Improving an Object Detector and Extracting Regions using Superpixels

Guang Shu, Afshin Dehghan, Mubarak Shah
Computer Vision Lab, University of Central Florida
{gshu, adehghan, shah}@eecs.ucf.edu

Abstract

We propose an approach to improve the detection performance of a generic detector when it is applied to a particular video. The performance of offline-trained objects detectors are usually degraded in unconstrained video environments due to variant illuminations, backgrounds and camera viewpoints. Moreover, most object detectors are trained using Haar-like features or gradient features but ignore video specific features like consistent color patterns. In our approach, we apply a Superpixel-based Bag-of-Words (BoW) model to iteratively refine the output of a generic detector. Compared to other related work, our method builds a video-specific detector using superpixels, hence it can handle the problem of appearance variation. Most importantly, using Conditional Random Field (CRF) along with our super pixel-based BoW model, we develop and algorithm to segment the object from the background . Therefore our method generates an output of the exact object regions instead of the bounding boxes generated by most detectors. In general, our method takes detection bounding boxes of a generic detector as input and generates the detection output with higher average precision and precise object regions. The experiments on four recent datasets demonstrate the effectiveness of our approach and significantly improves the state-of-art detector by 5-16% *in average precision.*

1. Introduction

With the prevalence of video recording devices nowadays, the demand of automatically detecting objects in videos has significantly increased. The state-of-art object detectors [5, 17, 7] have achieved satisfying performance when detecting objects on static images, however, their performance on a particular video is limited for two reasons: first, most detectors are trained off-line using a fixed set of training examples, which cannot cover all the unconstrained video environments with variant illuminations, backgrounds and camera viewpoints. And it is very expensive to manually label examples and re-train the detector for each new video. Second, most detectors are designed for a generic object class using Histograms of Gradient (HOG) [5] or Haar-like features [13] . When applied on a particular video, they are not able to fully leverage the information presented in different frames of the video such as the consistent color pattern of objects and background. In this paper we aim to improve the object detector's performance in these two aspects.

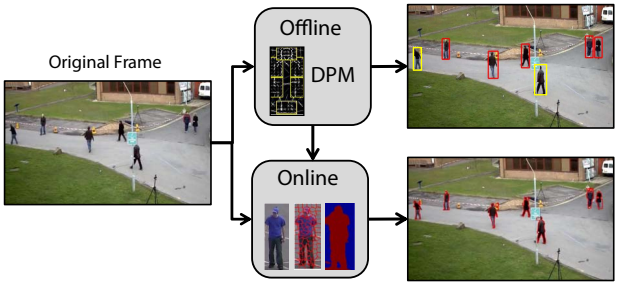

Figure 1. This figure shows how our approach improves the detection results of DPM detector. The left image is the input video frame. In the upper right image, the red bounding boxes show the detection results of the DPM and the yellow bounding boxes show the miss detections. The lower right image shows the results after refinement using our approach, in that all the objects are correctly detected and the object regions are extracted.

There has been a substantial amount of work that addresses the problem of learning from unlabeled data in a semi-supervised fashion [10, 11, 12, 4, 17, 19, 14]. A common technique of these approaches is to apply a coarse detector to the video and get initial detections, which are then added into the training set to improve the coarse detector. While these approaches have proven effective, they can only adapt their appearance models based on the coarse detections and so are not truly adaptive. On the other hand, detection-by-tracking approaches[18, 15, 3, 9] use trackers to improve the detection results for a particular video. However, they may introduce more noise to the detection results if the tracker is not reliable for the video.

To address the above mentioned problems, we propose the use of a online-learned appearance model to iteratively refine a generic detector. The intuition is that in a typi-

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.477

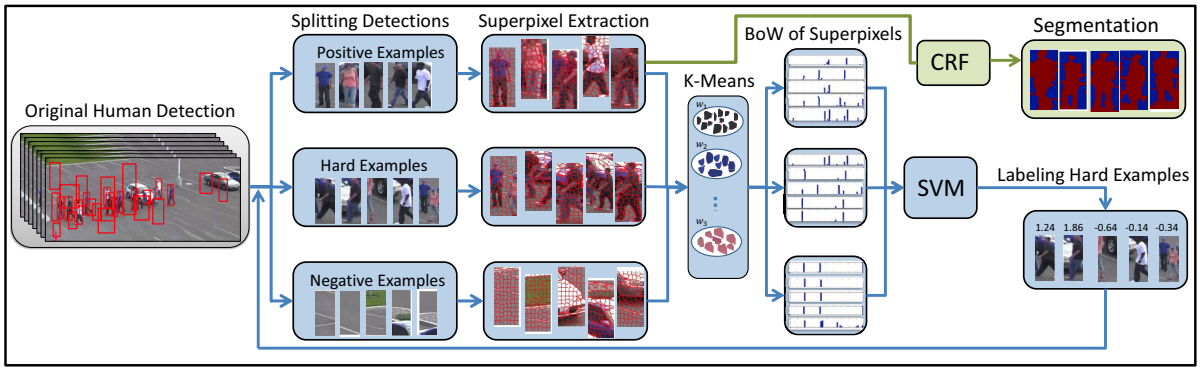

Figure 2. The framework of our approach.

cal surveillance video, each individual usually moves from one side of the scene to the other side, therefore typically many instances in the spatial-temporal domain are captured. These instances have variant poses but consistent color features. Based on this assumption, we transfer the knowledge from the generic detector to a more video-specific detector by using the superpixel-based appearance model.

The overview of our approach is illustrated in Figure 2. First we apply the original detector with a low detection threshold on every frame of a video and obtain a substantial amount of detection examples. Those examples are initially labeled as positive or hard by their confidences. Negative examples are collected automatically from background. Second, we extract superpixel features from all examples and make a Bag-of-Word representation for each example. In the last step, we train a SVM model with positive and negative examples and label the hard examples iteratively. Each time a small number of hard examples are conservatively added into the training set until the iterations converge.

Superpixels have been successfully applied in image segmentation [1], object localization [8] and tracking [16]. As the middle-level feature, Superpixels enable us to measure the feature statistics on a semantically meaningful subregion rather than individual pixels which can be brittle. On the other hand, the superpixels have great flexibility which avoids the mis-alignment of the HOG and Haar-like features on variant poses of objects.

Using the mentioned advantages of superpixels along with the proposed algorithm, we also extract the regions of objects, as shown in Figure 1. A confidence map which shows the likelihood of each pixel belonging to the target is made using a background generic model. Later CRF is employed to obtain a smooth object boundaries. Different from any background subtraction method, our method requires no background modeling, hence it is not sensitive to camera motion and will still work with a moving camera. In general, our algorithm can extract the object regions without prior knowledge of the object's shape, and the output could serve a more precise initialization for other applications such as tracking and recognition.

In this paper we take pedestrian detection as an example to illustrate our approach. The rest of the paper is organized as follows. In section 2, we introduce the related work. The details of our approach is presented in section 3. Section 4, shows experimental results on four challenging datasets. Finally section 6 concludes the paper.

2. Related Work

A substantial amount of work has been reported for building online learning approaches for object detection. Levin et.al [10] built a semi-supervised learning system using co-training, in which two different classifiers are used to train each other to improve the detection performance. In [12] a co-training based approach is proposed to continuously label incoming data and use it for online updates of the boosted classifier. However, both approaches require a number of manually labeled examples as the initial training examples. [11] presented a framework that can automatically label data and learns the classifier for detecting moving objects from video. Celik et.al [4] proposed an approach to automatically detect dominant objects for visual surveillance. However, in [11] and [4] the initial coarse detectors are based on background subtraction, hence they don't fit into the scenarios with complex background or moving camera.

Wang et.al [17] proposed a non-parametric transfer learning approach, in which they use a vocabulary tree to encode each example into binary codes. Wang et.al [17] only learns objects having the similar appearance to the initial examples. These approaches are likely to miss some hard examples with large appearance variations. On the other hand, the detection-by-tracking approaches improve the detections by using trackers [18, 15, 3, 9]. In these methods, object hypothesis are detected in all frames and then associated by trackers. By using the structural informa-

tion of the scenario, the tracker may find even the occluded objects. However, tracking itself is a challenging problem. If the tracker is not suitable to the scenario, it may lower the performance of the original detector.

Our solution lies between of the semi-supervised learning and detection-by-tracking. Compared to semi-supervised learning methods that learn a generic detector, Our method learns a video-specific detector that leverages the consistent color patterns in frames of a video; though it is more conservative than detection-by-tracking methods and will not introduce additional problems caused by the tracker itself. Moreover, our method obtains regions of objects, hence it provides more improved performance.

3. Our Approach

3.1. Initial Detection

We employ the deformable part-based model detector (DPM) [7] as the initial detector in our approach since it has shown excellent performance in static images. The detector is given a lower detection threshold t_d so we can obtain almost all true detections and a large amount of false alarms. According to the detector's confidence scores, we initially split all detections into two groups: the ones with confidence scores above a threshold are labeled as the positive examples; the rest are labeled as hard examples. In addition, a large number of negative examples are randomly collected in a way that they do not overlap with any positive or hard examples. All the examples are then resized to 128×64.

3.2. Superpixels and Appearance Model

Most object detectors use HOG or Haar-like features which can represent a generic object class well. However, they are not robust to different poses of an individual. As shown in Figure 3, an individual can have variant poses that renders a mis-alignment for HOG, Haar-like features or any other pixel-level features. To handle this problem, we need to transfer the knowledge from a generic detector to a video-specific detector. Therefore we build a statistics appearance model with the superpixels as units.

Figure 3. Variant poses in different frames.

We segment each detection output into N_{sp} superpixels by using the SLIC Superpixels segmentation algorithm in [1]. We choose an appropriate number N so that one superpixel is roughly uniform in color and naturally preserves the boundaries of objects. In order to encode both color and spatial information into superpixels, we describe each superpixel $Sp(i)$ by a 5-dimensional feature vector $f = (L, a, b, x, y)$, in which (L, a, b) is the average $CIELAB$ colorspace value of all pixels and (x, y) is the average location of all pixels.

An M-word vocabulary is assembled by clustering all the superpixels using the K-means algorithm. Then the superpixels are aggregated into an M-bin L2-normalized histogram for each example and later each example is represented in a BoW fashion.

3.3. Classification

We train a support vector machine (SVM) to classify the hard examples. Due to the limitation of the initial training samples, it is not easy to obtain a good decision boundary. Therefore we use an iterative way to gradually update the SVM model. After each iteration we obtain SVM scores for the hard examples, and we split the hard examples into three groups again. We move the examples with high scores into positive set and the examples with low scores into negative set, then re-train the SVM model. In this way, the decision boundary is gradually refined for each iteration. We will repeat this process until all the example labels are unchanged. In our experiments, it usually takes 5 to 7 iterations. After all the hard examples are labeled, we can project the positive examples back into the image sequence and generate the detection output.

3.4. Region Extraction

The superpixel-based appearance model enables us not only to improve the detector, but also to precisely extract the regions of objects. Since the superpixels can naturally preserve the boundary of objects, we develop an algorithm that takes the detection bounding box as input and calculate a confidence map indicating how likely each superpixel belongs to the target.

First, we cluster all superpixels of the negative samples into M_n clusters by $CIELAB$ color features. Each cluster $clst(j)$ is represented by the its center. Then we calculate the similarities between all superpixels from positive examples and all the clusters. The similarity is measured by the Equation

$$Sp(i,j) = \exp(\|Sp(i) - clst(j)\| \times prior(j)), \qquad (1)$$

in which $S_p(i)$ is the i-th superpixel from positive examples and $clst(j)$ is the j-th cluster center. They are all represented by color values. $prior(j)$ is the prior probability that j-th cluster belongs to the background; this is defined by the number of superpixels in the j-th cluster. The $prior(j)$ is used here as a regularizing term which discourages the small cluster of background. After obtaining the

similarity matrix $W(i,j)$, we can calculate the confidence of a superpixel belonging to the target by the equation

$$Q(i) = 1 - \max_j W(i,j). \quad (2)$$

Therefore we can obtain a confidence map Q for each positive example, as shown in the second row of Figure 4.

In order to extract a precise region in the confidence map, a conditional random field (CRF) model [2] is utilized to learn the conditional distribution over the class labeling. CRF allows us to incorporate constraints in the pairwise edge potentials and hence improve the classification accuracy around the boundary. Let $Pr(\mathbf{c}|G;\omega)$ be the conditional probability of the class label assignments c given the graph $G(Sp, Edge)$ and a weight ω, We need to minimize the energy equation

$$-\log(Pr(\mathbf{c}|G;\omega)) = \sum_{s_i \in Sp} \Psi(c_i|s_i) + \omega \sum_{s_i,s_j \in Edge} \Phi(c_i,c_j|s_i,s_j), \quad (3)$$

where Ψ is the unary potentials defined by the probability provided by the confidence map Q:

$$\Psi(c_i|s_i) = -\log(Pr(c_i|s_i)), \quad (4)$$

and Φ is the pairwise edge potentials defined by

$$\Phi(c_i,c_j|s_i,s_j) = \left(\frac{1}{1+\|Sp(i)-Sp(j)\|}\right)[c(i) \neq c(j)], \quad (5)$$

where $[.]$ is the one-zero indicator function and $\|Sp(i) - Sp(j)\|$ is the $L2$-norm of color difference between superpixels. After the CRF segmentation we will obtain a binary map on which the target and background is distinctly separated. Note that in some positive examples, there are usually some superpixels which belong to other near targets labeled as target. We show some segmentation results in Figure 4. more examples are shown in Figure 8(b).

4. Experiments

We extensively experimented on the proposed method using four dataset: Pets2009, Oxford Town Center [3], PNNL-Parking Lot [15] and our own Skateborading sequences. The experimental datasets provide a wide range of significant challenges including occlusion, camera motion, crowded scenes and cluttered background. In all the sequences, we only use the visual information and do not use any scene knowledge such as the camera calibration or the static obstacles.

We compare our method with the original DPM detector. We also compare the superpixel-based appearance model (SP) with HOG within our online-learning framework. In the HOG implementation, each detection window is represented by a standard 3780-dimensional feature vector as in [5]; the other steps are identical to our proposed approach.

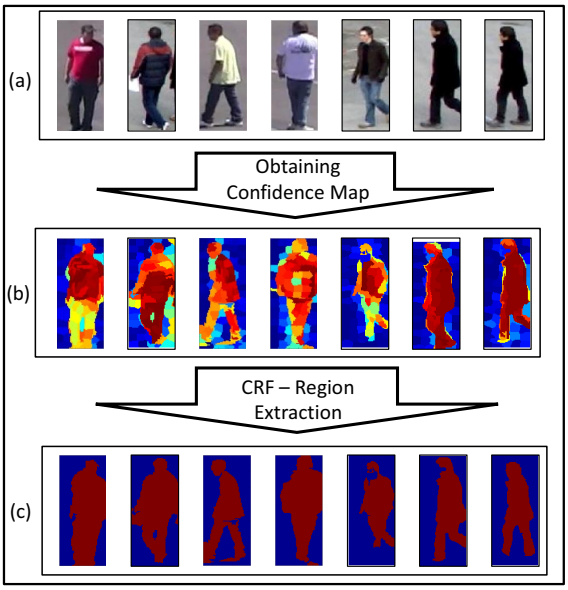

Figure 4. Examples of CRF segmentation. The first row shows some pedestrian examples, the second row shows the corresponding confidence maps and the last row shows the corresponding CRF segmentations.

We use the criterion of the PASCAL VOC challenge [6] for evaluations in our experiments. In that, a detection that has more than 0.5 overlap with the groundtruth is determined as true positive. We analyze the detector performance by computing Precision-Recall curves for all four datasets, as shown in Figure 6.

In our implementation, we use the pre-trained pedestrian model from [7]. We set a detection threshold $t_d = -2$ to achieve a high recall. For the superpixel segmentation, we set the number of superpixels for each examples to$N_{sp} = 100$. In the K-mean clustering we set the number of clusters $M = 400$. In the region extraction we set the number of negative clusters to be $M_n = 200$.

Our approach has achieved better performance in all four datasets. We also calculate the average precision for quantitative comparison which is used in [6]. The AP summarizes the characteristics of the Precision-Recall curve in one number and is defined as the mean precision at a set of equally spaced recall levels. We chose the levels to be the same as [6] to represent the overall performance on each video, as shown in Table 1.

In addition, the computational cost of our approach is relatively low. While the initial detector takes around 15 second for each frame, our additional steps takes only 3 seconds on average for each frame with a 3GHz CPU.

Finally, we show some qualitative results of our experiments. Figure 5 and 7 show the detection results; figure 8(a) shows the region extraction results. We analyze the results on the four datasets in the following.

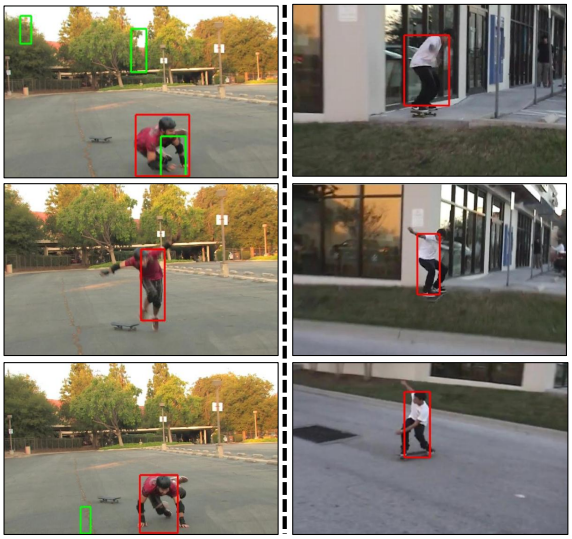

Figure 5. Detection results of the Skateboarding dataset. The green bounding boxes are the output by DPM detector; the red bounding boxes are the output by our approach. It is clear that our approach has fewer false positives as well as false negatives.

Skateborading Dataset (SB): This dataset consist of two video sequences captured by a hand-held camera. This is a very challenging dataset due to the camera motion and severe pose changes. Table 1 and Figure5 show our approach performs significantly better than the original DPM detector.

PNNL Parking lot Dataset (PL): This dataset consists of two video sequences collected in a parking lot using a static camera. Parking lot 1 is a moderately crowded scene including groups of pedestrians walking in queues with parallel motion and similar appearance. Parking lot 2 is a more challenging sequence due to the large amounts of pose variations and occlusions, hence the results on this dataset are lower than other datasets. However, our approach still performs significantly better than the DPM detector and HOG feature.

Town Center Dataset (TC): This is a semi-crowded sequence with rare long-term occlusions. The motion of pedestrians is often linear and predictable. However, it is still quite challenging because of the inter-object occlusions and the difficulty of detecting pedestrians on a bike or with a stroller. Table 1 and 2 show that we outperform the DPM detector both in precision and average precision by significant margin.

Pets2009 Dataset (PT): This is a relatively sparse scene including a few people walking in random directions. We upsampled the resolution of this video by 2 because the original low resolution is not suitable for the DPM detector. The original detector has already achieved satisfying results but our approach performs even better.

Table 1. Average Precision on our testing datasets. Second row shows the results of our method using HOG as descriptors and the third row shows the proposed method using bog of words of superpixels.

Dataset	PL1	PL2	TC	PT	SB1	SB2
Orig	86.4	55.1	86.9	93.7	59.9	70.6
Ours-HOG	91.6	66.1	93.6	97.9	73.8	83.3
Ours-SP	**93.0**	**67.6**	**94.7**	**98.0**	**75.8**	**85.6**

Table 2. The Precision = Recall points for our experiments on four different datasets.

Dataset	PL1	PL2	TC	PT	SB1	SB2
Orig	87.0	56.0	83.6	92.9	60.4	69.6
Ours-HOG	88.7	62.0	84.7	94.5	67.9	78.3
Ours-SP	**90.6**	**64.2**	**91.7**	**96.1**	**69.4**	**82.9**

5. Conclusion

We proposed an effective method to improve generic detectors and extract object regions using a superpixels-based Bag-of-Words model. Our method captures rich information about individuals by superpixels; hence it is highly discriminative and robust against appearance changes. We employ a part-based human detector to obtain initial labels and gradually refine the detections in a iterative way. We also present a region extraction algorithm that extracts the regions of objects. We demonstrated by experiments that our method effectively improves the performance of object detectors in four recent datasets.

6. Acknowledgment

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-09-1-0255.

References

[1] R. Achanta, A. Shaji, P. F. Kevin Smith, Aurelien Lucchi, and S. Susstrunk. Slic superpixels. In *EPFL Technical Report*, 2010.

[2] R. N. B. Yang, C. Huang. Segmentation of objects in a detection window by nonparametric inhomogeneous crfs. In *Computer Vision and Image Understanding*, 2011.

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

[4] H. Celik, A. Hanjalic, and E. A. Hendriks. Unsupervised and simultaneous training of multiple object detectors from unlabeled surveillance video. In *Computer Vision and Image Understanding*, 2009.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

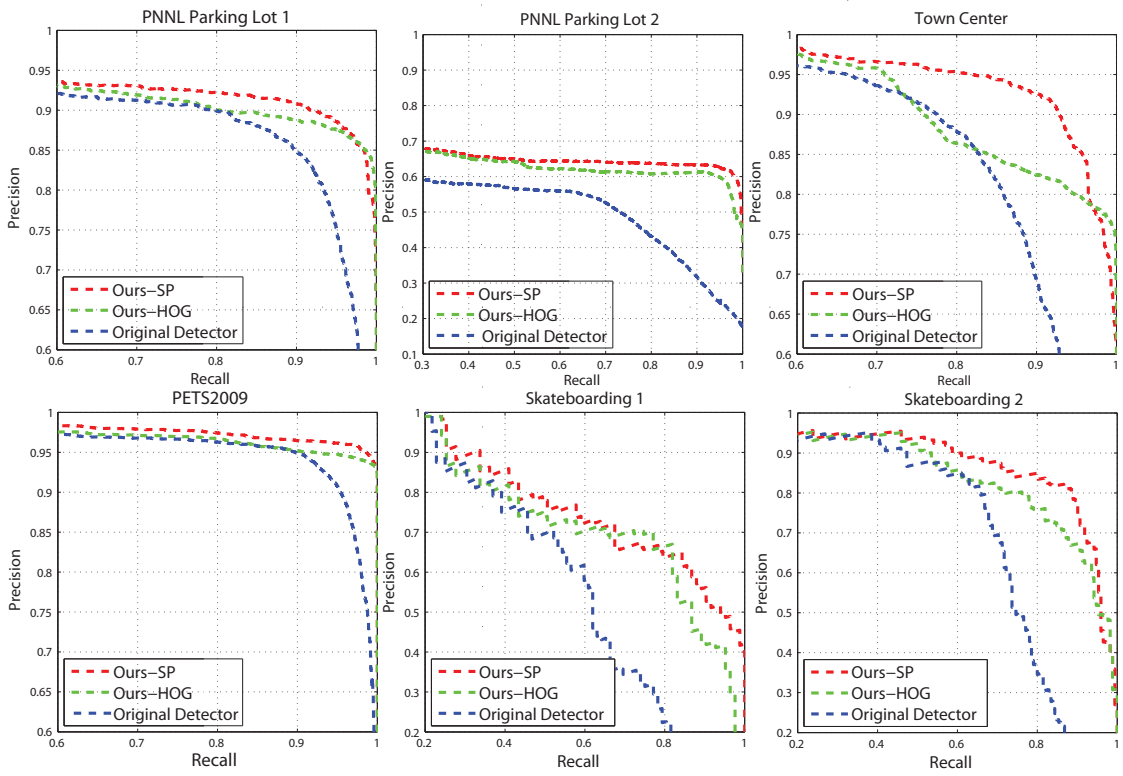

Figure 6. Performance comparison on four datasets. We compared our method against the original detector once by choosing HOG and once using bag-of-word of superpixels as the feature.

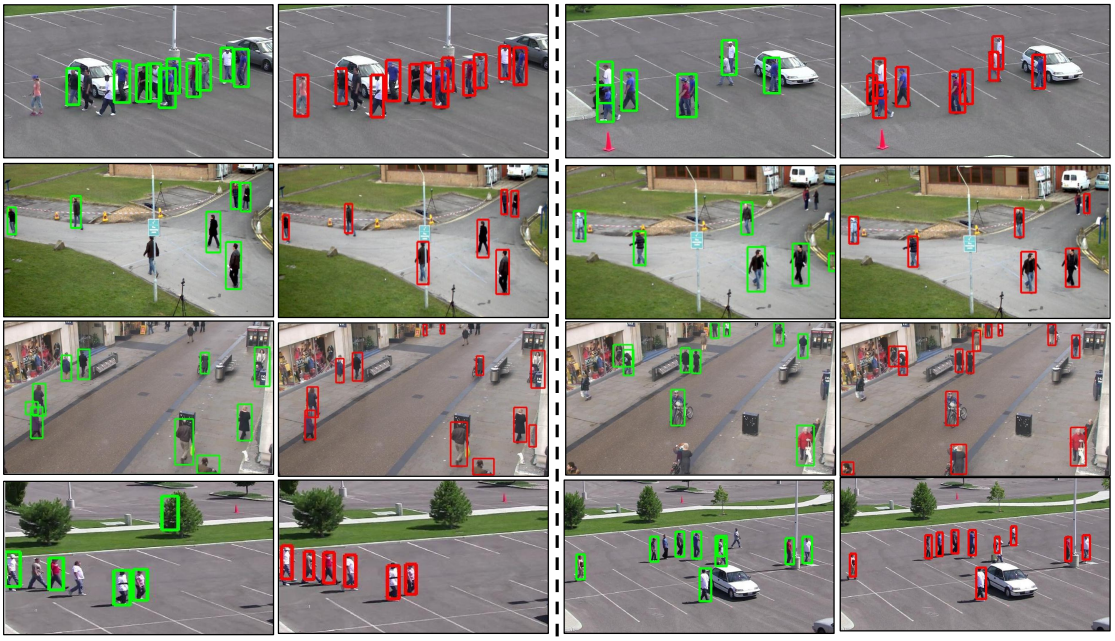

Figure 7. Detection results on videos. The datasets from the first row to the last row are: PNNL Parking Lot 1, Pets2009 Town Center and PNNL Parking Lot 2. The green bounding boxes are the output by DPM detector; the red bounding boxes are the output by our approach. It is clear that our approach has fewer false positives as well as false negatives.

Figure 8. (a) Region extraction results on datasets Parking Lot 1, Pets2009 and Parking Lot 2. We blended the extracted object region in red on the original image. (b) Individual examples. The rst row shows the original detection window; the second row shows our segmentation results using CRF.

[6] M. Everingham, V. Gool, L. Williams, C. K. I., J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. In *International Journal of Computer Vision*, 2010.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In *PAMI*, 2010.

[8] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[9] C. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 2010.

[10] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *CVPR*, 2003.

[11] V. Nair and J. Clark. An unsupervised, online learning framework for moving object detection. In *CVPR*, 2004.

[12] O.Javed, S. Ali, and M. Shah. Online detection and classification of moving objects using progresively improving detectors. In *CVPR*, 2005.

[13] P.Viola and M.Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[14] P. Roth, S. Sternig, H. Grabner, and H. Bischof. Classifier grids for robust adaptive object detection. In *CVPR*, 2009.

[15] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012.

[16] S. Wang, H. Lu, F. Yang, and M.-H. Yang. Superpixel tracking. In *ICCV*, 2011.

[17] X. Wang, G. Hua, and T. X. Han. Detection by detections: Non-parametric detector adaptation for a video. In *CVPR*, 2012.

[18] A. R. Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *ECCV*, 2012.

[19] C. Zhang, R. Hamid, and Z. Zhang. Taylor expansion based classifier adaptation: Application to person detection. In *CVPR*, 2008.