# Parts-based object recognition seeded by frequency-tuned saliency for Child Detection in Active Safety

Shinko Y. Cheng, Jose Molineros, Yuri Owechko
HRL Laboratories, LLC
3011 Malibu Canyon Road
Malibu CA 90265
{sycheng,jmmolineros,yowechko}@hrl.com

*Dan Levi, †Wende Zhang
*GM Advanced Technology Center-Israel
†GM Research
{dan.levi,wende.zhang}@gm.com

*Abstract*— **This paper proposes a novel system for automatically detecting children from a color monocular back-up camera, as part of a back-up warning device in passenger vehicles. We presented the use of an attentional mechansim that focuses compute-intensive bounding-box classifiers on a subset of all possible bounding-box solutions to enable real-time performance of 248ms per frame with negligible reduction in performance. The attentional mechanism called *Attention to Children* which consists of a window generation and verification cascade of based on Frequency-Tuned Saliency, Variational-Optical-Flow Obstacle Detection and finally a parts-based classifier. We also presented a method of reducing much of the cascade classifier evaluations by judicious sampling of the bounding-box solution space. The result is a reduction in the number of windows evaluated down to 439 from more than 12K windows in traditional sliding window techniques, a 97% reduction in the number of windows. The verification stages leading up to the parts-based classifier further reduces the number of windows to half. Together with a parallel processing and pipelining, the final processing time was 248ms per frame.**

## I. Introduction

Statistics show 2600 children are involved in backing collisions every year in the U.S., 104 children are injured fatally. Tragically, 70% of these incidents involve a parent or close relative behind the wheel [1], [2]. No consumer vehicle currently offers a system specifically to detect and warn of children being present in the collision path of a vehicle in reverse.

This paper proposes a system for automatically detecting children from a color monocular back-up camera in passenger vehicles. The child detector as a part of a back-up warning device warns the driver of the presence of children in the vehicle's back-up path. The output of the system estimates the position and scale of a bounding box in an image. Our main contribution is an attentional mechanism that focuses compute-intensive bounding-box classifiers to a subset of the entire space of possible bounding-boxes to enable real-time performance with marginal loss in accuracy. The detector consists of 4 modules falling into 2 functional groups: solution generation and verification. The parts are

1) An optimally tuned bounding-box candidate generator,
2) A neuroscience-inspired attention mechanism called Frequency-Tuned Saliency (FTS),
3) An above-ground obstacle detector based on optical flow called Variational Optical Flow Obstacle Detection (VOFOD), and
4) A parts-based object detector.

Bounding-box generator serves to present a finite set of possible bounding-box solutions to the subsequent components for analysis. The other 3 parts all have the role of analysing the pixels contained within the bounding-box and scores them high if the presence of a child within is likely, and low otherwise. This algorithm is generic to any bounding-box classifier, including parts-based ones. For testing, the Feature Synthesis model was used [3].

Th proposed attentional mechanism is referred to as Attention to Children (A2C), named such for the definition for *interesting* – and therefore attention worthy – is defined by the true presence of a child and not the fidelity to neural processes, although the computational model for attention remains consistent with our understanding of the mammalian visual pathway.

Automatically detecting children from video is a challenging problem that the state-of-the-art in vision-based object detectors cannot practically address because children naturally appear in a variety of poses in a back-up camera and promising approaches are computationally prohibitive. Children may be found running, sitting on the ground, riding a toy, lying down. The challenge of detecting children in a variety of poses is in addition to the typical challenges for a vision-system, such as varying lighting conditions, occlusion, perspective distortion. It is imperative that a warning system be able to detect children in the presence of all of these challenges, and to do so at a sufficiently high update rate and high accuracy (high probability of detection and low probability of false alarm). A high update rate allows the system to track the position of the child through time and predict the probability of collision. To address the accuracy challenge, a parts-based classifier has been selected because it is shown in literature as one of the most accurate detectors in existence today. Our contribution is a method that increases the speed of processing by the parts-based classifier at a cost of a marginal amount of accuracy. The proposed system reduces the amount of processing by 50% over exhaustive processing of bounding-boxes, and 97% over dense sliding window-based processing.

The proposed system achieves the high update rate by employing a cascade architecture that prunes away more and more bounding-boxes from each stage in the cascade, until only a set of highly probable candidates remain. The system also achieves the high update rate by sampling the image with a minimal number of bounding-boxes as a starting set to achieve the desired accuracy.

The remaining sections are organized with a brief review of related work in sec. II. It follows by a description of the algorithm in Sec. III, followed by testing results in Sec. IV, and conclusions in Sec. V.

## II. Related Work

Existing production backup warning systems based on ultrasound and radar sensors are able to measure distances of objects up 2m behind the vehicle [4]. However, since the systems sense and advise on a discrete number of zones over the volume of space behind the vehicle, the coarseness makes the task of disambiguating humans from general obstacles difficultin contrast to the proposed system, which uses vision sensors containing 300K data points (640x480 image resolution) over the same space. The coverage extends beyond 2 meters to the horizon, enabling the system to determine very finely the location of the child behind the vehicle. Besides the proposed use of only the vision camera to detect obstacles in the rear, there is to the best of our knowledge no current production automotive pedestrian/child detection system that explicitly detects children in arbitrary poses.

Parts-based detection is known to be highly accurate, but computationally prohibitive. There have been efforts to increase the efficiency by way of a cascade of detectors. Each stage of the cascade analyses a particular aspect of the target object in the bounding-box and determines whether the sample is a non-target and rejected immediately, or a passed to the next stage for further analysis. The samples that survive to the final stage are the predicted targets [5], [6], [7], [8]. Another method analyses the scene geometry [9], textures [10], or appearance of targets in the image with a faster procedure to first detect likely pixel blobs of targets with which to focus the processing of the slower parts-based detector. Alexe *et al.* utilizes saliency, color-contrast, and superpixel straddling features to score bounding boxes according to their likelihood of containing an object within [8], measuring the *objectness* of targets.

Our method extends these concepts and utilizes a cascade to reject negative candidates early and introduce the use of Frequency-Tuned Saliency (FTS) [11]–an improved saliency algorithm–and a novel above-ground residual optical flow obstacle detector to direct the attention of the parts-based classifier. The speed of the objectness model is impractically slow at 4 seconds per image of size 350x500 [8]. Our proposed structure allows operation at nearly 270ms per frame on the same machine. We also present analysis on bounding-box candidate generation and show that a significant savings in processing can be obtained merely by judiciously sampling the solution space.

## III. Attention to Children

Figure 1 illustrates the proposed algorithm. The system takes color images as input, and directs the attention of bounding-box classifiers on salient boxes in the image. The method consists of 2 functional groups: window generation and window verification, where we use window and bounding-box interchangeably. Three separate modules perform the function of window verification, namely Frequency-Tuned Saliency, Variation Optical Flow Obstacle Detector, and a parts-based detector.

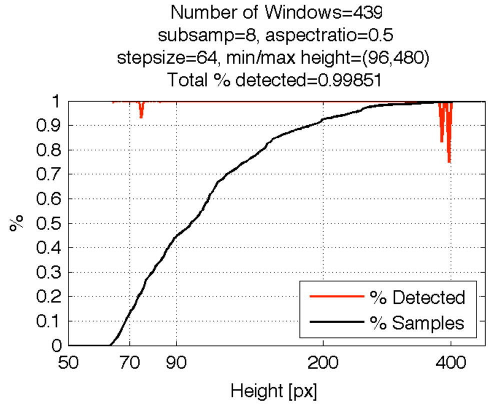

Fig. 2. Plot in red shows the percentage of humans detected.

Window generation module generates an exhaustive set of bounding-boxes as potential solutions to the child detection problem. The objective is to include all target boxes, and to the furthest extent possible, prune away the obviously non-target boxes. The procedure begins by undistorted the image using the cameras calibration parameters. Then, place a bounding-box with an aspect ratio of 0.5 starting from the top left, and slide it across and down the image with 64 pixels between steps. Repeat the process for a range of heights from 96 to 480 at 1.35x steps between heights. Altogether 439 windows were generated for an image of size 640x480.

With a video dataset containing ground-truth annotations of children and adults taken from a back-up camera, this set of parameters was shown to generate a minimal set of bounding boxes that still detects 99.85% of all targets in the dataset. Figure 2 shows the percentage of humans in the scene detected by this initial set of bounding-boxes corresponding to box height. The criteria for detection is if the area of the intersection between the ground-truth bounding-box and the detection bounding-box divided by the union of their areas exceed 1/3, the detection bounding-box *detects* the human. Shown in the same plot is the cumulative distribution of the humans present in the children dataset. The 0.15% of the humans missed in this dataset is the very far (short height) and very close (tall height) objects. In particular, the very close objects are typically cropped and only part of the human is visible.

Following window generation, the subsequent processing steps sequentially remove windows until only the windows containing the human remains. The first stage of filtration is performed by Frequency-tuned Saliency (FTS). FTS was first proposed by Achanta *et al.* as a computational algorithm that mimics the behaviour of the mammalian visual pathway but analyses the saliency of contiguous groups of pixels and

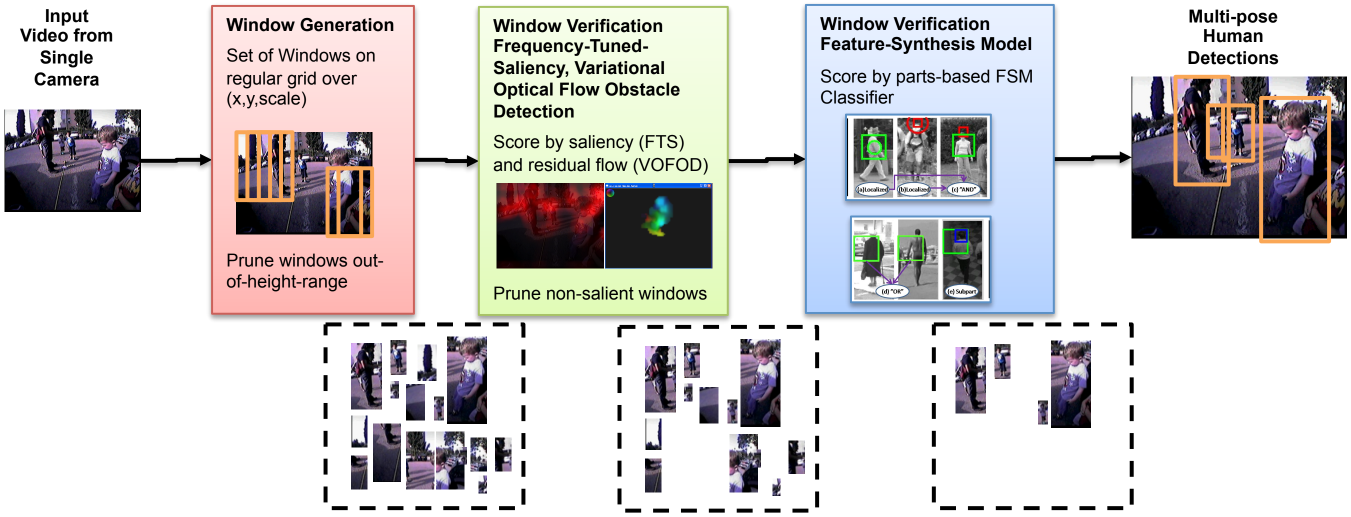
Input Video from Single Camera
Window Generation
Set of Windows on regular grid over (x,y,scale)
Prune windows out-of-height-range
Window Verification Frequency-Tuned-Saliency, Variational Optical Flow Obstacle Detection
Score by saliency (FTS) and residual flow (VOFOD)
Prune non-salient windows
Window Verification Feature-Synthesis Model
Score by parts-based FSM Classifier
(a)Localized
(b)Localized
(c) "AND"
(d) "OR"
(e) Subpart
Multi-pose Human Detections

Fig. 1. Shown are the processing steps in the parts-based detection of children with Attention to Children. Input images are used to generate a minimal exhaustive set of candidate boxes. Each box is verified with Frequency-tuned Saliency, Variational-optical-flow-obstacle-detection, and a parts-based classifier (Feature Synthesis model shown). Final output is bounding-boxes that survived the filtration process, leaving behind only the bounding-boxes over children (and adults).

utilizes the global statistics of the images as opposed to local neighbourhoods around the pixels. The first step of FTS is to compute the saliency image by taking the norm over the color channels of the difference between average pixel value and a Gaussian blurred version of the input image. Separately, the input image is processed to obtain a mean-shift segmentation, where each pixel joins a neighbouring, contiguous group of pixels where the mean pixel value of the group is similar to the value of the individual pixel, more than any other group in the neighbourhood. For each segment, if the mean saliency score of the segment exceeds twice the saliency score of the average over the entire image, the segment returned. Further details can be found in [11]. Each of the 439 windows is scored using the output segments returned from the image. The score is computed by taking the average value of the saliency within the window. The integral image approach is used to reduce the computations to constant time for computing integral image and 4 memory accesses to compute the average value for each window [5]. Windows exceeding a predetermined threshold are allowed through for further processing.

The next stage of filtration is Variational Optical Flow Obstacle Detection (VOFOD). This module detects obstacles above the ground from a pair of images taken in succession. VOFOD consists of 3 stages: feature detection, ego-motion registration of the images and non-ground motion segmentation. In the first stage, CenSurE image landmarks are detected in the two input images. The pair of images and landmark locations are transformed into a top-down or birds-eye view given the camera calibration parameters and a planar homography model of the ground. In the second stage, the collection of landmarks are clustered into groups by computing minimum-spanning-tree and pruning the very long edges, leaving the clusters of landmarks that are close together. The 2D ground-motion is computed with each cluster of landmarks, and the cluster with the least reprojection error is selected, and used to estimate the global 2D ground-motion vector. With this motion vector, the 2 images in birds-eye view can be registered. In the third and final step, a dense optical flow map using Variational-optical-flow is computed from the original pair of input images. Each flow vector is projected onto the bird-eye view and the difference between each projected flow vector and the 2D ego-motion vector is computed, resulting in the residual flow vector. More details on the VOFOD algorithm can be found in [12]. Each of the remaining windows from FTS filtration is scored using the residual flow by computing the mean magnitude of the residual flow vectors within each window. Again, the integral image is used to compute the mean. Windows exceeding a second predetermined threshold are allowed through for further processing.

In the final stage, a parts-based classifier is used to classify the remaining windows. For testing, the Feature Synthesis model is used for analysis, but any window classifier may be used, including Deformable Parts Model [6], Multi-scale HoG detector [13].

Depending on the aggressiveness of the threshold used at each window verification stage, a certain number of windows are allowed through. A threshold is chosen by selecting the value that allows the majority of samples through. Figure 3 shows the Probability of Detection vs the percentage of boxes remaining to be analysed. Assuming the desired detection rate is >95%, we can see that a threshold allowing 97% probability of detection eliminates 50% of the windows with the use FTS. At this operating point, we also see a significant reduction in the number of windows as compared to random scoring of windows of 25%.

FTS is best able to eliminate obvious non-human windows at high probability of detection range (>0.5), while VOFOD is best able to eliminate windows at low probability of

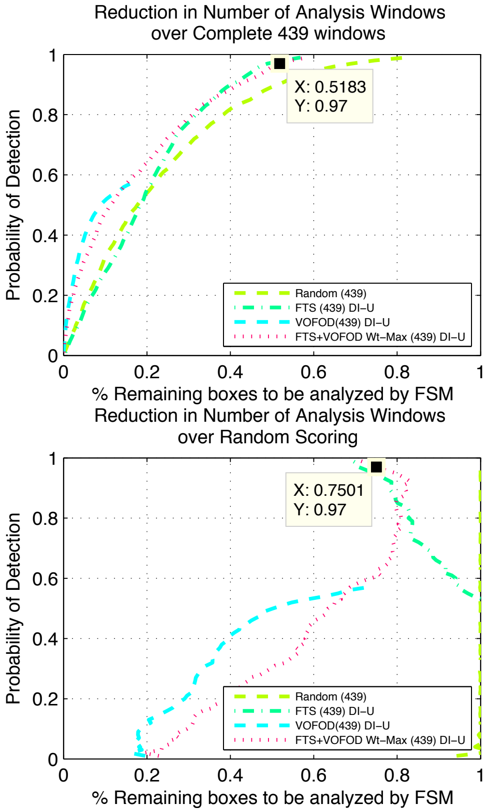

Fig. 3. Probability of Detection as a function of the percentage of boxes remaining for subsequent stage analysis. FTS, VOFOD, FTS+VOFOD are compared against the exhaustive initial set of windows (439) and random window scoring.

detection ($\leq 0.5$). Depending on the desired operating range, high probability of detection or very low false positive per image rate, either FTS or VOFOD will be the best verification module to use. However, if the desired operating point cannot be determined a priori, taking the maximum saliency value for each pixel to generate the fused saliency map from both VOFOD and FTS will yield a performance curve that is superior to random window scoring for the entire operating range. The curve for using the fused saliency map is shown in red, dotted plot in Figure 3.

## IV. Evaluation

We define the metric for performance on the probability of detection and false-positives per image. A detection is defined as a ground-truth human overlapping with a detection window by a sufficient amount. A detection window sufficiently overlaps a ground-truthed human if the areas of their intersection divided by the area of their union exceeds 1/3. The percentage of ground-truthed humans detected equals the probability of detection. Detection windows that do not overlap any ground-truth humans are false-positives. The mean number of false-positives per image is the false-positives-per-image. Each detection window has a confidence score. A threshold is used to determine which sets of detection windows are allowed through at the output. We sweep this threshold over the range of confidence scores to obtain a ROC curve with the axes probability of detection vs. false-positives-per-image (FPPI).

The dataset of children and humans taken from a backup camera is used for performance evaluation, consisting of approximately 26,000 frames of video with 0-6 individuals in each frame of video.

For performance purposes, we assume the high probability of detection range is the desired operating range. In this case, the VOFOD module is removed. The performance of the proposed system is shown in Figure 4. Four performance curves are shown: Random scoring, FTS scoring, FSM (parts-based classifier) scoring, and the cascade with FTS and FSM. Random scoring of the 439 windows achieves 300 FPPI at 95% probability of detection. With FTS, the number decreases to 220 FPPI, eliminating 80 false positives per image on average. FTS is thus shown to improve both accuracy and speed. FSM performance is shown in the dotted, green line and the cascade FTS-FSM model operates nearly at the same accuracy with 100 FPPI at 95% probability of detection. At the cost of a marginal decrease in accuracy, we make use of FTS to prune out 50% of the windows. It is worthwhile to note that state-of-the-art detectors evaluate systems at 1 FPPI with Pd of at most 40%. The plots emphasize the impact of the attentional mechanisms in reducing the number of windows that needs to be evaluated by the parts-based-classifier, which have been shown to be among the best performing detectors [14], [9], [6].

Because each window is operated on independently of the other windows, windows may be analysed in parallel, particularly useful for parts-based classifiers that require lengthy computations when operating sequentially. Because each stage can also operate independently of the other stages in the cascade, a pipeline architecture may be used to trade throughput (frames per second) for latency (time to result). Figure 5 shows one such parallel, pipelined architecture where each block indicates processing performed over time, and thin rectangles indicate the gates of the pipeline, where input images are placed in position for FTS processing, and the output of FTS is placed in position for parallel FSM processing. The timings are shown in the bar graph in Figure 6. Sequential processing of FTS and FSM are 84ms and 653ms, respectively. Parallel-FSM on a dual-core machine cut the time nearly in half to 327ms, reducing the total processing time to 327ms from 653ms. By applying pipelining, the processing time is further reduced to 248ms or 4 frames-per-second.

## V. Discussion and Conclusions

It is important to note that the usual box densities for sliding-window object detectors in literature is much greater [7], [15]. These bounding-box densities amounts to at least 12K candidate bounding-boxes, typically in the order of 100K. This minimum of 2 orders-of-magnitude reduction in the number of bounding-boxes is enabled by looser overlap criteria of 1/3 instead of 1/2, but more critically by the realization that a denser grid does not improve detection performance when the NMS is not used and additional

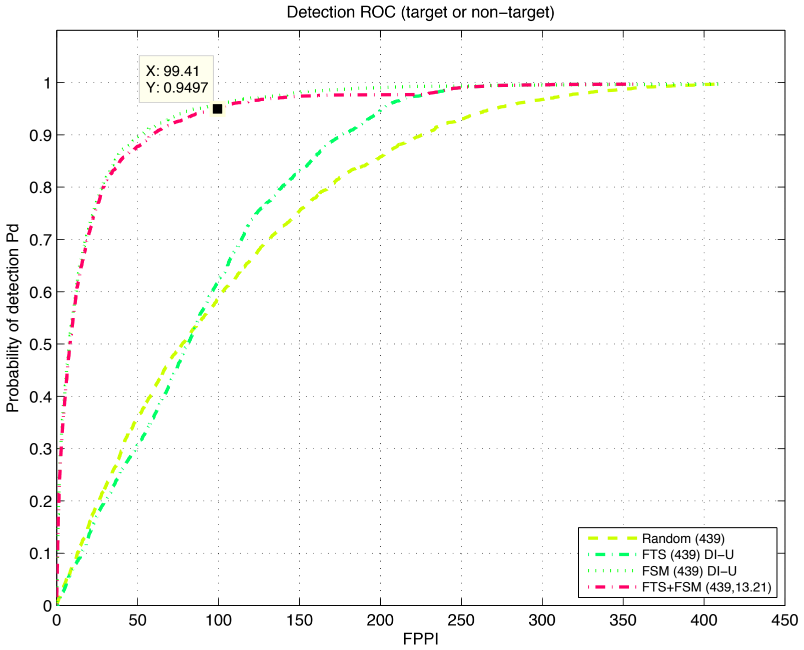

Fig. 4. Detection performance of Attention to Children with FSM parts-based classification.

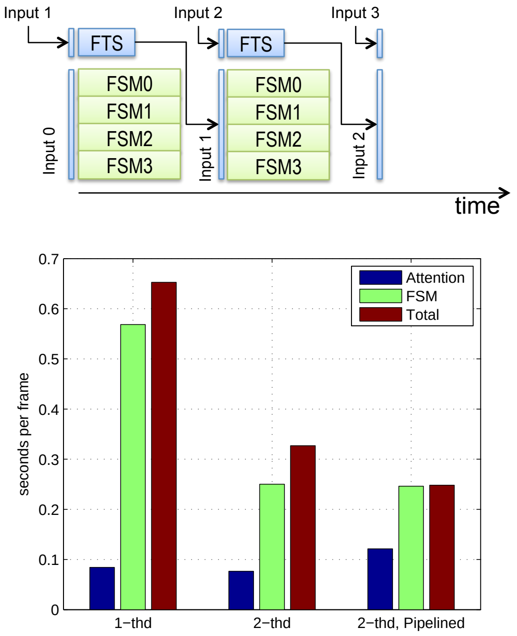

Fig. 5. Parallel computation and pipelining and the resulting compute times.

bounding-boxes overlapping a target is not counted as a false-positive. Without the use of NMS, denser grid of boxes only increases the false positives per image after the grid is dense enough to overlap all the objects. It becomes more important that the density is just enough to overlap all the targets.

This paper proposes a novel system for automatically detecting children from a color monocular back-up camera, as part of a back-up warning device in passenger vehicles. We presented the use of an attentional mechanism that focuses compute-intensive bounding-box classifiers on a subset of all possible bounding-box solutions to enable real-time performance of 248ms per frame with negligible reduction in performance. The attentional mechanism called *Attention to Children* which consists of a window generation and verification cascade of based on Frequency-Tuned Saliency, Variational-Optical-Flow Obstacle Detection and finally a parts-based classifier. We also presented a method of reducing much of the cascade classifier evaluations by judicious sampling of the bounding-box solution space. The result is a reduction in the number of windows evaluated down to 439 from more than 12K windows in traditional sliding window techniques, a 97% reduction in the number of windows. The verification stages leading up to the parts-based classifier further reduces the number of windows to half. Together with a parallel processing and pipelining, the final processing time was 248ms per frame.

## References

[1] Centers for Disease Control, “Nonfatal Motor-Vehicle-Related Back-over Injuries Among Children – United States, 2001-2003,” *MMWR Weekly*, vol. 54, no. 06, pp. 144–146, Feb. 2005.

[2] ——, “Injuries and Deaths Among Children Left Unattended in or Around Motor Vehicles -United States, July 2000-2001,” *MMWR Weekly*, vol. 51, no. 26, pp. 570–572, July 2002.

[3] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, “Part-based feature synthesis for human detection,” *European Conference on Computer Vision*, 2010.

[4] Ford Motor Company, “Hybrid ultrasonic and radar based backup aid,” no. 5754123, May 1996.

[5] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, “Cascade object detection with deformable part models,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2241–2248.

[8] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object?” *International Conference on Computer Vision and Pattern Recognition*, 2010.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 304–311.

[10] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context,” *International Journal of Computer Vision*, vol. 81, no. 1, Jan. 2009.

[11] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[12] J. Molineros, S. Y. Cheng, and Y. Owechko, “Monocular Rear-view Obstacle Detection Using Residual Flow,” in submission, 2012.

[13] T. Poggio and S. Bileschi, “A Multi-Scale Generalization of the HoG and HMAX Image Descriptors for Object Detection,” 2008.

[14] D. Levi and S. Ullman, “Learning Model Complexity in an Online Environment,” in *Computer and Robot Vision, 2009. CRV ’09. Canadian Conference on*, 2009, pp. 260–267.

[15] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.