Long-Range Pedestrian Detection using Stereo and a Cascade of Convolutional Network Classifiers*

Zsolt Kira, Raia Hadsell, Garbis Salgian, and Supun Samarasekera[1]

***Abstract*— In this paper, we present a system for detecting pedestrians at long ranges using a combination of stereo-based detection, classification using deep learning, and a cascade of specialized classifiers that can reduce false positives and computational load. Specifically, we use stereo to perform detection of vertical structures which are further filtered based on edge responses. A convolutional neural network was then designed to support the classification of pedestrians using both appearance and stereo disparity-based features. A second convolutional network classifier was trained specifically for the case of long-range detections using appearance only. We further speed up the classifier using a cascade approach and multi-threading. The system was deployed on two robots, one using a high resolution stereo pair with 180 degree fisheye lenses and the other using 80 degree FOV lenses. Results are demonstrated on a large dataset captured in a variety of environments.**

I. INTRODUCTION

Pedestrian detection is an important topic with a variety of applications including surveillance and collision avoidance on vehicles. There are many sensors that have been used for this task depending on the application, including radar, EO cameras (both monocular and stereo), infrared cameras, and LIDAR. While additional sensors can enhance detection, cameras still provide the widest applicability for robotics applications due to low cost, low payload size and weight, and availability of data.

The current literature on pedestrian detection using monocular or stereo cameras is typified by hand-designed features extracted from the images, the most successful being the Histogram of Oriented Gradients (HOG) for appearance [1]. These features are used to train classifiers, for example support vector machines. While these approaches can achieve some level of success, there are several challenging aspects to the problem that have not been represented in the common public datasets, including occlusion, pose variation, and pedestrians at long ranges. As has been demonstrated in several recent survey papers (e.g. [4]), an overall limitation of most approaches is that they are computationally intensive and, more importantly, perform poorly on longer-range detections with fewer than 50 pixels on target.

This paper focuses on detection pedestrians at long ranges. To motivate this, Figure 1 shows a qualitative comparison between the typical datasets used (top) and our dataset (bottom). We achieve this through several novel differences from related work. Specifically, many of the current approaches perform classification by scanning windows at multiple scales across entire images. This is impractical for longer-range detections for two reasons: 1) larger-resolution images are necessary for such detections to be possible, making the processing time much larger, and 2) pedestrians that are further than 50 or more meters away will have less than 50 pixels on target, even when using higher resolution images than typical, and hence the scale-space search must be increased to include very small scales that dramatically increase the computation time.

Fig. 1. Qualitative comparison between images (all scaled to 50% original) from the PASCAL VOC (top left), the INRIA pedestrian dataset (top right), and our long range detection data (bottom).

In order to overcome these challenges, we first find candidate pedestrians in stereo disparity maps and further filter them based on edge responses of the candidate detection windows. An important contribution of this paper is the design of a convolutional neural network to accurately classify pedestrians using both appearance and disparity information, without having to hand-design additional features to leverage stereo depth information. The second contribution is the use of a second classifier specifically designed for long-range detections in order to increase recall and decrease false positives. We show that these two classifiers outperform current HOG-based methods for both medium-range and long-range detections. Finally, we cacade a third, lower-resolution classifier that is faster with the higher-resolution classifier, the computation of which is spread across multiple threads, in order to speed up classification. We demonstrate the positive effects that each of these techniques afford on a large set of data collected in varying outdoor environments

This work was supported by TARDEC under the Robotic Technology Consortium CombatID program (69-200907 T05).

[1]Z. Kira, R. Hadsell, G. Salgian, and S. Samarasekera are at SRI International Sarnoff, Princeton, NJ. zsolt.kira@sri.com

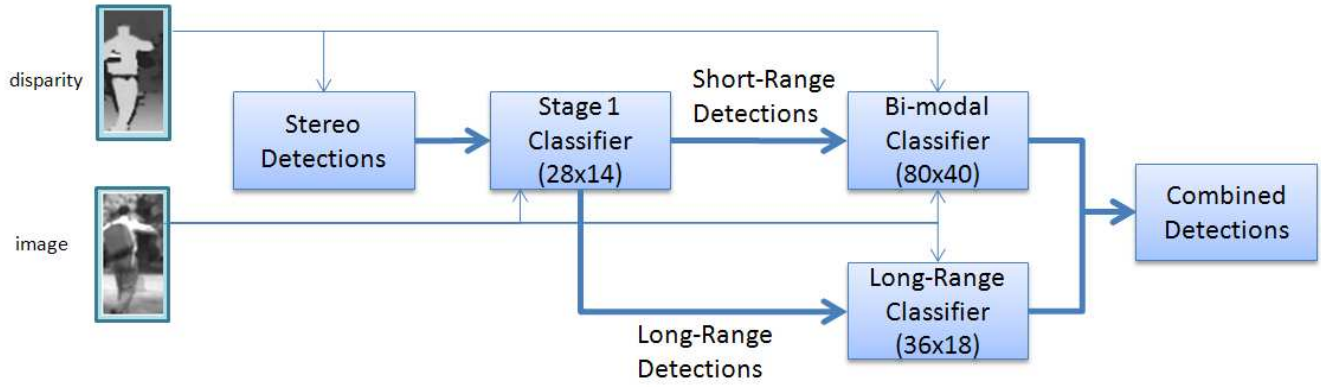

Fig. 2. Depiction of overall system, beginning with a stereo detector that finds vertical objects above locally estimated ground patches. A series of convolutional network classifiers are then used, starting with a fast 28x14 classifier that removes easier instances, a bimodal appearance and stereo classifier for shorter range detections, and a long-range appearance-only classifier for detections with less pixels. Detections that are positively classified by the bimodal classifier and long-range classifier are then combined to form the final list of hypothesized pedestrians.

and containing a significant amount of people far away.

II. RELATED WORK

Pedestrian detection is an active area of research, with dominant approaches utilizing monocular cameras [1][2] and a more limited amount of work on the use of stereo cameras [3][6]. The most common approach used as a baseline for comparison is one that uses Histogram of Oriented Gradients (HOG) features that are then fed to a linear SVM classifier. This approach has been since augmented with other features such as parts-based features [2] or motion features [5]. Several works have explored combining multiple features to achieve performance that is better than any single feature [7]. See [4] for a survey of methods that have been explored.

While several of these methods outperform a HOG-based classifier in the tested datasets, they are often much slower than HOG features that are amenable to efficient implementations, including through the use of CUDA. Several feature types, such as parts-based detections, are also difficult to obtain with lower resolution image patches. Stereo-based detectors have also been used in, e.g., [3] that use stereo to produce detections by segmenting the scene based on range and using various shape and appearance features. Even when using stereo, however, these works are extremely limited in their ability to detect pedestrians when the number of pixels on target is low.

III. SYSTEM DESIGN

The overall system design is depicted in Figure 2, consisting of a stereo detector and a cascade of three convolutional network classifiers. We will now describe each portion of the system in detail.

A. Camera Calibration and Stereo Detection

In order to compute accurate stereo disparity maps, the cameras must first be calibrated to obtain intrinsic camera parameters and extrinsic stereo parameters. For this paper, two types of camera lenses were used. The first was a 180-degree fisheye lens, where we used the Scaramuzza MATLAB toolbox to obtain intrinsic parameters [9]. Given these parameters, we estimated extrinsic parameters using images containing a checkerboard pattern in both views, and finally mapped the fisheye image onto a cylindrical image. The second camera pair used an 80-degree lens which was calibrated using a traditional camera model and checkerboard pattern.

Given a calibrated stereo pair, we compute stereo disparity maps using a fast CUDA implementation [12]. The resulting disparity image is then used to find vertical structures in the scene that could potentially correspond to pedestrians. Specifically, we discretize the image into a fixed number of patches and estimate the ground plane, converted into three-dimensional Cartesian space, using the disparity information. This is done by building a histogram of disparity values for each horizontal scanline in the patch and finding the best orientation estimate for the patch across all of these histograms using a Hough transform. This orientation is used to estimate the ground as a function of disparity for each patch. A mask of all disparity pixels that are above this estimated ground is then created. Connected components is used to group these above-ground pixels together, separated into multiple disparity levels (distances) to provide a final estimate of all objects that are vertical. An ROI (region of interest) in the image space is created from the bounding box of each component.

The technique described above finds all objects that are estimated to be above local ground patches, and hence can return a large number of ROIs. For example, large structures such as buildings and trees can be picked up, although they are contiguous structures that are unlikely to be pedestrians. In order to further prune these detections, it was observed that some of these false positives present no vertical edges in the camera image, while pedestrian silhouettes contain strong edges. Hence, we also use a Sobel edge detection filter on the entire image, and prune ROIs that have little vertical edge response. Section IV will show that this can significantly reduce the amount of detections with very little loss of recall.

Fig. 3. Edge-reponse example with a negative instance mistakenly detected due to the curb and stereo noise having very little vertical edge response (left) while a positive instance has a strong vertical edge response (right).

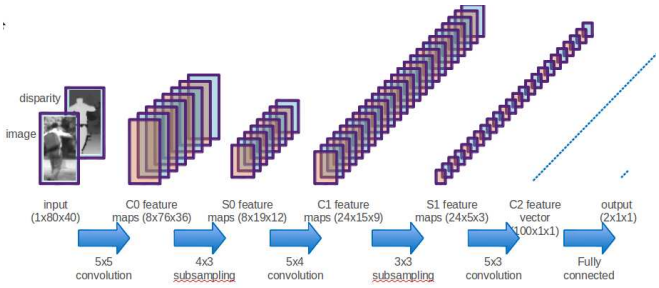

Fig. 4. Architecture for 2 input convolutional neural network.

Figure 3 shows an example of a false positive with little edge response and a typical true positive that presents strong edges.

B. Classification using Convolutional Networks

After the stereo-based detection, a classifier is used to determine whether each ROI is a pedestrian or not. While previous appearance-based classifiers have achieved some success for cases where there are more than 50 pixels on a pedestrian, they typically degrade heavily beyond this range [4]. While we mitigate this effect somewhat by our use of higher-resolution cameras with an 80-degree lens, the fisheye lens has significantly less pixels on target and even the narrower field of view results in pedestrians at our maximum target range (approximately 100 meters) occupying around 30 pixels or less. In fact, we will show in the experimental section that approximately a third of our dataset consists of detections with less than 50 pixels in height.

Instead of hand-crafted features, we use convolutional neural networks, a deep learning method that has achieved success on a wide array of modalities for tasks such as object recognition [10] and speech [11]. It is considered a deep learning technique due to its use of a hierarchical architecture that consists of alternating filters (that are learned) and pooling, resulting in the processing of data at multiple levels. In addition to learning the underlying filters and connection weights, labeled training data can be used to learn a discrimination function that is able to classify data into multiple classes.

There are several advantages to using this technique over traditional approaches. First, it is a flexible framework which supports the use of multiple modalities, in this case EO images and stereo disparity images. We design a neural network architecture that utilizes an asymmetrical connection map with inputs to ensure that features are learned from each modality independently as well as jointly over both modalities. In order to incorporate this additional information, there is no need to hand-design additional features for the new modality.

A second advantage is that this framework allows the features to be learned for the particular task of pedestrian detection, taking long-range instances into account when learning them. As mentioned earlier, hand-designed features have thus far not been as successful in these situations, and we will show in Section IV that the convolutional network classifier outperforms standard HOG features. Finally, the processing time required during run-time (i.e. on test data after training) lends itself to close to real-time application.

Figure 4 shows the architecture. The inputs to the classifier consist of two channels: An 80x40 8-bit grayscale camera image and a second 80x40 integer disparity map. Both inputs are normalized to zero mean and unit variance before being applied to the subsequent layers. Beyond the input, a 6 layer hierarchy was designed consisting of 3 convolutional layers, 2 pooling layers, and a fully connected layer that produced two numerical outputs representing scores for each class (pedestrian and non-pedestrian). In all, there are 8,000 trainable parameters describing the weights of the connections. These network parameters are optimized using stochastic gradient descent given a training set consisting of labeled positive and negative instances.

In addition, in order to further increase the classifier speed, we trained a second classifier that takes in 28x14 inputs instead of the original 80x40. The threshold for this classifier is conservatively chosen to achieve a high recall. While the false alarm rate will not be as low as with the full-sized classifier, it is able to filter out a significant portion of the easier negatives. Since the smaller classifier is faster, a net time savings can be gained. Note that we also run the two classifiers across multiple threads since the classification of ROIs can be independently processed, yielding further speed savings on multi-core systems.

C. Long-Range Classifier

There are several challenging aspects to classifying detections of pedestrians at long range. First, there are few pixels on target (e.g. less than 50) and hence the resulting stereo disparity map will be of much lower quality. This can result in errors in the ground estimation as well as in the stereo calculations themselves, which require some texture that can be eroded by blur. Second, these problems can cause detection windows that are not fully aligned with the pedestrians, a situation that can reduce the classifier's success.

In order to mitigate these problems, we use a second appearance-only classifier, trained exclusively on pedestrians with 40 pixels or less on target. Unlike the earlier classifier,

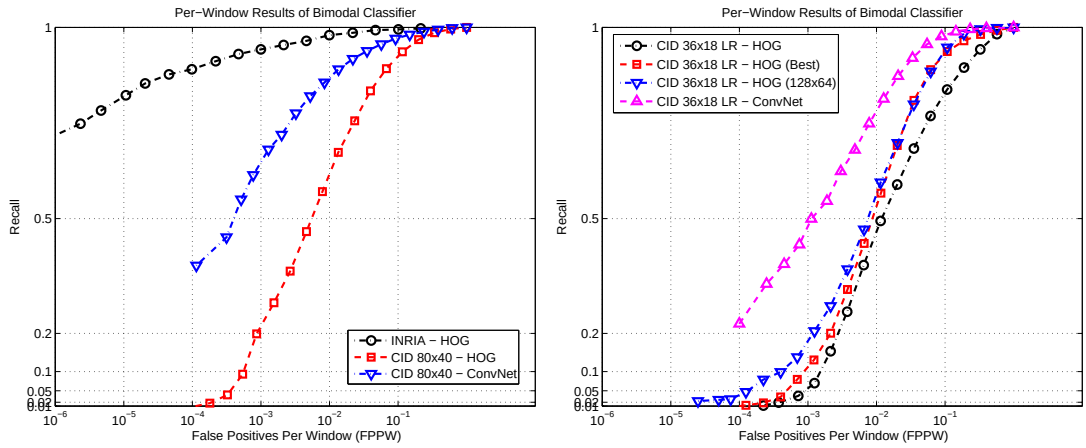

Fig. 5. Left: Recall/FPPI curve demonstrating the performance of convolutional networks (”ConvNet”) over HOG features combined with linear SVMs [1] (”HOG”) on our dataset (”CID - 80x40”). We also demonstrate the difficulty of the dataset by showing the results of the HOG algorithm on the INRIA dataset; our dataset is significantly more challenging due to longer-range pedestrians. Right: Recall/FPPI curve for results obtained using the long-range classifier trained on pedestrian samples that were 40 pixels in height or less. We compare our results to the original HOG-based classifier (”HOG”), after optimizing the parameters of the HOG-based classifier (”HOG (Best)”), and after upsampling the 36x18 samples to 128x64 for which the HOG classifier was originally designed (”HOG (128x64)”).

the windows with which training occurs is not resized or resampled, and pedestrians that are smaller or larger are allowed. In other words, the center of the detection is taken and a window in the size of the classifier input (in our case 36x18) is taken around the detection. This training regime specifically builds in robustness to partial overlap with the pedestrian and reduced image blur.

In order to combine the two classifiers, the detections are split based on their height in pixels and fed to each respective classifier. The results are then combined together to form the final pedestrian detections. Note that since each classifier has its own characteristics in terms of the operating curve, different thresholds are used for each classifier.

IV. EXPERIMENTAL RESULTS

In order to test the system, over 45 video sequences were gathered in various outdoor environments over a period of 10 months and subsequently ground-truthed (we use the name ”CID” to label this dataset in subsequent graphs). Figure 6 shows examples of mid-range detections, and table I summarizes some statistics of the dataset. The environments ranged from a helipad spanning approximately 60 meters, a parking lot with long rows of cars, an open area with buildings and vegetation, a forest environment, and an open parking lot spanning more than 100 meters.

Two sets of stereo cameras, each operating at either 1 Hz or 5 Hz, were used to capture the data, and classifier training consisted of data from both cameras. Both rigs consisted of a stereo pair of Prosilica GB2450 cameras operating at a resolution of 2448x2050; one setup used a 180 degree fisheye lens, resulting in an image of size 2881x657 after conversion to the cylindrical image, while another rig with 80-degree lens and more pixels on target was used, with the image cropped to size 2448x625. The pedestrians in the dataset were captured at ranges of 10 to 140 meters. In all,

Fig. 6. Sample mid-range inputs showing image and disparity maps.

over 39,000 frames with over 160,000 annotated pedestrians resulted, with about a third being 50 pixels or less. The datasets also increase in complexity, with the last one (”Set 4”) consisting of people performing various maneuvers and running, conditions that did not exist in the training set. We will now describe our results on the training of the bimodal and long-range classifiers on ground-truthed positives and negatives sampled from images. The positive and negative instances were extracted from the data sets and separated into training, testing, and validation sets. After showing results on the testing set on a per-instance basis, we will show results on entire sequences on a per-image basis.

TABLE I
DESCRIPTION OF DATA SETS

Attribute	Fisheye	80 Degree
Frames	24,572	14,533
Detections	99,522	64,204
Mean Height (Pixels)	87.78	88.37
% l.t. 50 pixels	32.1	35.6
% l.t. 40 pixels	14.4	11.6

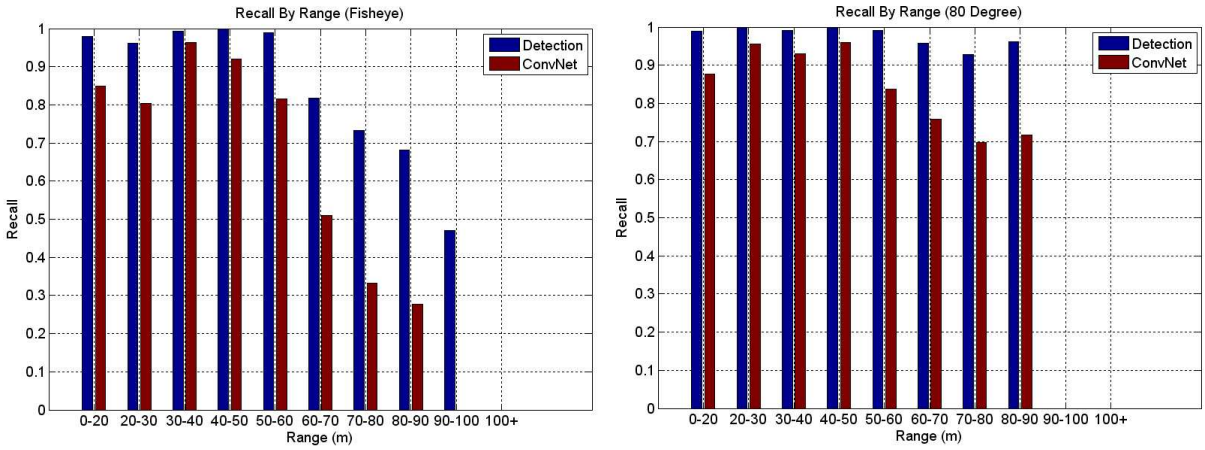

Fig. 8. Results on one sequence breaking down the detection and classification results for the fisheye camera (left) and 80 degree camera (right). Over half of the pedestrians are picked up at greater than 50 meters for the fisheye, representing approximately 30 pixels. The 80 degree camera performs even better, picking up approximately 60% of the pedestrians at over 80 meters.

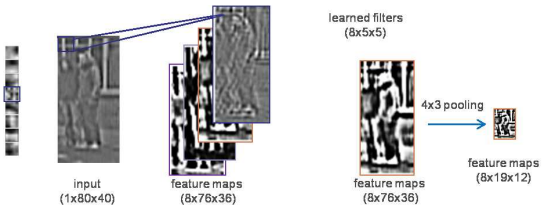

Fig. 7. Example filters that were learned and the features maps after training the convolutional network.

A. Bimodal Convolutional Network

To train the convolutional network classifier, 25 sequences were combined and all of the positive and negative image patches were shuffled into training, testing, and validation sets. To obtain positive samples, ground-truth ROIs were extracted from the images. Note that a larger border was used during extraction so that the original positive set could be expanded by applying translational and rotational jitter. This process serves to expand the positive set and make it robust to some transformation. Negative patches were extracted by taking random patches that did not intersect with the ground truth pedestrian windows but that had similar scales and image vertical positions. In all, after jittering the images 800,000 labeled positives (ROIs with pedestrians) and negatives (ROIs with no pedestrians) were used. Training took about 2 days running on a single cpu.

Figure 7 shows an example of the resulting filters that were learned, and Figure 5 (left) shows the result on the testing set. The results are displayed as Recall/FPPI curves, where the class decision is made by comparing the two numerical outputs of the network classifier. In this case, a bias on the positive class is applied and varied to produce the curve. For comparison, we have also trained a HOG-based SVM classifier with the same training set and tested on the same testing set. The same scoring methodology was used as well. As can be seen, the convolutional network achieves significantly better performance on the testing patches (note that the x-axis is log-scale). Note that some of this gain may be due to the additional modality, but as we will see in the next subsection the convolutional network outperforms HOG even when both use only appearance. We also display results of the HOG-based classifier on the original INRIA dataset to demonstrate that our dataset is significantly more challenging than INRIA. This is due to significant amount of long-range pedestrians.

B. Long-Range Classifier

The long-range classifier used 29 sequences, with a similar methodology for extracting positive samples. In this case, however, the samples were not resampled to the classifier input size; instead, the center of the positive ground truth windows that were 40 pixels in height or less was calculated and a 36x18 patch around this center was extracted. Negative samples consisted of stereo detections that did not intersect with the ground-truth. In all, after expansion through jittering, approximately 53k positive and negative samples each were used for training, and approximately 40k positive and negative samples each were used for testing. Training and testing sets were obtained by randomly shuffling the input data.

Figure 5 (right) shows the per-window results of the long-range classifier. Again, we compare these results to the original HOG features with an SVM classifier. Since the image samples were much smaller than the HOG algorithm was designed for, we also varied the number of cells, cell size, and distance metrics used in the HOG algorithm. The ”HOG (Best)” condition shows the results of the best parameter set, consisting of 4 cells, each with a size of 4 pixels, and the L2Hys metric. We also performed a second condition where the smaller patches were upsampled to the original 128x64 that the classifier was designed for. This latter condition performed the best for the HOG-based classifier, but was still outperformed by the convolutional neural network.

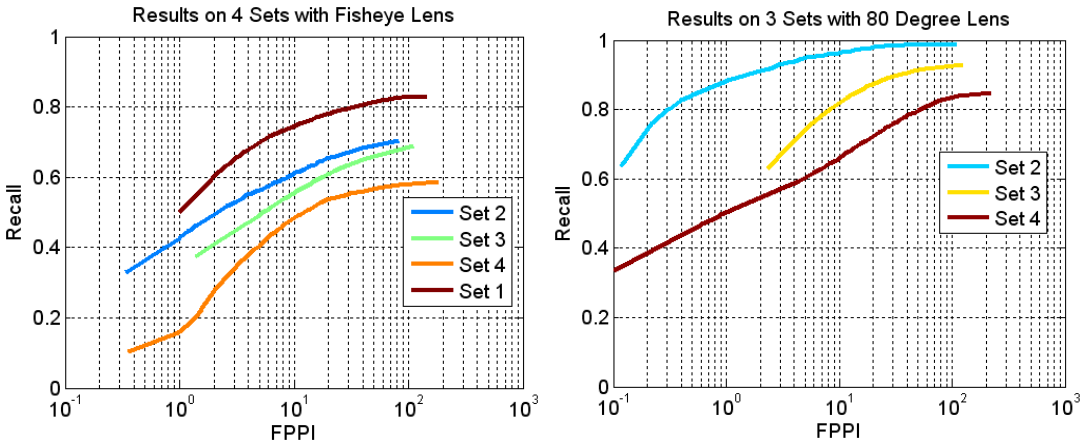

Fig. 9. Results of the classification system on a per-image basis, on 4 subsets of the data. We show results for the fisheye (left) and 80-degree camera (right), and demonstrate competitive results for long-range detections.

C. Results on Data Sequences

Although the results on a per-window basis is informative, the practical deployment of a detection system is determined by the per-image performance. As a result, we have taken four subsets of the data based on the day they were captured, with increasing level of difficulty. As mentioned earlier, the last set consisted of people running and performing various maneuvers. Figure 8 shows the results for one sequence only and at one particular operating point (resulting in approximately 1.6 false positives per frame), focusing on the recall as the estimated distance, obtained from stereo, varied. We show results for the fishey camera (left) and the 80 degree camera (right), showing that the latter can detect pedestrians at further ranges. Even the fisheye, however, can detect pedestrians at greater than 50 meters, corresponding to approximately 30 pixels in height for this lens. The 80 degree camera can detect greater than half of the pedestrians beyond 80 meters. Figure 9 shows our results for the four subsets using the fisheye camera (left) and latter three subsets for the 80 degree camera (right, note that the first subset was only collected using the fisheye camera). Overall, competitive detection rates at fewer than 1 false positive per frame can be obtained, especially when using the 80 degree camera and in the earlier datasets that did not involve running and other motions.

These results did not utilize the long-range classifier. In order to show that the long-range classifier can increase performance on a per-image basis, we used the entire pipeline on a challenging sequence from Set 4. Figure 10 shows results at different pixel height-based splits for one operating point (top) and the entire Recall/FPPI curve (bottom) for the sequence, demonstrating that higher operating points can be achieved in the performance curve if the long-range classifier is used for detections with a height of 40 pixels or less. There is a steeper dropoff at very low FPPI levels, however, possibly resulting from the fact that the long-range classifier does not utilize stereo information.

Figure 11 shows results on a subset of the sequences to

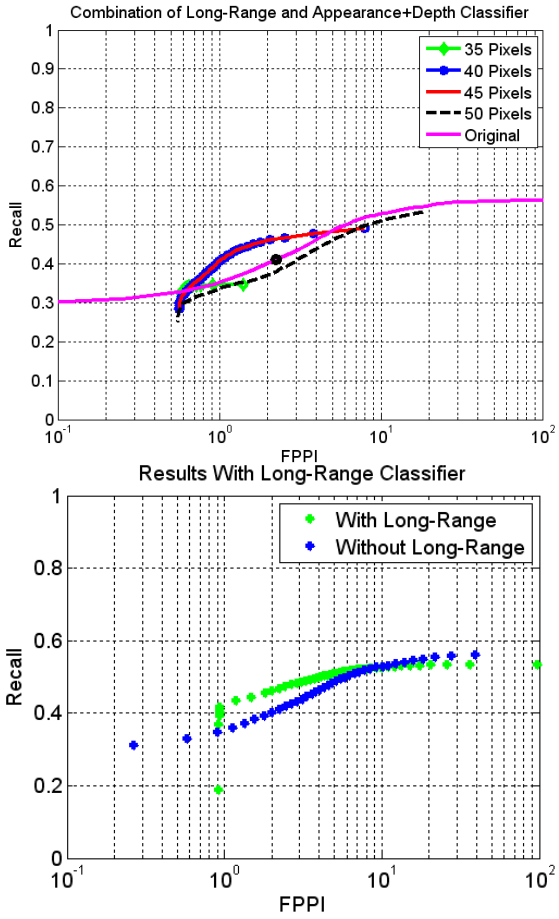

Fig. 10. Results of incorporating the long-range classifier, with varying pixel sizes used to split which detections go to which classifier, at one operating point (black dot, top) and the entire curve when using a pixel height split of 40 pixels (bottom). As can be seen, a boost in performance can be achieved over the original system when the long-range classifier handles detections at 40 pixels or less.

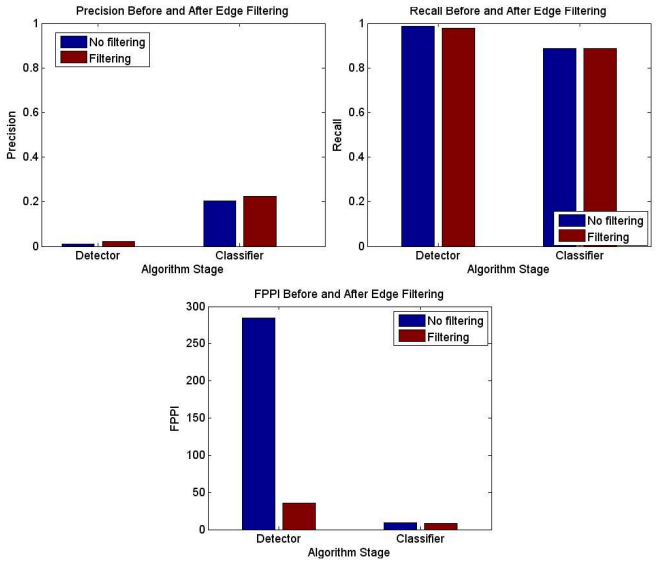

Fig. 11. Results of edge-based filtering of detections in terms of precision (top left), recall (top right), and false positives per frame (bottom) after stereo-based detection and classification. As can be seen, a drastic reduction in the number of detections can be achieved for these datasets.

show the effect of edge-based filtering of stereo detections, for the three metrics. The left column shows results after the detection stage, and the right column shows results for the classification stage. As can be seen, edge-based filtering significantly reduced the amount of detections while resulting in very slight recall decreases. This significantly decreases the number of detections that have to go through the classification stage. After the classifier, recall is very slightly decreased while some false positives that would have passed through the classifier are removed.

Finally, we show timing results from the cascade of the lower-resolution (28x14) classifier that is used to filter out easy examples early on to avoid having to run it through the more computationally expensive bimodal classifier. Note that we do not use the long-range classifier or edge-based filtering for these timing results, which would only improve the speed as the long-range classifier is faster. Figure 12 shows timinig results on a quad-core i7 Dell Precision M6500 laptop. Although we do not show it here, there is little to no effect on the accuracy metrics after the cascade, as the lower-resolution classifier is extremely high in recall. As can be seen from the timing results, the cascading and multi-threading of the detections obtained from stereo significantly increases the running time, resulting in rates that are close to the capture rate of 5Hz.

V. CONCLUSIONS AND FUTURE WORK

This paper demonstrated a system designed for the detection of pedestrians at varying distances, including ranges that have typically not been explored in the literature. The design of the system is targeted towards these longer-range detections through several novel techniques. First, a stereo-based detector is used to avoid having to classify the entire image at many scales, a method that has typically been

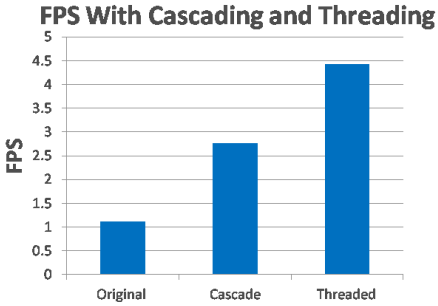

Fig. 12. Timing results on sample frames using the base system, two-classifier cascade, and after threading.

employed but is computationally expensive due to the higher resolution of the images and smaller scales of the pedestrians. A second edge-based filtering step is then employed to reduce the number of false positives significantly.

A cascade of three convolutional network classifiers are then employed, each of which is designed for a particular subproblem. A bimodal classifier is first trained to leverage both appearance and stereo disparity information. We use a deep learning approach to obviate the need for having to hand-design new stereo features that are instead automatically learned as part of the optimization process. A second classifier that takes in much smaller inputs is also trained, designed to quickly filter out easier false positives. Finally, a classifier designed specifically for longer ranges, where there are 40 pixels on target or less, in order to increase robustness to image blur and misalignments of the underlying detections that occur in such cases. We have shown that each of these design decisions have led to a system that is either more accurate or computationally faster. We have also shown that the bimodal and long-range classifier perform significantly better on a per-window basis than the HOG-based SVM classifier.

There are several avenues of future work that remain. First, there are more sophisticated methods to combine the short-range and long-range classifiers that may yield additional improvements. Second, while this work has made significant headway in the area of pedestrian detection, and as mentioned in recent surveys [4], it remains a significant challenge to deploy these systems for applications such as safety where very few false positives are tolerable and very high recalls are desired. One topic that has not been explored in this paper is tracking, where temporal information is used to reduce false positives. Another significant challenge for pedestrian detection is the presence of persistent false alarms, for example on objects that resemble humans such as poles, signs, etc. Here, additional modalities such as dense LIDAR or classification using multiple views obtained by a mobile robot may be helpful.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", CVPR, 2005.

[2] P. Felzenszwalb and R. Girshick, and D. McAllester, ”Cascade Object Detection with Deformable Part Models”, CVPR, 2010
[3] M. Bajracharya and B. Moghaddam and A. Howard and S. Brennan and L.H. Matthies, A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle, The International Journal of Robotics Research, pp. 1466–1485, v. 28:11-12, 2009.
[4] P. Dollar, C. Wojek, B. Schiele and P. Perona , ”Pedestrian Detection: An Evaluation of the State of the Art”, PAMI, 2011.
[5] S. Walk, N. Majer, K. Schindler, and B. Schiele, New features and insights for pedestrian detection, in IEEE Conf. Computer Vision and Pattern Recognition, 2010.
[6] S. Walk, K. Schindler, and B. Schiele, Disparity statistics for pedestrian detection: Combining appearance, motion and stereo, in European Conf. Computer Vision, 2010.
[7] C. Wojek and B. Schiele, A performance evaluation of single and multi-feature people detection, in DAGM Symposium Pattern Recognition, 2008.
[8] Y LeCun, S Chopra, R Hadsell, M Ranzato, F Huang, A tutorial on energy-based learning, in Predicting Structured Data, ed by G. Bakir, T. Hofman, B. Scholkopf, A. Smola, B. Taskar, MIT Press, 2006.
[9] Scaramuzza, D., Martinelli, A. and Siegwart, R., (2006). ”A Toolbox for Easy Calibrating Omnidirectional Cameras”, Proceedings to IEEE International Conference on Intelligent Robots and Systems (IROS 2006), Beijing China, October 7-15, 2006.
[10] MA Ranzato, FJ Huang, YL Boureau, Y LeCun Unsupervised learning of invariant feature hierarchies with applications to object recognition, CVPR 2007.
[11] Honglak Lee, Peter T. Pham, Yan Largman, Andrew Y. Ng: Unsupervised feature learning for audio classification using convolutional deep belief networks. NIPS 2009: 1096-1104.
[12] Sizintsev, M. and Kuthirummal, S. and Samarasekera, S. and Kumar, R. and Sawhney, H.S. and Chaudry, A., ”GPU accelerated realtime stereo for augmented reality”, in Proceedings of the 5th International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT), 2010.