People in the Weeds: Pedestrian Detection Goes Off-road

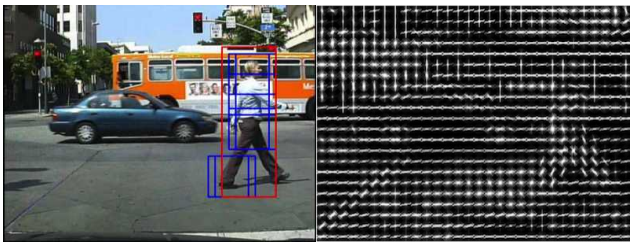
Trenton Tabor, Zachary Pezzementi, Carlos Vallespi, and Carl Wellington
National Robotics Engineering Center
Carnegie Mellon University Robotics Institute,
Pittsburgh, Pennsylvania 15201.
Email: ttabor@nrec.ri.cmu.edu

***Abstract*— Robotics offers a great opportunity to improve efficiency while also improving safety, but reliable detection of humans in off-road environments remains a key challenge. We present a person detector evaluation on a dataset collected from an autonomous tractor in an off-road environment representing challenging conditions with significant occlusion from weeds and branches as well as non-standing poses. We apply three image-only algorithms from urban pedestrian detection to better understand how well these approaches work in this domain. We evaluate the Aggregate Channel Features (ACF) and Deformable Parts Model (DPM) algorithms from the literature, as well as our own implementation of a Convolutional Neural Network (CNN). We show that the traditional performance metric used in the pedestrian detection literature is extremely sensitive to parameterization. When applied in domains like this one, where localization is challenging due to high background texture and occlusion, the choice of overlap threshold strongly affects measured performance. Using a permissive overlap threshold, we found that ACF, DPM, and CNN perform similarly overall in this domain, although they each have different failure modes.**

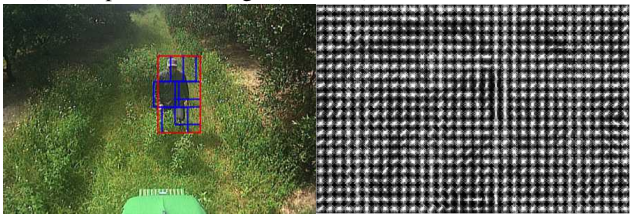
I. Introduction

Enabling the full promise of robotics in off-road environments requires reliable detection of human workers so that people and machines can effectively and safely carry out tasks together. Many machines in this domain are powerful and potentially dangerous and must be able to operate near humans for certain tasks. Some applications may need to enforce a safety buffer, and these areas often have minimal access controls. Even for smaller off-road robots, it is often important for them to understand where the people in their environment are to effectively complete their tasks. Additionally, the same techniques used for safe operation around people can be applied to enforce perimeter security, detect people in protected areas, or find people in need of assistance. This would enable the use of patrolling robots [1] in a greater range of off-road scenarios.

Our previous work resulted in a spatially distributed multi-vehicle system of autonomous tractors overseen by a remote supervisor to accomplish agricultural operations in a citrus orchard [2]. This system has demonstrated over 2400 km of autonomous operation and performed significant useful work at a higher productivity level than current methods. The system includes a sophisticated obstacle detection system, but a key limiting factor was the reliable detection of people when partially occluded by tree branches and weeds or when lying on the ground or in other non-standard poses.



(a) Sample urban image (from [3]) and its HoG visualization



(b) Sample off-road image and its HoG visualization

Fig. 1: Sample images with deformable parts model (DPM) detections and HoG visualizations. Note the difference in gradient orientation alignment between the two scenes.

Recently there has been substantial work in pedestrian detection for urban and indoor environments. However, relatively little focus has been given to the unique challenges and opportunities of human detection in off-road domains, including agriculture, outdoor rescue, surveillance, and perimeter security. As shown in Figure 1, traditional urban or indoor settings generally focus on walking or standing people that are geometrically distinct from their surroundings, whereas complex off-road environments often feature people who may be obscured by the environment in a greater variety of poses and activities. Similar occlusion and unusual poses are also major challenges in victim detection in search and rescue scenarios, where people may be covered by debris [4]. In the work presented here, people are commonly obscured by trees, weeds, crops, and other vegetation while working in or near these natural materials.

In this paper, we provide a brief review of the related work in both the off-road and urban settings and discuss some fundamental differences between the domains. We then

introduce a new dataset taken in an orchard environment and compare the performance of three pedestrian detection algorithms on this data, with a focus on the suitability of standard pedestrian detection metrics in this domain. This environment provides examples of many of the challenging conditions that safety, security, and rescue applications will need to deal with when detecting people, making it a good testbed for operation in general unstructured settings.

II. Off-Road and Urban Person Detection

A. Off-road person detection

Although there has been significant work in general off-road navigation, relatively little research has focused on the detection of people in these challenging environments. A common approach is to look for clusters of points above the ground in a LIDAR scan. This has been demonstrated in an orchard [5], but it is insufficient when people are standing next to trees or in vegetation. Pulling from the remote sensing community, the ratio of near-infrared and red camera data has been used to discriminate between vegetation and other objects [6], but this signature can be fooled by certain man-made fabrics and paints. Ray-tracing LIDAR returns through a 3D voxel representation can generate an approximate density measure, and this has been used within several autonomous systems to help operate in off-road areas with vegetation, often combining density with other appearance features within a learning framework [7], [8]. Some outdoor environments have common repetitive textures, and an anomaly detection approach can be used to find objects that have a different appearance [9].

In our previous work [2], we performed person detection using a combination of local color and texture patch features from camera images with 3D data features from a scanning LIDAR sensor or a stereo camera. As shown in Figure 2, these combined features were used to classify individual 3D voxels without additional context, so there are some voxels classified as obstacles within the trees and some tree voxels within the person. We used a spatio-temporal filter to collapse the 3D voxel classifications into a 2D obstacle map that was used to autonomously control a tractor.This system performed well when people were fully visible, but like other existing literature in the area, it was limited by its lack of contextual reasoning, especially when people were occluded by trees. Given the success of pedestrian detection in urban environments using image-only techniques that include the larger context of an entire person, we wanted to explore how these approaches would perform in an off-road application.

B. Urban pedestrian detection

There has been substantial progress in urban pedestrian detection in the last several years, motivated largely by applications in surveillance and automotive safety and autonomy. Recent survey papers summarize results across the field [3], [10], and show how work in this domain leans heavily on common labeled datasets with standard metrics, such as the Caltech Pedestrian Detection Benchmark [11] and the KITTI Benchmark [12]. In [10], the authors group approaches into three main algorithmic categories: classical rigid object detectors, deformable parts models, and the recently growing set of deep convolutional neural networks. We chose one implementation from each of these classes of detectors to apply to this domain for comparison, described further in Section IV.

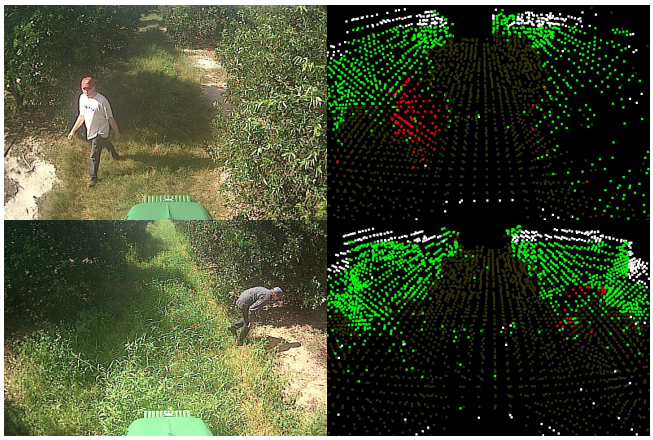


Fig. 2: Example classifications from our previous system using a combination of 3D and image features. Voxels are classified as either ground (brown), tree (green), or obstacle (red). Invalid voxels are marked white.

C. Comparison between Urban and Off-Road Domains

While the appearance of people is highly variable across both urban and off-road environments, the appearance characteristics of their surroundings have some significant differences that are likely to affect algorithms' discrimination and localization power. Visualization of the histogram of oriented gradients (HoG) features for typical images provides one view of the texture characteristics of the background in each domain, as shown in Figure 1. Urban backgrounds commonly have large regions of low texture and strongly oriented edges, particularly in vertical and horizontal directions. In contrast, backgrounds in off-road domains often have strong texture throughout the image, with high gradient energy spread out across many different orientations.

Off-road environments also can include people in more challenging poses, especially if they are bent over working with materials on the ground, such as in an agricultural, rescue, or surveillance applications.

III. Data and Labeling

Data were collected in a 1,300 hectare orange orchard in Florida, USA, described in detail in our previous work [2]. As shown in Figure 3, the tractor used for data collection includes a set of three stereo cameras mounted on the front of the roof. The side cameras are utilized to cover the sharp turns at the end of the rows, but the analysis in this paper focuses only on monocular data from the front camera. Data logs for this system contain image sequences from these cameras and other information such as 6D pose using RTK GPS, but this work considers only image-based detection.

Fig. 3: Autonomous tractor used for data collection

Logs were collected with different people in varied clothing carrying out several standard motions at different distances (with mannequins used at very close distances, for safety). The same set of standard motions and positions was used for each variation in the environment or the appearance of the person. Logs cover both static and moving people with various amounts of occlusion by trees and weeds. People are seen wearing a wide range of different types of clothing that are commonly encountered in different off-road environments, including challenging examples of a person in full camouflage.

The collected data comprise a total of 1172 logs, approximately 19 total hours of video, or nearly 1 TB of data. From that set, we selected 144 logs with people present, and we generated ground truth labels for these logs, consisting of a rectangular bounding box surrounding the visible portion of the person in each image. Labeled positive examples were then randomly divided between train and test sets at the level of individual logs. When logs were part of a set containing a person in the same outfit in several different positions or motions, then the entire set was placed in either train or test. Logs containing no obstacles were also collected, and these logs were then randomly sub-sampled to create a broad set of negative images. The training set contains 3925 labeled positive images and 2500 negative images, and the test set has 1724 labeled positive images, with the non-person background regions of these images (the vast majority of each) allowing evaluation of false-positive rates.

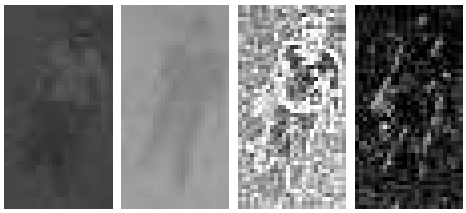
Further information on the data set is available on the project website.[1]

IV. Algorithms Compared

A. Aggregated Channel Features (ACF)

A common class of learners for general object detection is a group of weak decision trees combined with a variant of Adaboost. Decisions are made on features from fixed locations in the window, so these are considered rigid detectors [10]. As an example from this family, we evaluate Aggregated Channel Features [13], available in Piotr's Matlab Toolbox [14]. This detector uses a sliding window approach; candidate bounding boxes are considered at regular intervals throughout the image. During training, Adaboost selects from a set of features, called "channels", to generate a fixed size decision forest for classifying candidate bounding boxes.

[1] http://www.nrec.ri.cmu.edu/projects/usdapersondetection



(a) L (b) V (c) GM (d) GD2

Fig. 4: Example ACF channels for the person in Figure 7

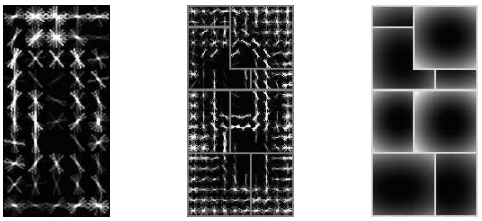
Figure 4 demonstrates some of the features used in this method, when applied to a portion of the image in Figure 7. Color features are computed by converting the image to LUV space and sampling at a low resolution. Figures 4a and 4b show the Luminance and Value color channels respectively. Gradient features at the same level of resolution provide 7 more channels: the overall magnitude (GM), shown in Figure 4c, and 6 directional magnitudes, one of which is shown in Figure 4d (GD2). This model was trained and evaluated using the pipeline from [14], using the default settings for the Caltech benchmark, but with tree depth lowered to avoid over-fitting on this dataset.

B. Deformable Parts Model (DPM)

An additional common person detector we applied to our data was the Deformable Parts Model [15], which uses a two-stage classification process to model the ability of parts of an object to move relative to each other and to the object centroid. This model consists of three elements, illustrated for our data in Figure 5. First, a "root filter", effectively a HoG detector with a weak threshold, acts as a detection proposal generator; this detector scans the whole image and generates candidate person bounding boxes. Next, a set of additional HoG detectors are applied at higher resolution. These are referred to as "part filters" and are used to scan the bounding box proposals from the first step looking for specific details that are learned from the training data. For intuition, these may correspond to a person's arms and legs. Finally, there is a set of cost functions for each of the part filters, representing how much displacement of the detected parts was seen in the training data. A final detection score is determined by considering both the goodness-of-fit of the detected parts and their distance from their canonical locations.

Our model was trained using the original pipeline designed for the PASCAL Visual Object Challenge [16]. It allows for specifying a separate negative and positive training set with bounding boxes. The model is evaluated using the same pipeline with small modifications for figure generation.

For both ACF and DPM, non-maximal suppression is performed by iterating over detections by decreasing score and suppressing all other detections with overlap, O_{D_1,D_2}, above a threshold [14], measured for a pair of detections D_1 and D_2 as $O_{D_1,D_2} = \frac{D_1 \cap D_2}{D_1 \cup D_2}$. The left images in



(a) Root filter (b) Parts filter (c) Displacements

Fig. 5: Deformable Parts Model trained with [16] on our data

Figure 7 show typical results on this dataset using the default overlap thresholds for ACF (0.65) and DPM (0.5). With their default settings, both algorithms commonly placed multiple bounding boxes on the person, resulting in significant false positive counts, because the standard metric only allows a single correct detection. In this application, overlapping people are rare, and the priority is to detect when any person is present, so we reduced the required overlap. After tuning on a small validation set independent of the test set, a non-maximal suppression overlap value of 0.2 was selected, because it produced significantly better performance than the defaults, and lowering it further had negligible effect.

C. Convolutional Neural Network (CNN)

With the recent renewed interest in Convolutional Neural Networks (CNNs), many papers have been published in the area of pedestrian detection, starting with [17] and focusing increasingly on application in real-time systems [18]. In this paper we apply our own initial implementation of a CNN. Local contrast normalization is independently applied to each color channel of the input images. The network takes as input a window of 60×45 color pixels and uses 4 convolution layers followed by pooling layers. We denote a convolution layer with k filters of size n by m by $C^k_{n\times m}$, and an average pooling layer with a pooling kernel of size n by m as $PA_{n\times m}$. Our network has the following architecture, chosen to support embedded use: $C^{10}_{5\times 4} - PA_{2\times 2} - C^{15}_{5\times 4} - PA_{2\times 2} - C^{20}_{5\times 4} - PA_{2\times 2} - C^{25}_{4\times 3}$. Leaky Rectified Linear Units (LReLU) follow all convolution layers and the last is connected to a softmax.

To train the network we grow all labeled positive training examples by 10 percent in each direction to capture additional context. We then resize all adjusted positive training example regions to 80×60, and then we extract 60×45 patches from this region at different translations and augment the data by applying horizontal flips. For the negative class, we initially use 60×45 patches extracted randomly from different scales of images without people. We randomly select 5% of the training data for a hold-out set that we use to tune the number of filters in each layer.

The network weights are initialized as described in [19], and the weights are optimized using stochastic gradient descent with mini-batches of 25 samples. The learning rate is adjusted using Adadelta [20] with $\epsilon = 10^{-6}$ and $\rho = 0.99$. Every two iterations, we augment the negative class by harvesting examples from the training set that produce errors. We repeat this process 5 times. After this, we perform 10 more iterations of fine tuning with $\epsilon = 10^{-8}$.

To convert the raw CNN responses to bounding box detections, we convolve the responses from 5 different scalings of the image (see Figure 6) with a box filter to accumulate over the local area and take high scoring responses as candidate detections. These are converted to a set of bounding boxes across all scales, which are grouped using a greedy clustering. The non-maximal suppression used for the other methods did not produce good results with these inputs, with small detections often suppressing better, larger detections. For each cluster we therefore produce a single detection as the average of the parameters for all bounding boxes (location and scale) weighted by their score multiplied by area, thereby preferring larger detections more likely to contain the entire person.

V. Performance Evaluation

Although input and ground truth data are image sequences, we follow the conventions of previous work in evaluating performance on individual images [14], without any temporal reasoning. Additionally, only the left camera image from the frontal stereo pair is used by these algorithms.

A. Performance Metric

For any choice of performance metric, we must compare algorithm detections to ground truth labels and mark each as correct or not. The standard metric used in the pedestrian literature [3] is bounding box overlap between the detection, D and the label, L, evaluated as $O_{D,L}$ as in Section IV-B. Then a threshold, T, is applied, and detections are considered correct if $O_{D,L} > T$. The standard choice of T for most benchmarks is 0.5, so we report values with this threshold to match other work. However, as discussed further in Section VI, we find that the results in this domain are very sensitive to this threshold, so we report results for other metric overlap thresholds as well.

There are many options for plotting detection performance, including popular choices like Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves. For easy comparison to the existing literature, we show performance using the same variant of the ROC as is commonly used by others in pedestrian detection: miss rate versus false positive rate on a log scale.

B. Alternate Metrics

Although the ROC curve may be helpful for evaluating expected detections in an image, it is difficult to use this curve to predict the behavior of a system that uses the detector. For this purpose, it is important to capture effects like the persistence of true and false positives over a period of time and the distance at which detections are made.

One way to evaluate such performance is to model a full autonomous system using the person detector in a safeguarding system to limit the speed of the vehicle [21]. Then statistics on vehicle speed and stops can be compared to a ground truth model of expected behavior using

Fig. 6: CNN responses for center of detection bounding box (in blue) at all 5 scales for the person in Figure 7. The true size is between the first two scales.

Fig. 7: Comparison of ACF (top, green) and DPM (bottom, red) detections using (left) default non-maximal suppression settings and (right) our modified settings

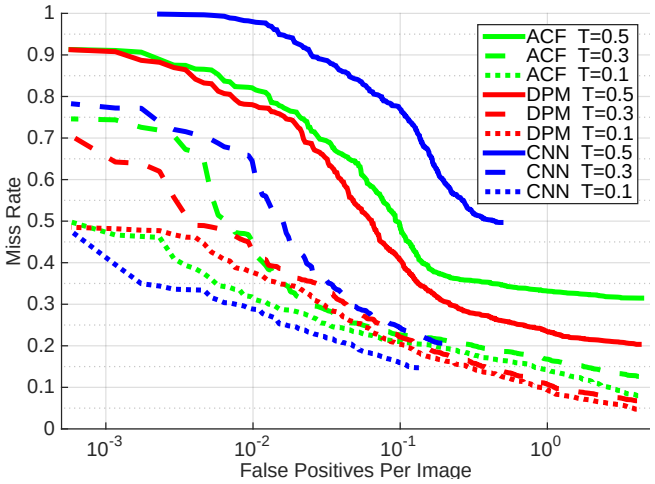


Fig. 8: Algorithm performance comparison using different performance metric overlap thresholds. Generated with [14].

(a) Detections failing $T = 0.5$ (b) Detections failing $T = 0.3$

Fig. 9: Typical person detections that are counted as false positives for different performance metric overlap thresholds. CNN detections (blue) and ground truth labels (white).

an ideal detector. This requires modeling aspects of the problem beyond pure detection (temporal filtering, vehicle dynamics, closed-loop control, etc.), but it allows for much more intuitive metrics of vehicle behavior, such as correct stops, statistics on stopping distance, false positive stop and slow-down rates, and others. However, appropriate intuitive metrics vary by application, and they may be quite different for a security or search and rescue system. We wish to avoid assuming particular dynamics or task control, but still evaluate detection with a view toward expected system-level performance. We therefore report results using the standard metric, but with requirements for localization accuracy more appropriate to the domain.

VI. Results

Each of the three algorithms was run with the training and test data described in Section III to produce independently scored detections on each image, and the resulting performance is shown in Figure 8. As shown in Figure 9, using the default performance metric overlap threshold of $T = 0.5$ counts many detections as false positives that are not perfectly localized with respect to labels (which can not themselves be assumed to be perfect, due to occlusion, interpolation, and labeler error), but visually appear to be good detections that would produce desirable behavior for a robotic system making use of them. Figure 8 shows how the scores improve for the different algorithms as we relax the requirement on precise bounding box localization. For some tasks in off-roads environments, achieving a precise bounding box is not a high priority. For instance, for many deployable systems operating in an orchard, the tractor motion is constrained, and its only option when a person is present is to slow down and stop. These results show the metric's strong sensitivity to the overlap threshold in domains like this, which have high background texture and include significant occlusion, making precise bounding box localization difficult. This is in contrast to urban pedestrian detection, where an analysis of the metric overlap threshold

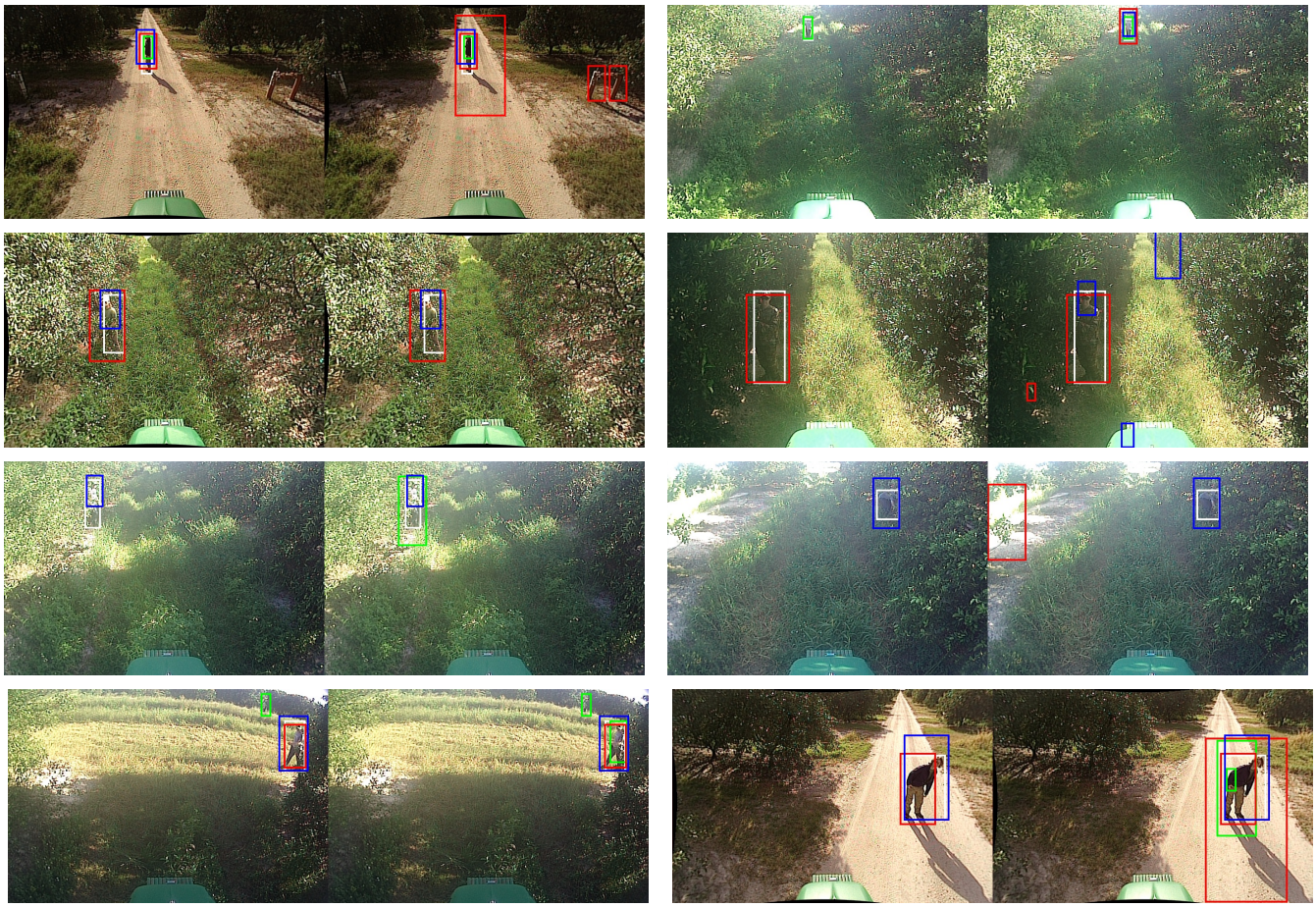


Fig. 10: Example detections in challenging conditions for different false positive rates. Each example shows bounding boxes for detections from ACF (green), DPM (red), and CNN (blue), along with the labeled ground truth (white). The left image in each pair uses a threshold that corresponds to an average false positive rate of 10^{-2} per image and the right image corresponds to 10^{-1} per image over this dataset, for the dotted lines in Figure 9. Generated with [14].

was found to have relatively little effect in those domains as long as it was below around 0.6 [3].

Figure 10 shows example detections in challenging conditions for all methods at thresholds corresponding to average false positive per image rates over the entire test set of 0.01 (left) and 0.1 (right), using a metric overlap threshold of $T = 0.1$. A video of these detections on the entire test set is also available online.[2] All three approaches were able to detect standing, visible people well with low false positives. The behavior of ACF and CNN is similar on most images. For example, they both have some more trouble detecting people with color similar to vegetation, but they also both achieve impressive detections on some challenging cases. CNN more often returns detections on people that are highly occluded or in non-standing poses. ACF detects some people very far away but misses more close people than CNN. DPM's behavior is noticeably different. It is often able to localize the full person well, even when the person is partially occluded, and it detects the camouflaged person while ACF and CNN fails. False positives not due to poor localization are most common in strong shadows and unusual lighting conditions, particularly in the presence of strong vertical edges. Note that Figure 10 shows examples of false detections from branches, weeds, and pipes.

When we require a high overlap, DPM provides the best bounding box localization and therefore has the highest metric performance, followed by ACF and then CNN. When we focus on detection instead of localization by relaxing the required overlap, the ordering nearly reverses. The preferred detector therefore depends strongly on localization requirements.

VII. Conclusions and Future Work

Off-road environments have a number of differences from urban environments, but we have shown that two reference pedestrian detection implementations can effectively find people in these settings. The only change we made besides retraining was to make the non-maximal suppression more aggressive, showing the generality of these approaches for person detection across domains. We also presented preliminary results using a CNN that produces competitive results with ACF and DPM with less susceptibility to occlusion and

[2]Full test set detections: https://youtu.be/TzzJC8sOe60

unusual poses. For all methods, achieving precise bounding box locations in this high texture environment with occlusions is challenging, and the results were very sensitive to the exact overlap threshold chosen for the standard metric. Considering only one overlap threshold, as is prevalent in the literature, can hide this effect. The CNN bounding box generation is not as mature as the other approaches and showed the largest performance change as the required overlap threshold was relaxed.

All three of these approaches show impressive detections in challenging conditions. More work is needed to achieve image-level performance acceptable for a robust robotic system, but when the output of these classifiers is viewed as a video instead of independent images, it appears that most false positives are brief and unique to one detector and so could potentially be filtered out with additional context. For instance, in future work, we plan to add information from stereo and motion. As discussed earlier, a system level performance analysis would be needed to determine the expected behavior of an actual robotic vehicle using these approaches, appropriate to the application undertaken.

ACKNOWLEDGEMENT

This work was supported by the USDA National Institute of Food and Agriculture as part of the National Robotics Initiative under award number 2014-67021-22171. The authors would like to thank John Deere for the use of the autonomous tractor used to collect data for this paper.

REFERENCES

[1] D. Portugal and R. P. Rocha, "On the performance and scalability of multi-robot patrolling algorithms," in *Int. Symposium on Safety, Security, and Rescue Robotics (SSRR)*, IEEE, 2011, pp. 50–55.

[2] S. J. Moorehead, C. K. Wellington, B. J. Gilmore, and C. Vallespi, "Automating orchards: A system of autonomous tractors for orchard maintenance," in *Int. Conf. on Intelligent Robots and Systems (IROS) Workshop on Agricultural Robotics*, IEEE/RSJ, 2012.

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[4] M. Andriluka, P. Schnitzspan, J. Meyer, S. Kohlbrecher, K. Petersen, O. Von Stryk, S. Roth, and B. Schiele, "Vision based victim detection from unmanned aerial vehicles," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, IEEE, 2010, pp. 1740–1747.

[5] G. Freitas, B. Hamner, M. Bergerman, and S. Singh, "A practical obstacle detection system for autonomous orchard vehicles," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE/RSJ, 2012, pp. 3391–3398.

[6] D. M. Bradley, R. Unnikrishnan, and J. Bagnell, "Vegetation detection for driving in complex environments," in *Int. Conf. on Robotics and Automation (ICRA)*, IEEE, 2007, pp. 503–508.

[7] C. Wellington, A. Courville, and A. Stentz, "A generative model of terrain for autonomous navigation in vegetation," *Int. J. of Robotics Research (IJRR)*, 2006.

[8] J. Bagnell, D. Bradley, D. Silver, B. Sofman, and A. Stentz, "Learning for autonomous navigation," *IEEE Robotics Automation Magazine*, pp. 74–84, 2010.

[9] D. Ball, P. Ross, A. English, T. Patten, B. Upcroft, R. Fitch, S. Sukkarieh, G. Wyeth, and P. Corke, "Robotics for sustainable broad-acre agriculture," in *Field and Service Robotics*, Springer, 2013, pp. 439–453.

[10] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" In *Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving*, 2014.

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, p. 0 278 364 913 491 297, 2013.

[13] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[14] P. Dollár, *Piotr's Computer Vision Matlab Toolbox (PMT)*, `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 1627–1645, 2010.

[16] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, *Discriminatively trained deformable part models, release 5*, `http://people.cs.uchicago.edu/~rbg/latent-release5/`.

[17] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2013, pp. 3626–3633.

[18] A. Angelova, A. Krizhevsky, and V. Vanhoucke, "Pedestrian detection with a large-field-of-view deep network," in *Int. Conf. on Robotics and Automation (ICRA)*, IEEE, 2015, pp. 704–711.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, 2015.

[20] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *ArXiv preprint arXiv:1212.5701*, 2012.

[21] C. Dima, C. Wellington, S. Moorehead, L. Lister, J. Campoy, C. Vallespi, B. Jung, M. Kise, and Z. Bonefas, "PVS: a system for large scale outdoor perception performance evaluation," in *Int. Conf. on Robotics and Automation (ICRA)*, IEEE, 2011, pp. 834–841.