crowded scenes while methods in [10] only deal with trolley luggage.

IV. Conclusions

We have presented a framework for detecting sling bags and backpacks from surveillance cameras with low resolution and various lighting conditions. We exploit geometrical features of bags from images directly rather than silhouettes. Since our method does not rely on the silhouettes computed from background subtraction, it can handle more crowded condition and applicable to any mobile platforms such as humanoid or sentry robots. The method also does not require lateral camera view to capture the protruding parts of the bags. Furthermore, the proposed framework is able to detect bags from different body orientations. Another advantage of the method is that the detection speed is very fast for real-time computation. This makes it suitable to run concurrently with other more computationally intensive video analytics. Although the proposed approach provides promising results, the framework require good contrast between clothings and bags. In future work, we will be investigating the use of more robust segmentation algorithm that could better separate the bag regions from the clothings. Notwithstanding the existing limitations, the framework presents a novel way of detecting bags at the torso region in multi-directional views through the geometrical features.

[1]We consider the person carries a sling bag in such case.

References

[1] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, "Backpack: Detection of people carrying objects using silhouettes," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 1999, pp. 102–107.

[2] D. Damen and D. Hogg, "Detecting carried objects in short video sequences," in *Proc. of European Conference on Computer Vision (ECCV)*, 2008, pp. 154–167.

[3] A. Branca, M. Leo, G. Attolico, and A. Distante, "Detection on objects carried by people," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, 2002, pp. 317–320.

[4] C. B. Abdelkader and L. Davis, "Detection of people carrying objects: A motion-based recognition approach," in *Proc. of IEEE Conference on Automatic Face and Gesture Recognition (FGR)*, 2002, pp. 378–383.

[5] R. Chayanurak, N. Cooharojananone, S. Satoh, and R. Lipikorn, "Carried object detection using star skeleton with adaptive centroid and time series graph," in *Proc. of International Conference on Signal Processing (ICSP)*, 2010, pp. 736–739.

[6] Y. Qi, G.-C. Huang, and Y.-H. Wang, "Carrying object detection and tracking based on body main axis," in *Proc. of International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, 2007, pp. 1237–1240.

[7] V. A.-Vanacloig, J. R.-Ortega, G. A.-García, and J. M. V.-González, "People and luggage recognition in airport surveillance under real-time constraints," in *Proc. of International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[8] J. R.-Ortega, G. A.-García, V. A.-Vanacloig, and J. M. V.-González, "Feature sets for people and luggage recognition in airport surveillance under real-time constraints," in *Proc. of International Conference on Computer Vision Theory and Applications (VISAPP)*, 2008, pp. 662–665.

[9] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Human carrying status in visual surveillance," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1670–1677.

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

[11] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 971–987, July 2002.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[13] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 511–518.

[14] N. T. Pham, K. Leman, and T. W. Chua, "Sequential particle filter for multiple object tracking," in *Proc. of IAPR Conference on Machine Vision Applications (MVA)*, 2011, pp. 63–66.

[15] N. Kiryati, Y. Eldar, and A. M. Bruckstein, "A probabilistic hough transform," *Pattern Recognition*, vol. 24, no. 4, pp. 303–316, 1991.