# Learning to Detect Carried Objects with Minimal Supervision*

Radu Dondera, Vlad Morariu and Larry Davis
University of Maryland
College Park, MD USA
{rdondera, morariu, lsd}@cs.umd.edu

## Abstract

*We propose a learning-based method for detecting carried objects that generates candidate image regions from protrusion, color contrast and occlusion boundary cues, and uses a classifier to filter out the regions unlikely to be carried objects. The method achieves higher accuracy than state of the art, which can only detect protrusions from the human shape, and the discriminative model it builds for the silhouette context-based region features generalizes well. To reduce annotation effort, we investigate training the model in a Multiple Instance Learning framework where the only available supervision is "walk" and "carry" labels associated with intervals of human tracks, i.e., the spatial extent of carried objects is not annotated. We present an extension to the miSVM algorithm that uses knowledge of the fraction of positive instances in positive bags and that scales to training sets of hundreds of thousands of instances.*

## 1. Introduction

In the field of visual surveillance one of the important problems that has received increased attention in recent years is the detection of objects carried by people. The train bombings carried out in Madrid and London in recent years are strong incentives for a computer vision solution, but there are also other applications, especially military, that require awareness of object presence. While significant progress has been made in detecting and tracking humans, the variability in the appearance of the objects people can carry makes carried object detection a very challenging problem. Capturing the relationships of the object with the human silhouette is also hard, as objects may or may not have color contrast with clothing; may occupy a small fraction of the human silhouette or can be comparable in height with the human; may be carried by hand, under the arm, with both arms, or on the back. Finally, objects may be swung or held still and they may be occluded in some of the frames in which the human is observable.

The most successful approaches so far to finding carried objects have extracted a foreground mask of the human and then matched and subtracted a generic body template (either 2D [9] or 3D [22]), returning the protrusions as objects. While this approach is intuitively appealing, it cannot detect objects in the frequent case when they are mostly inside the human silhouette, in the 2D setting, and it requires a stereo camera moving among people, in the 3D setting. Directly using other cues such as color and motion to find carried objects is bound to produce numerous false alarms corresponding to the head, feet, hands, or just noise, but for human vision it is easy to distinguish body parts from carried objects when displayed together with the human silhouette. We propose a method to detect carried objects that applies three types of low level detectors inside human bounding boxes (based on protrusions, color contrast and occlusion boundaries) and models the resulting image regions as carried objects with a kernel SVM on features related to the human silhouette context.

As the performance of the classifier is directly related to the size of the training set, and as the object annotation process is time consuming (roughly 40,000 precise bounding boxes are needed for one of the datasets in this work), we investigated using a multiple instance learning (MIL) framework. MIL, introduced by [10], departs from the classic supervised learning setting by making labels available for sets of instances (bags) rather than individual instances; in each positive bag there is at least one positive instance while all the instances in negative bags are negative. In our setting, instances are image regions produced by low level detectors and bags are sets of instances from intervals of human tracks annotated as "walk" (no carried object) or "carry" (at least one object), and we focus on instance level classification. Most MIL approaches are computationally intractable for our datasets (our problems range from approximately 12,000 to 192,000 instances), and the few that are tractable–miSVM [3] and sbMIL [6]–can have significantly lower test set accuracy than a fully supervised classifier. Observing

*This research is supported by ONR grant N000141010766.

that our low level detectors produce a roughly constant fraction of correct regions when the human is carrying an object, we extend miSVM to adjust the fraction of positive labels in positive bags accordingly at each iteration.

Our contribution is two fold: (1) we propose a novel learning-based method for carried object detection with accuracy exceeding state-of-the-art and with good generalization capability; (2) we extend the miSVM algorithm to account for an expected positive bag density, achieving improved accuracy for virtually the same computational cost.

## 2. Related Work

The majority of papers on carried object detection follow the pattern of estimating the pixel mask of the person and object, subtracting from it a human template (either abstract or learned from data) and returning the remaining regions. Haritaoglu et al. [16] used background subtraction, averaged human masks temporally, and relied on the symmetry of the walking human silhouette around a principal axis and on the periodic nature of limb motions. Lee and Elgammal [18] proposed a generative silhouette appearance model parameterized by viewpoint, body proportions and gait phase, and iteratively estimated these parameters together with holes in the foreground mask and outlier regions (carried objects). Noting the sensitivity of Haritaoglu et al.'s method to the principal axis estimate, Damen and Hogg [9] matched and subtracted synthetically rendered templates of unencumbered humans. To select the correct template, they require a ground plane homography and an estimate of the walking direction. The most recent work related to carried objects utilized a cylindrical 3D shape representation of humans both in a tracking-before-detection framework and for carried object detection [22]. 2D template subtraction approaches are limited to discovering objects that significantly protrude from the silhouette and their accuracy is dataset dependent – the results section shows Damen and Hogg's method [16] performing poorly when people wear robes. To improve both the recall and the precision of 2D carried object detection, we propose using multiple sources of candidate object regions and then pruning these candidates in the context of the human silhouette.

Interesting context modeling work by Zheng et al [28] effectively combines the appearance of an object with that of its neighborhood. Other efforts focus on deciding whether people carry something or not, without providing an actual location for the object [26] [23]. While knowing carrying status is valuable, precise object masks are directly usable in important higher level tasks like detecting abandoned objects, theft or object exchange. Unfortunately, much more annotation effort is involved in learning-based methods that explicitly localize objects, but we adopt a MIL framework and still require only weak supervision in the form of carry status.

The Multiple Instance Learning literature is extensive, covering aspects as varied as discovering a single concept shared by positive but not negative bags [21], finding the most appropriate exemplar embedding [8], explicitly factoring in the cost of false positives in the classification task [15], to give just a few examples. As Li et al. [20] noted, most approaches that can classify instances have prohibitive training cost. An exception is the miSVM framework of Andrews et al. [3], who cast MIL as a mixed integer program involving the labels of instances in positive bags and the parameters of the separating hyperplane, and solved it with an iterating heuristic with good performance in practice. Gehler and Chappelle [13] added to the SVM formulation of [3] a term correlated to label uncertainty that allows finding better local minima of the objective function. However, this leads to very high computational cost if the number of instances in positive bags is large, since the SVM solver sees these instances duplicated as both positive and negative. The approach most directly applicable to our setting is due to Bunescu and Mooney [6], who loosened a constraint in their SVM formulation so that as few as one instance per positive bag can be labeled positive. The results of their approach are inferior to miSVM [3] in our problems, which we believe is because too few of the actual positive instances are labeled positive.

A few researchers used MIL to cope with noisy labels when learning from images retrieved with search engines [27] [20] [19]. Li et al. [20] leveraged the constraint that the fraction of positives in a positive bag is relatively large ($0.6$) and proposed an iterative scheme that trained on an increasingly larger number of bags. In [19], they reduced the high computational cost of the optimization run in each iteration and updated a separating hyperplane incrementally. It is very unlikely that these two methods would be applicable to our problem setting, as the positive bag density varies from $0$ to $0.5$ and the decision surface has to make multiple local distinctions between various objects and body parts.

Lastly, two papers bear superficial resemblance to our work. Fathi et al. [11] used egocentric video to learn to discriminate between object appearances with little supervision. While both works learn to classify image regions in a MIL framework, the problems considered are significantly different: [11] employs multi-class MIL for relatively small training sets, while we use two class MIL for large amounts of data. Ghanem and Davis [14] also adopted a learning approach in connection to carried objects, but could only predict object appearance/disappearance events holistically.

## 3. Low Level Detectors

Our method assumes that human tracks are available and runs background subtraction [17] and optical flow [24]. Next, three types of image region detectors are run: an optical flow-based protrusion detector, a segmentation-based

(a) (b) (c)

Figure 1: Sample output of low level detectors: (a) optical flow-based protrusion (b) segmentation-based color contrast (c) occlusion boundary-based moving blob. Each of these is too noisy as a carried object detector, but human silhouette context can be used effectively to filter its output.

color contrast detector and an occlusion boundary-based moving blob detector, see Figure 1. The three detectors are simple but have high probability of finding carried objects if they exist; if none of them fires during an interval in which a person is carrying an object, then most likely the object does not protrude, has poor contrast, and is static with respect to the body – an extremely hard target to find. We ignore such cases and instead address the problem of disambiguating between image regions corresponding to body parts/noise versus those that are carried objects, using the context of the human silhouette. With respect to [9], we additionally require optical flow but we improve on their detector (section 3.1), use two additional detectors and employ a mechanism to select the correct regions.

### 3.1. Optical Flow-based Protrusion Detector

The optical flow-based protrusion detector builds a probabilistic mask for each human bounding box that reflects how close the motion of a pixel is to the average translation in the box. We call this the carried probability mask (CP) and we define it by assuming that the projection of a pixel's velocity on the average translation is normally distributed:

$$CP(p) \propto \exp\left(-\frac{\left(\frac{\mathbf{w}(p)\cdot\bar{\mathbf{w}}}{|\bar{\mathbf{w}}|_2^2}-1\right)^2}{2\sigma^2}\right) \quad (1)$$

where $\mathbf{w}(p) = (u(p), v(p))$ is the optical flow vector at pixel $p = (x, y)$ and $\bar{\mathbf{w}}$ is the mean optical flow of the points in the human bounding box. (To compensate for camera motion, the average optical flow over the image is subtracted from all optical flow vectors.) We visualized CPs for a range of $\sigma$'s, observed that smaller values produce holes and larger values overestimate the human shapes, and chose $\sigma = 0.4$ for all videos we process. Limbs swinging opposite to the walking direction tend to be removed, which is advantageous over using background subtraction masks as in [9], since the temporal aggregation for noise reduction can be done effectively on a smaller time interval, e.g. 9 frames as opposed to 50 needed by [9]. We aggregate the CP masks by simply translating them opposite to the average optical flow vector and call the thresholded resulting mask average carrying shape (ACS). The ACS's of unencumbered pedestrians tend to be urn-shaped regardless of viewpoint, which allows matching against a single urn+head template with shape contexts [4] and then retrieve protrusions, see supplemental material. Compared to our protrusion detector, Damen and Hogg [9] incur the disadvantage of needing a ground plane homography and an estimate of the walking direction to select the proper template.

### 3.2. Segmentation-based Color Contrast Detector

The color contrast detector runs mean shift clustering on the foreground mask obtained with background subtraction. Foreground pixels are represented with $[0, 1]$ normalized rgb and image positions (the positions are normalized with respect to the human bounding box). The clustering bandwidth is set to 0.2 for all videos in all datasets; other values do not lead to significantly different segmentations with respect to the carried objects. This detector is designed for situations when the object's color clearly stands out from the colors of the human silhouette, as in Figure 1b. As the figure shows, many false positives occur, but a large portion are meaningful parts of the silhouette e.g., body and head.

### 3.3. Occlusion Boundary-based Moving Blob Detector

If the person moves the carried object with respect to the body or changes viewpoint while walking, occlusion boundaries will likely appear around the object. To detect them we employ criteria from [25]: occlusion boundaries are pixels where the flow forward from a frame is inconsistent with the flow back into the frame or where the flow gradient has large magnitude. With respect to [25], we tighten the first condition and loosen the second, requiring more consistency but allowing for larger gradient magnitudes:

$$|\mathbf{w}(p) + \mathbf{w}'(p')|_2^2 > 0.01(|\mathbf{w}(p)|_2^2 + |\mathbf{w}'(p')|_2^2) + 0.01 \quad (2)$$

$$|\nabla u(p)|_2^2 + |\nabla v(p)|_2^2 > 0.01|\mathbf{w}(p)|_2^2 + 0.01 \quad (3)$$

where $p' = p + \mathbf{w}(p)$ and $\mathbf{w}'$ is the backward optical flow field. Superimposing the boundary mask on the foreground mask from background subtraction segments the latter into candidate regions. Empirically we observe this detector frequently finds people's heads and feet.

## 4. Learning a Model for Carried Object Regions

The candidate image regions retrieved by the low level detectors are filtered to remove noise: regions less than 10 pixels in width or height, or greater than half the size of the human mask are eliminated. We also use a compactness filter requiring a region to occupy at least half its minimum area (not necessarily axis aligned) enclosing rectangle. The method might miss some types of objects (e.g. semi-automatic weapons), but since compactness is one of the features we compute for regions, the choice can be reverted by simply removing this filter. The cost is introducing more types of negatives and making learning harder.

### 4.1. Region Features

The inspiration for features comes from common sense knowledge about body parts, e.g. the head is near the top of the silhouette, shares contour points with it and is relatively small. We compute twelve features and use a Gaussian kernel SVM for classification. Three features characterize the shape of a region and nine capture its relation to the human silhouette. (To clarify, we use the term silhouette to denote all points inside a shape as opposed to just its contour.) The silhouette produced by background subtraction is processed with a morphological "open" prior to feature computation to reduce the noise of the estimated silhouette height. The features are:

- compactness: ratio of the region size to the area of its enclosing rectangle
- orientation: the angle of the largest side of the enclosing rectangle with the vertical direction ($\in [0, \frac{\pi}{2}]$)
- aspect ratio: the ratio of the larger side of the enclosing rectangle to the smaller side
- relative size: the ratio of the region size to the silhouette size
- relative x: the absolute difference between the $x$ of the region centroid and the $x$ of the silhouette centroid, normalized by silhouette height (the width is too noisy)
- relative y 1: minimum $y$ of the region normalized with respect to vertical silhouette span
- relative y 2: maximum $y$ of the region normalized with respect to vertical silhouette span
- fraction of horizontal occupancy: the ratio of the region size to the silhouette area between the region's smallest and largest $y$
- fraction of vertical occupancy: the ratio of the region size to the silhouette area between the region's smallest and largest $x$
- fraction of contour points 1: the fraction of points on the region contour that are at most 5 pixels away from the silhouette contour
- fraction of contour points 2: the fraction of points on the silhouette contour that are at most 5 pixels away from the region contour
- local color contrast: $\chi^2$ distance between the color histogram of the region and the color histogram of the silhouette pixels in a bounding box four times larger than the region bounding box (like the CC cue from [2] but projected on the silhouette)

Note that due to different video resolutions, the 5 pixel threshold represents roughly the same quantity relative to the silhouette height.

## 5. A Multiple Instance Framework for Learning a Model for Carried Object Regions

One of the typical ways to apply MIL to computer vision is to treat images as bags and their segments as instances. In our framework, the instances are still image regions but the bags are sets of regions produced by the low level detectors in human track intervals annotated as "carry" or "walk". The label "carry" means that the walking human has at least one visible object in some frames of the interval and "walk" means no object visible. The annotations are independent of region detector output, so a slight complication arises that some bags labeled positive may not contain any positive instances at all due to low-level detectors failing to retrieve carried objects. However, a more important aspect is problem size: the smallest problem in this work has approximately $12,000$ instances, about twice more than the well known MIL dataset MUSK-2, and the largest is approximately $192,000$, two orders of magnitude larger. Another difficulty is that the union of positive bags has $51\%$ to $85\%$ of the training instances while the fraction of actual positives is between $7\%$ and $14\%$ of the training instances. (Note that the latter is different from the expected fraction of actual positives in each positive bag, $25\%$.) Learning an instance level classifier requires overriding bag labels for large numbers of instances in positive bags with the support of a limited number of known negatives.

Numerous MIL methods assume the existence of a few prototypical positive instances common to many positive bags and/or a meaningful Euclidian distance, assumptions which do not hold for our datasets. The most suitable approach is miSVM [3], which iterates two steps: (1) compute

**Algorithm 1** miSVM-Positive Fraction Shift

**input** : instances, bags, bag labels; $T$, $\alpha_0$, $\theta$
label all instances with their bag labels
**for** $i = 1 \rightarrow T$ **do**
  compute separating hyperplane with SVM solver
  compute decision values for instances in positive bags
  **for** each positive bag **do**
    $\alpha \leftarrow$ fraction of instances with decision value $\geq 0$
    order the instances by decision values
    relabel top $(1-\theta)\alpha + \theta\alpha_0$ instances as positive
    relabel rest of bag instances as negative
  **end for**
**end for**
**return** separating hyperplane computed with SVM solver for current labels

the separating hyperplane given all instance labels (initialized with bag labels) and (2) relabel the instances in positive bags according to the current separating hyperplane, correcting so that each positive bag has at least one positive instance. A characteristic of our problem is that the low level detectors produce a fraction of correct regions close to $\alpha_0 = 0.25$ when the person carries an object, so we adapt miSVM to reflect an expectation of the fraction of positives in positive bags, see Algorithm 1. The relabeling is now done so that the fraction of positive instances shifts towards $\alpha_0$ and we call this extension miSVM-Positive Fraction Shift (miSVM-PFS).

The algorithm minimizes the modified SVM objective

$$\begin{aligned}\mathscr{L}(w, b, y_{1..N_+}) &= \frac{1}{2}||w||_2^2 \\ &+C_1 \sum_{i=1}^{N} max(0, 1 - y_i(wx_i + b)) \\ &+C_2 \sum_{j=1}^{n_+} \left| \sum_{k \in B_j} \frac{y_k + 1}{2} - \alpha_0 n_j \right| \quad (4)\end{aligned}$$

where $y_{1..N_+}$ are the labels of the instances in positive bags, $N$ is the total number of instances ($N_+$ in positive bags, $N_-$ in negative bags), $n_+$ is the number of positive bags, $B_j$ are the indexes of instances in the j-th (positive) bag and $n_j = |B_j|$. In each iteration, the SVM training minimizes the sum of the first two terms over $w$ and $b$, and the subsequent instance relabeling minimizes the sum of the second and third over $y_{1..N_+}$. To see why the latter is true, consider the change in the second loss term when label $y_k$ switches:

$$\Delta\mathscr{L}_{2k} = \begin{cases} y_k \cdot 2dv_k & |dv_k| < 1 \\ y_k \cdot (dv_k + sign(dv_k)) & |dv_k| \geq 1 \end{cases} \quad (5)$$

where $dv_k = wx_k + b$. For each positive bag, minimizing the second term is achieved by switching the labels of instances with decision values of opposite sign to the old labels ($\Delta\mathscr{L}_{2k} < 0$). Any set of label changes can be composed as a set that minimizes the second term of the objective and then some other set of changes. The other set will strictly increase the second term while potentially decreasing the third, so to minimize their sum, it must include only instances with $dv$ between 0 and a threshold depending on $\alpha_0$ and $\frac{C_2}{C_1}$ (smallest $|dv|$'s). This is because $\Delta\mathscr{L}_{2k}$ is monotonically increasing in $dv_k$ and the third loss term does not depend on which labels are switched but on how many. The algorithm implements the two sets of label changes together, by sorting instances by $dv$ and relabeling them in relation to a threshold between 0 and the $dv$ of the top $\alpha_0$-th instance. Parameter $\theta$ equivalently models the effect of $\frac{C_2}{C_1}$.

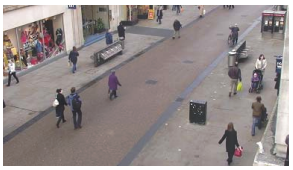
We observe that the algorithm changes very few labels after 20 iterations in all problem settings, so we fix $T$ to this value. We also set $\theta$ to 0.333. Note that $\theta = 0$ does not make our algorithm equivalent to miSVM but makes it overfit (miSVM counters overfitting by switching a label when no positives are left in a positive bag). By relabeling instances in positive bags in a controlled manner, with $\theta > 0$ bias towards fraction $\alpha_0$, miSVM-PFS smoothes the trajectory in label space and so is less likely to find local minima. The ALP-SVM version of Gehler and Chapelle's deterministic annealing approach [13] has a similar smoothing effect and employs a similar objective function, but it incurs far higher computational cost in the SVM training step because it duplicates the instances in positive bags as both positive and negative. Running ALP-SVM on a subsampled version of one of the smallest MIL problems in our datasets (thousands of instances) took over one hour while miSVM-PFS finished in under one minute; for the other problems the disparity will be much greater due to SVM training time increasing roughly quadratically with the number of instances.

## 6. Experimental Results

We ran experiments on three datasets: Pets2006, Cd2a and Towncenter, see Figure 2 for representative images. Pets2006 [12] is a well known visual surveillance benchmark that contains videos of people walking with luggage in a busy train station. For comparison with the method of Damen and Hogg [9], we ran our system on the 7 videos from the third camera. These videos range from 2,371 to 3,401 frames in length, with an average of approximately 25 people in the scene. Cd2a consists of 16 videos we selected from a corpus collected to highlight carry and exchange actions [1]. The Cd2a videos show people in varied viewpoints in two types of outdoor scenarios: a country road and a safe house. There are few object types (small packets, large boxes, duffel bags and backpacks), but people wear robes and head scarves, which complicate silhouettes. The videos are between 2,430 and 18,023 frames and

(a) Pets2006

(b) Cd2a



(c) Towncenter

Figure 2: Datasets used in the paper.

show an average of approximately 14 people. The Towncenter dataset [5] consists of a single high resolution video of a busy pedestrian-only zone near store fronts. We evaluate our approach on the first 4500 of the 7500 video frames, for which [5] provide (noisy) ground truth human bounding boxes; we annotate the objects carried by the 230 people.

We manually annotate the human tracks which are input to our method and perform training and testing on image regions detected in the parts of the tracks where humans move. These are annotated as "walk" or "carry" according to object presence and also used by the MIL version of our approach. Note that these two settings reflect the scope of our method: given human tracks, the goal is to detect objects carried by walking people, as is done in prior work [9] [16]. Each region feature is normalized by subtracting its training set mean and dividing by standard deviation. For classification, we use libsvm [7] with a Gaussian kernel with $\sigma$ the mean of pairwise distances between instances in the training set.

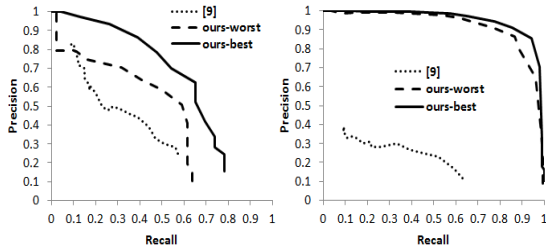
We compare our method against Damen and Hogg's [9]; note that a comparison with the more recent method due to Mitzel et al [22] is not meaningful because they use video and depth data. [9] evaluate carried object detections with a criterion requiring that the bounding box of a detected region overlap at least $15\%$ with a ground truth object bounding box. The threshold is much lower than typically used in human detection ($50\%$) in order to recognize correct matches when the protrusion is small, but this has a serious flaw – a method can return random large parts of the human silhouette and score high when most people carry objects. We remedy the criterion: a detected region is correct if it covers at least $20\%$ of a ground truth object bounding box **and** at least $66.6\%$ of its area is inside the box. We measure the performance of carried object detection methods in terms of region precision and of object track recall. Precision is defined as the fraction of regions (out of all regions eventually returned) that match ground truth and recall as the fraction of object tracks (out of all object tracks) for which there are correct detections in at least $10\%$ of the frames. We perform non-maximum suppression by removing any region that has high pixel mask overlap with another region with higher detection score. The low recall threshold allows detections to be sparse in time (3/s), but since our method is very precise, a blob tracking extension of it can achieve both high frame level recall and high precision.

### 6.1. Fully Supervised Learning

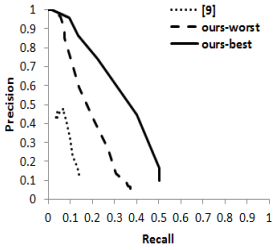
Precisely annotated object bounding boxes determine labels for the image regions from the low level detectors: a region is positive if and only if it matches a box by the criteria in the preceding paragraph. In the fully supervised setting, the label of each training region is given. We perform cross validation experiments on all three datasets. We randomly divide the sets of videos 10 times into roughly half for training and half for testing; for Towncenter, the split was on persons. For each of the three datasets, Figure 3 shows two precision-recall curves: the curve with the smallest and the largest area among the 10 splits. The curves were obtained with $C = 100$ for the kernel SVM for all datasets; values 10 and 1000 virtually did not change any of the Pets2006 curves and two curves of Cd2a, showing no need for cross validation. The Towncenter video is especially hard because people walk in all directions, wear very diverse clothing, have vastly different body builds, and carry many types of baggage, all while the number of training regions is about twice that for Pets2006. Also, it is difficult to obtain accurate foreground masks as the scene is densely populated.

To compare against Damen and Hogg's [9] full method (using spatial prior and continuity) we modify the code made public by the authors to return correctly aligned carried object pixel masks for all video frames. As was done in [9], we vary parameter $\lambda$ representing the pairwise cost in an MRF-based segmentation and trace the PR curves displayed in dotted black in Figure 3. Their method tends to return large parts of the human silhouette together with the carried object, which is significantly less precise than our method. A qualitative analysis complementing numerical results can be found in the supplemental material, which we urge reviewers to consult.

Given large differences between the three datasets in the appearance of people and objects, it is legitimate to doubt that a model learned on one dataset would work well on the other two, but experiments in which we train on a complete dataset and test on the others highlight the generalization capability of our models, see Figure 4. The PR curves are below those obtained when training and testing on subsets of the same dataset (Figure 3), but good precision-recall values are achieved when we train on people wearing tight clothing and test on people wearing robes (Pets2006→Cd2a) or when we train with 4 object types and then test on more

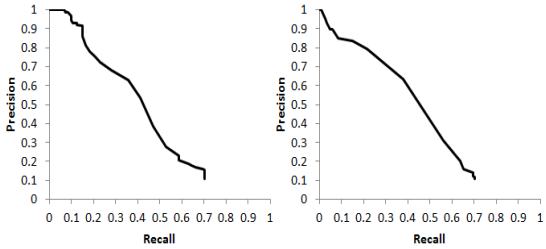



(a) Pets2006 (b) Cd2a

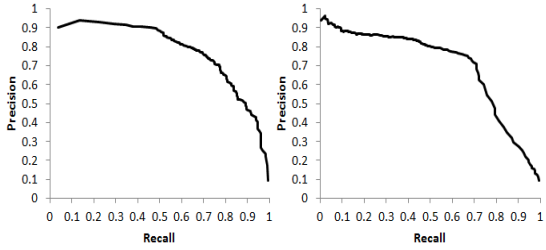



(c) Towncenter

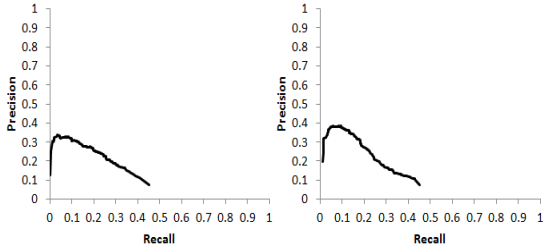
Figure 3: Pairs of representative precision-recall curves for the fully supervised version of our method (solid and dashed) and the curve for [9] (dotted). The solid curve has largest area among the curves obtained on the 10 training-test splits, dashed smallest. The curves for [9] do not extend right more than shown; in particular, on Pets2006 [9] does not obtain more than 0.57 recall.




(a) Pets2006 (Cd2a) (b) Pets2006 (Towncenter)




(c) Cd2a (Pets2006) (d) Cd2a (Towncenter)




(e) Towncenter (Pets2006) (f) Towncenter (Cd2a)

Figure 4: Precision-recall curves when training and testing on different datasets. Format: test dataset (training dataset).

than 10 object types (Cd2a→Pets2006), for example. The models learned on Pets2006 and on Cd2a (Figure 4, e and f) perform poorly on Towncenter but there is strong reason to believe this is because Towncenter is more complex than Pets2006 and Cd2a: the model learned on the former tests well on both latter datasets (Figure 4, b and d).

### 6.2. Multiple Instance Learning

In this setting, the only supervision is labels "carry" and "walk" associated with intervals of human tracks, partitioning the image regions retrieved by low level detectors into positive and negative bags. We compare the peformance of miSVM [3], our extension miSVM-PFS ($\alpha_0 = 0.25$), sMIL and sbMIL [6] in Table 1, which includes the results of a fully supervised SVM (labels available for each image region) for reference. We use the same 10 training-test splits as in the fully supervised experiments and report the mean area under the PR curve. In many of the splits the total number $N_+$ of instances in positive bags is larger than the total number $N_-$ of instances in negative bags (sometimes drastically so), biasing classifiers towards false positives. In the SVM formulations of both miSVM and miSVM-PFS, we kept the weights of the instances in positive bags 1 and assigned weights $\frac{N_+}{N_-}$ to instances in negative bags. We set $C$ to 1 for both miSVM and miSVM-PFS; other values produce little change in the relative peformance of the two.

Note that since some positively labeled bags may not actually contain any positive instances due to low level detectors failing to find any object regions, it is inappropriate to set $C$ by bag-based cross validation. For sMIL and sbMIL, we report the best mean area under PR curve over a number of parameter combinations. In particular, $\eta$ in sbMIL for the expected positive bag density ($\alpha_0$ in our work) varied in the set $\{0.1, 0.25, 0.5\}$. Table 1 shows the effectiveness of miSVM-PFS compared to other approaches on Pets2006 and Cd2a. All approaches perform poorly on Towncenter, confirming the difficulty of this dataset. sbMIL is slightly better on Towncenter; we attribute this to imbalance in positive bag densities (many more values close to 0 and 0.5 than to $\alpha_0 = 0.25$) due to errors in background subtraction.

The training times obtained on an Intel Core2 Quad at 3GHz for miSVM and miSVM-PFS are very similar. On Cd2a, the largest dataset, the training time averaged over the 10 splits was 20.7 minutes for miSVM and 16.3 minutes for miSVM-PFS; the average training set size over the 10 splits was 154,000 instances. The two algorithms both took about 1 minute on Pets2006 and about 10 minutes on Cd2a per training set respectively.

| | Pets2006 | Cd2a | Towncenter |
|---|---|---|---|
| SVM (fully supervised) | 0.5238 | 0.9410 | 0.2488 |
| miSVM | 0.3526 | 0.8158 | 0.0881 |
| miSVM-PFS | 0.4098 | 0.8496 | 0.0971 |
| sMIL | 0.1413 | 0.3031 | 0.0507 |
| sbMIL | 0.3086 | 0.6878 | 0.1019 |

Table 1: Mean area under PR curve for different learning methods. The second row of the table shows results when object bounding boxes are available, while for the other rows only "carry" and "walk" information is given.

| dataset$\backslash\alpha_0$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|
| Pets2006 | 0.2481 | 0.3791 | 0.4298 | 0.4159 | 0.3933 |
| Cd2a | 0.7065 | 0.8370 | 0.8508 | 0.8325 | 0.8221 |
| Towncenter | 0.0552 | 0.0832 | 0.0903 | 0.0959 | 0.0963 |

Table 2: Mean area under PR curve when the expected positive bag density in miSVM-PFS is varied.

Because the positive bag density is only approximately known, we characterize the sensitivity of miSVM-PFS to parameter $\alpha_0$. Table 2 shows good mean area under the PR curve for the range $[0.1, 0.5]$; on Pets2006 and Cd2a we still outperform competing approaches, suggesting that an accurate estimate of $\alpha_0$ is not critical.

## 7. Conclusion

We proposed a learning-based method for carried object detection that finds objects even when they do not protrude, achieves high accuracy, and has good generalization capabilities. Our method obtains candidate image regions from three cues (protrusions, color contrast and occlusion boundaries) and selects the plausible object regions with a kernel SVM classifier on features characterizing the context of the human silhouette. To avoid annotating tens of thousands of carried object bounding boxes, we investigated training the classifier in a MIL framework which only required hundreds of "walk" and "carry" labels for intervals of human tracks. We extended the miSVM algorithm [3] to effectively account for a known fraction of positive instances in positive bags and this extension consistently improved accuracy while keeping computational cost low.

## References


[1] `www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx`.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, 2012.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[5] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.

[6] R. C. Bunescu and R. J. Mooney. Multiple instance learning for sparse positive bags. In *ICML*, 2007.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[8] Y. Chen, J. Bi, and J. Z. Wang. Miles: Mil via embedded instance selection. *PAMI*, 2006.

[9] D. Damen and D. Hogg. Detecting carried objects from sequences of walking pedestrians. *PAMI*, 2012.

[10] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *AI*, 89(1-2):31–71, 1997.

[11] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.

[12] J. Ferryman. *WPETS*, 2006.

[13] P. V. Gehler and O. Chapelle. Deterministic annealing for multiple-instance learning. *JMLR*, 2:123–130, 2007.

[14] N. Ghanem and L. Davis. Human appearance change detection. In *ICIAP*, 2007.

[15] Y. Han, Q. Tao, and J. Wang. Avoiding false positive in multi-instance learning. In *NIPS*, 2010.

[16] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis. *Backpack*: Detection of people carrying objects using silhouettes. *CVIU*, 81(3):385–397, 2001.

[17] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. S. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 2005.

[18] C.-S. Lee and A. Elgammal. Carrying object detection using pose preserving dynamic shape models. In *AMDO*, 2006.

[19] W. Li, L. Duan, I. W.-H. Tsang, and D. Xu. Batch mode adaptive multiple instance learning for computer vision tasks. In *CVPR*, 2012.

[20] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang. Text-based image retrieval using progressive mil. In *ICCV*, 2011.

[21] O. Maron and T. Lozano-Prez. A framework for multiple-instance learning. In *NIPS*, 1998.

[22] D. Mitzel and B. Leibe. Taking mobile multi-object tracking to the next level. In *ECCV*, 2012.

[23] T. Senst, A. Kuhn, H. Theisel, and T. Sikora. Detecting people carrying objects utilizing lagrangian dynamics. In *AVSS*, 2012.

[24] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010.

[25] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

[26] D. Tao, X. Li, X. Wu, and S. J. Maybank. Human carrying status in visual surveillance. In *CVPR*, 2006.

[27] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In *CVPR*, 2008.

[28] W.-S. Zheng, S. Gong, and T. Xiang. Quantifying contextual information for object detection. In *ICCV*, 2009.