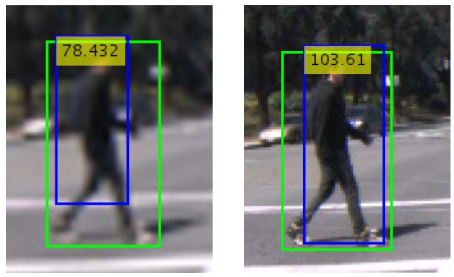
Looking at Pedestrians at Different Scales: A Multiresolution Approach and Evaluations

Rakesh Nattoji Rajaram, Eshed Ohn-Bar, and Mohan Manubhai Trivedi

***Abstract*—Typically, in a detector framework, the model size is fixed at the size of the smallest object to be detected, and larger objects are detected by scaling the input image. The information lost due to scaling could be vital for accurately detecting large objects, which is an essential task for vision-based driver-assistance systems. To this end, we evaluate a multiresolution detector framework by training models at different sizes and demonstrate its effectiveness on a state-of-the-art pedestrian detector. Our comprehensive evaluation demonstrates meaningful improvement in detector performance. On the KITTI dataset under moderate difficulty settings, we achieve a 6% increase in the detector's average precision over the baseline single-resolution result on the KITTI benchmark. Further insights into the detector's improvements are provided using a fine-grained analysis of the detector's performance at various threshold settings.**

***Index Terms*—Computer vision, fine-grained analysis, intelligent/safe vehicle, multiresolution model, pedestrian detection.**



Small model size. Large model size.

Fig. 1. Both small (25 pixels) and large (100 pixels) models are trained on the same dataset. Green box is the ground truth, whereas blue is the bounding box from the detector. Notice that the bounding box from the smaller model is poorly localized. In addition, corresponding confidence scores are different for the same object. Images are resampled to the same size for visual appeal.

I. Introduction

PEDESTRIAN safety is an important issue in the intelligent transport systems domain. Over the past decade, the essential role of computer vision in active safety systems for accident prevention is analyzed in detail by the authors in [1]. This in turn has led to a number of innovative ideas for pedestrian safety. Analysis on impact of appearance pattern on pedestrian detection [2], integrated framework for pedestrian trajectory prediction [3], a part-based pedestrian detection and feature-based tracking for driver assistance [4], estimating pedestrian orientation for improved path prediction [5], active pedestrian safety by automatic automobile maneuvering [6], driver attention monitoring [7] are some of the most recent progress in enhancing pedestrian safety.

Looking at pedestrians from the intelligent transportation system's point of view is very different from the classical case of object detection in a generic computer vision setting. Under this setting, images captured using a camera by an end user is typically well focused, is taken under adequate lighting condition and the primary object of interest occupies a significant portion of the image. Compared to this, frames extracted from a video sequence that is captured by a camera mounted on an intelligent vehicle has fixed focal length and hence cannot output sharp images of all the objects in the scene. Also, an object could appear differently under varying intensity and direction of natural light. Typically, pedestrians travel in groups and are constantly occluded by other pedestrians or objects such as trees, poles or cars. Even in such constrained settings, systems should be able to detect pedestrians in realtime to enhance their safety. In conclusion, we need a fast detector that is robust to object transformation, occlusion and truncation.

Objects appear differently when observed at different spatial resolutions. A person standing 25 pixels tall will look very differently from a person 100 pixels tall. The traditional pipeline for object detection involves learning a detector using the features extracted by scaling all the positive samples to a fixed size (called model size). This elegant approach lacks the necessary structure to exploit high resolution features, when available. Choosing a smaller template size will allow for the detection of smaller pedestrians at the cost of lower detector accuracy. On the other hand, a bigger model size will yield better detector accuracy for large pedestrians at the cost of missing out on smaller ones (unless the image is up-sampled at the expense of additional computational cost). This phenomenon is explained better with the example in Fig. 1. In this paper, we evaluate a pipeline for combining multiple models trained at different resolutions by studying its effect on detector AP.[1] Each model consists of decision trees trained using AdaBoost scheme with pixel lookup features for fast detection. Our framework is

Manuscript received December 6, 2015; revised March 2, 2016; accepted April 16, 2016. The Associate Editor for this paper was F.-Y. Wang.

The authors are with the Laboratory for Intelligent and Safe Automobiles (LISA), University of California San Diego, La Jolla, CA 92093-0434 USA (e-mail: rnattoji@ucsd.edu; eohnbar@ucsd.edu; mtrivedi@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2561262

[1] AP is defined as the area under precision vs recall curve times 100. This metric is chosen to standardize the comparison between the metrics in this paper with KITTI benchmark results.

implemented on top of the detector making it possible for any detector that uses pixel lookup features to be used. To demonstrate its effectiveness, we perform our experiments using the publicly available Piotr's toolbox [8].

The key contributions of this paper are as follows.

1) A principled study of multi-resolution detector framework on KITTI dataset is presented.
2) Learning parameters are studied for different scales showing optimal learning parameters to differ among scales.
3) A fine-grained performance analysis on different pedestrian sizes separately to better isolate and understand detection performance improvement. This analysis is inspired from [9].
4) Multi-resolution model training is carefully examined on KITTI, demonstrating significant detection performance gains on large pedestrians. This finding would be of great importance for driver assistance and collision warning systems.
5) An evaluation of the detector's performance with data collected in LISA lab to test the generalizability of the trained models.

Other detection schemes are expected to benefit from key findings of this paper. Aggregated channel features (ACF) detector [10] is used as an example, but the problem of scale is general over domains and objects and is important for complex real world on-road object detection tasks.

II. Related Research

State-of-the-art pedestrian detectors can be broadly divided into different categories based on channel selection, feature computation and classification stages. Most of the recent work [10]–[14] useRGB images as input but, some have ventured into including stereo [18], [19] or even 3D point clouds [20], [21] as an additional modality.

With RGB images, most of the recent work were based on using some combination of LUV color channels, Gradient Magnitude (GM), Histogram of Oriented Gradients (HOG) [22] and/or it's derivatives as the computed channels. This include, but not limited to integral channel features ICF [12], aggregated channel features (ACF) [10] and SubCat [17].

While ACF used a simple channel aggregation technique as features, ICF used harr features on top of LUV + a HOG variant (vHOG) + GM. Recent studies demonstrate the use of more complex filters on top of these channels to achieve higher detection accuracy. LDCF [23] implements a feature transformation function to remove the correlation between neighboring features. SquaresICF [24] learns multiple irregular sized windows to aggregate the extracted features. FilteredICF [25] uses combination of checkerboard, LDCF, PCA and square channel filters to extract features that improve pedestrian detector accuracy.

With convolution neural networks (CNN) leading the pack in multi-object classification tasks [26], some of the recent methods make use of the features derived from a CNN. For example, R-CNN [13] first minimizes the search space from millions of windows to a few thousand probable windows and then extracts CNN features from each window using a model that is fine tuned on a particular dataset. This high dimensional feature is then passed on to a support vector machine classifier. CCF [14] uses the same framework as ACF but, adds additional channels from a CNN. This leads to improved detector performance but, the CNN features disregard for power-law makes the feature pyramid construction step computationally expensive.

All the work mentioned above used a single rigid template to perform detection. But, pedestrians have a wide variation in appearance due to varying orientation, clothing and physical activity. To tackle such intra class variations, Deformable parts based models (DPM) [11] formulates an object as a root template and a number of associated parts whose position is flexible relative to the root template. Regionlets [27], [28] introduces appearance flexibility in the feature space. It operates by minimizing the search space to a few thousand windows (similar to R-CNN), extracting features from a fixed number of regions inside these windows, and then pooling them to establish invariance to localization, scale and aspect ratio. Next, the detected objects are re-localized using a localization model. SubCat [17] introduces modifications on top of the detector. Here, objects are sub-categorized into a fixed number of clusters based on geometric features such as height, width, aspect ratio, occlusion etc. and aggregated channel features. Then, a separate model is trained for each of these clusters. Along with improving detector accuracy, SubCat also improved orientation estimation accuracy.

While all the above methods also apply for generic object detection, pedestrian detection using a camera mounted on moving vehicle can benefit from additional geometric constraints. MT-DPM [16] introduced resolution aware transformations to map pedestrians in different resolutions to a common space and also makes use of a context-aware model to suppress false positive windows. ICF-MR [18] trains a model for each scale in the feature pyramid and detects pedestrian at multiple scales without re-sizing the image. It also makes use of stereo information to estimate depth which is used to reduce the search space. This leads to a faster detector. ACF-SC [29] makes use of semantic segmentation and context information to improve the ACF detector's detection quality.

In this paper, we evaluate the performance of our multi-resolution approach. Some of the recent work making use of similar approach includes MultiRes [15] and MT-DPM [16]. MultiRes modifies DPM to incorporate multiple models trained at different scales and fuse context information to improve detector AP. This modification is integrated within the DPM framework making it difficult to be implemented on top of other detectors. MT-DPM uses two models trained at different resolution and remodels DPM training framework to add information from both models. On the other hand, our proposed approach examines separate scale-specific models, as opposed to learning a joint two-resolution model over multi-resolution features. Both MultiRes and MT-DPM have not reported results on KITTI dataset. Our approach differs from existing multi-resolution approach, ICF-MR [18] in the following ways.

1) ICF-MR employs inverted pyramid and is only reported on Caltech dataset. We use only 2 additional scales

TABLE I
COMPARISON OF SELECTED STATE-OF-THE-ART PEDESTRIAN DETECTORS

Study	Name	Modality	Channels	Features	Classifier	Parts	Multi-Res	Speed
Felzenszwalb et al., 2010 [11]	DPM	RGB	-	HOG	Latent-SVM	Yes	No	10sec (640x480)
Dollar et al., 2009 [12]	ICF	RGB	LUV+HOG+GM	Haar	Decision Tree	No	No	2sec (640x480)
Dollar et al., 2014 [10]	ACF	RGB	LUV+HOG+GM	Aggregated Channel	Decision Tree	No	No	0.2sec (1242x375)
Girshick et al., 2014 [13]	R-CNN	RGB	-	CNN	SVM	No	No	4sec (1242x375)
Yang et al., 2015 [14]	CCF	RGB	LUV+HOG+GM	Aggregated Channel + CNN	Decision Tree	No	No	16sec (1242x375)
Park et al., 2010 [15]	MultiRes	RGB	-	HOG	Latent-SVM	Yes	Yes	Not Reported
Yan et al., 2013 [16]	MT-DPM	RGB	-	HOG	Latent-SVM	Yes	Yes	Not Reported
Ohn-Bar et al., 2015 [17]	SubCat	RGB	LUV+HOG+GM	Aggregated Channel	Decision Tree	No	Yes	1.2sec (1242x375)
Benenson et al., 2012 [18]	ICF-MR	RGB+Stereo	LUV+HOG+GM	Haar	Decision Tree	No	Yes	0.01sec (640x480) on GPU
This study	ACF-MR	RGB	LUV+HOG+GM	Aggregated Channel	Decision Tree	No	Yes	0.6sec (1242x375)

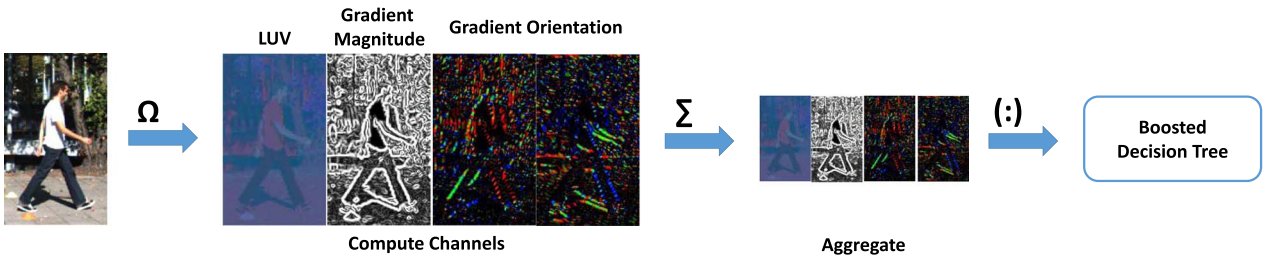


Fig. 2. Pipeline of the vanilla ACF detector [10]. LUV color channels, gradient magnitude, and histogram of gradient orientations are computed from the input RGB image as $\mathrm{C} = \Omega(\mathrm{I})$. These feature channels are sum pooled and passed through a smoothing filter, resulting in lower resolution aggregated channels. Boosting is used to learn decision trees using a vectorized aggregated channel to classify as an object or background.

and report our results on both KITTI and Caltech datasets.

2) While ICF-MR has not reported any score calibration experiment, we show its importance in improving detector AP.
3) The aim of ICF-MR was to implement a faster detector whereas our aim is to perform experiments and analysis to answer improvement in detector AP.

Selected methods are compared in Table I.

III. PROPOSED MULTI-RESOLUTION MODEL

In this section, the proposed multi-resolution detector framework is described. The idea is to train multiple models in different sizes and then during testing, run all these models on the corresponding scales of the feature pyramid and concatenate the bounding boxes derived from each model. The subsequent sections will provide an in-depth explanation.

This approach can be easily implemented with any detector that uses pixel lookup features and hence improvement in detector accuracy due to better feature selection or development of a better classifier is orthogonal to our analysis.

A. Vanilla ACF Detector

Our multi-resolution detector consists of several single resolution model trained using the ACF detector [10]. A brief outline of this single resolution detection pipeline is shown in Fig. 2.

Let $\mathbf{I}$ be an RGB input image patch of size $m \times n \times 3$. Channels are computed as $\mathbf{C} = \mathbf{\Omega}(\mathbf{I})$, where $\mathbf{C}$ is a matrix of size $m \times n \times 10$. Ten channels used are LUV color space, 6 bin unsigned oriented gradient histogram and normalized gradient magnitude. Features are single pixel look-ups in the aggregated channels which are computed with Equation (1) where, w_s is the aggregation local window size and $\mathbf{F}$ is a 3D matrix of size $\left\lfloor \frac{m}{w_s} \right\rfloor \times \left\lfloor \frac{n}{w_s} \right\rfloor \times 10$.

$$\mathbf{F}(i,j,k) = \sum_{x=iw_s}^{iw_s+w_s-1} \sum_{y=jw_s}^{jw_s+w_s-1} \mathbf{C}(x,y,k)$$
$$0 \le i < \left\lfloor \frac{m}{w_s} \right\rfloor, 0 \le j < \left\lfloor \frac{n}{w_s} \right\rfloor, 0 \le k < 9 \qquad (1)$$

To remove noise, smoothing filters are applied to the input image and also to the computed features. Boosting is used to learn random forest classifier over these vectorized features.

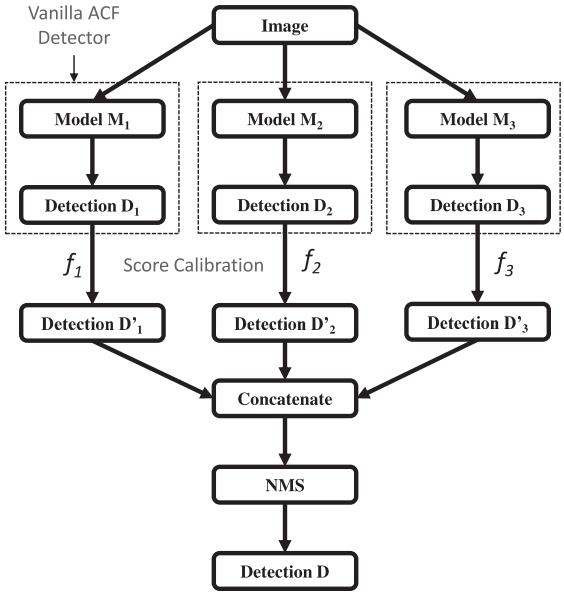


Fig. 3. Testing framework of the multiresolution ACF detector.

During the testing stage, some scales in the feature pyramid are approximated using power-law to find a trade-off between detector accuracy and runtime.

B. Training the Multi-Resolution Detector

Training process is identical to vanilla ACF training with multiple models trained independently using different parameters. Let M_i be the ACF model trained on the training set using parameter set P_i and annotations A_i. P_i includes model size, model padding, decision tree depth, aggregation local window and other constants (i.e., not varied according to i). Annotations A_i is derived from the original annotations by ignoring objects that are shorter than the model height. For example, a model with model size of 50 pixels by 20 pixels will be trained only on pedestrians who are taller than 50 pixels in height. Since each model is trained independently, the overall training time is roughly scaled by the number of models to be trained.

C. Testing the Multi-Resolution Detector

Identical to vanilla ACF, during the testing stage, detections are generated for each model independently and later combined using one of the methods discussed below. This pipeline is represented in Fig. 3. Since all the models use the same channels, channel pyramid can be shared among different model. In-fact, if the ACF feature pooling local window size (w_s) is the same for different models, then the feature pyramid itself can be shared between multiple models. Or, if they form a geometric progression, feature pyramid can be recursively calculated without additional computation time. Since feature pyramid computation is the most expensive operation during runtime, multi-resolution detector runtime does not scale by the total number of models.

The method proposed to generate the joint detection D is as follows. Let D_i be the detection bounding boxes generated by model M_i. $D_i = \{d_{i1}, d_{i2}, \ldots d_{ij}, \ldots\}$, $1 \leq j \leq n_i$, where n_i is the number of objects detected by model M_i. Each detected bounding box is written as $d_{ij} = \{R_{ij}, c_{ij}\}$, where R_{ij} defines the object boundary with confidence score c_{ij}. We create a new detection set for each model M_i as $D'_i = \{d'_{i1}, d'_{i2}, \ldots d'_{ij}, \ldots\}$ where $d'_{ij} = \{R_{ij}, c'_{ij}\}$ and $c'_{ij} = f_i(c_{ij})$. Here f_i is the score transformation function for model M_i. We experiment with the following score transformation functions and report their impact on detector performance.

1) *No Transformation*: $c'_{ij} = c_{ij}$
2) *Linear Transformation*: $c'_{ij} = k_i c_{ij}$ where k_i is the scaling factor for each model.
3) *Affine Transformation*: $c'_{ij} = k_i c_{ij} + o_i$ where k_i is the scaling factor and o_i is the score offset for each model.
4) *Min-Max Normalization*:

$$c'_{ij} = \frac{\left(c_{ij} - c_i^{\min}\right)}{\left(c_i^{\max} - c_i^{\min}\right)}$$

$$c_i^{\min} = \min_j\{c_{ij}\}, \quad c_i^{max} = \max_j\{c_{ij}\} \quad 1 \leq j \leq n_i$$

5) *Sigmoid Normalization*:

$$c'_{ij} = \frac{1}{1 + e^{-k_i c_{ij}}}$$

Why a simple concatenation may not work is reasoned in Fig. 1 where the confidence scores (and the bounding boxes) generated by different models on the same pedestrian produced dissimilar scores. Scores can also be calibrated to give more importance to higher-resolution pedestrians as they are close to the vehicle. For a further study of model score normalization and combination for possible further gains in performance, reader is refered to [30]. Yet, simple transformations were shown to work well (Table V).

The scaled detections D'_i from each models are concatenated and Non-Maximal Suppression (NMS) is applied to this superset to generate the final detections D.

IV. Datasets

A. KITTI

KITTI dataset [31] is captured by driving around the city of Karlsruhe, Germany, in rural areas and on highways. Frames are extracted from videos captured at 10 fps, thereby generating 7481 training images and 7518 test images. These images are then cropped to resolution of 1242 $\times$ 375 and thereafter randomized. To run our experiments a validation set was created as follows. Apply the inverse mapping to remove the random jumbling of images and then separate them into corresponding video clips. This generates 144 video sequences with an average of 52 frames each. They are divided into training and validation sets such that each set contains comparable number of frames and pedestrians. This is achieved by sorting the video according to the number of pedestrians and manually assigning videos to different sets. We had 3742 and 3739 images in our training and

(a) (b)

Fig. 4. Sample images with annotations from the two studied datasets. On comparison with images from the Caltech dataset, images from the KITTI dataset are captured using a camera that has a higher field of view and produces sharper images.

validation sets respectively. Sample images with annotations from KITTI dataset are shown in Fig. 4. Since multi-resolution approach and the comprehensive evaluation depends on the distribution of pedestrian height and aspect ratio,[2] we have included their histogram plots in Fig. 5. KITTI differentiates the difficulty in identifying pedestrians based on height, occlusion and truncation which is summarized in Table II.

B. Caltech-USA

The common evaluation split is performed [32], where the first six out of the 10 available sets of data are split into training and the remaining for testing. Each video clip has a resolution of 640 $\times$ 480 and is recorded at 30 frames per second. By periodically sampling at 1 second, we extract about 60 frames per video clip. This translates to 4250 and 4024 images in the training and testing set, respectively.

First, for performing the experimental analysis, we seek to standardize evaluation among KITTI and Caltech. For instance, Caltech does not annotate truncation, hence the moderate settings constraints cannot be enforced directly on Caltech. Furthermore, KITTI has qualitatively annotated occlusion as 3 different categories, whereas Caltech has annotated occlusion by providing bounding box of occluded portion of the ground truth.

A qualitative mapping between the evaluation constraints across the datasets is described below. By observing the occluded samples from KITTI, a minimum visibility of 65% was enforced on Caltech. Truncation was handled using the following algorithm to avoid manual annotation. First, a canonical aspect ratio (a), was assumed. Next, to get the truncation value for each pedestrian, a subset of all pedestrians who are very close to the left or right or bottom boundaries was constructed. Using bounding box height h (or width w if truncation is at the bottom), the expected width $w_e = ha$ (height $h_e = w/a$) is calculated. Finally, truncation is estimated as $t = (w_e - w)/w$ $(t = (h_e - h)/h)$. All pedestrians with truncation greater than 0.3 are ignored. Under these constraints, Caltech has 3106 and 2425 pedestrians in training and testing sets respectively. On the other hand, KITTI has 1785 and 1784 pedestrians in training and validation sets respectively. So with comparable number of images in both Caltech and KITTI, Caltech has significantly higher number of pedestrians.

Note that in [32], the author had to evaluate many detectors that were trained using different aspect ratio and hence was standardized. However, in this paper, as we have full control over the training parameters, no aspect ratio standardization is performed in the experiments.

[2]Aspect ratio is defined as the ratio of bounding box height to bounding box width.

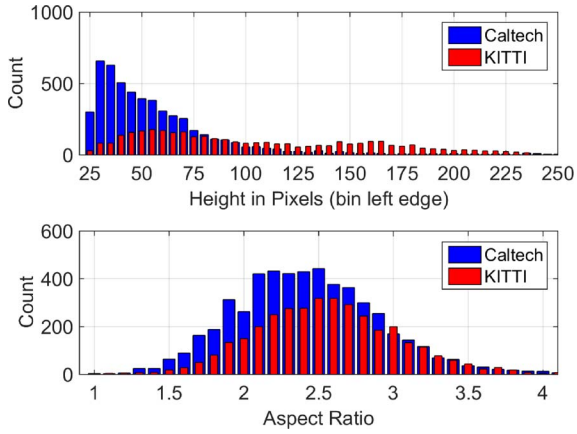


Fig. 5. Comparison of the pedestrian height and aspect ratio across datasets. Although the Caltech dataset has more annotated pedestrians, the KITTI dataset has a better distribution of pedestrian heights. Caltech dataset has a larger fraction of pedestrians shorter than 50 pixels in height. In addition, the Caltech dataset has a slightly larger variation in the aspect ratio.

TABLE II
DISTRIBUTION OF PEDESTRIANS IN THE KITTI DATASET INTO DIFFERENT DIFFICULTY SETTINGS BASED ON HEIGHT, OCCLUSION, AND TRUNCATION

Difficulty	Height	Occlusion	Truncation
Easy	40	Fully Visible	15%
Moderate	25	Partially Occluded	30%
Hard	25	Difficult to See	50%

TABLE III
SINGLE-RESOLUTION MODEL DETECTION RESULT AT DIFFERENT DIFFICULTY SETTINGS ON THE KITTI DATASETS. ADDING HARD POSITIVE SAMPLES TO THE TRAINING SET APPEARS TO HAVE A LIMITED EFFECT ON DETECTOR ACCURACY

		Testing Set		
		Easy	Moderate	Hard
Training Set	Easy	77.29%	58.12%	50.03%
	Moderate	**78.37%**	**61.96%**	**54.14%**
	Hard	77.20%	61.14%	53.12%

V. EXPERIMENTAL ANALYSIS

A. Overall Training Parameters

On both the datasets, ACF detector was trained with a maximum of 4096 decision trees using AdaBoost. Four rounds of hard negative mining were performed. In each round, 25,000 negatives were randomly mined and up to 50,000 of the hardest negatives were employed. Horizontally flipped versions of pedestrians windows were also included as positive samples.

B. Single Resolution ACF Result

1. KITTI: Various parameters of the detector were grid-optimized to maximize AP at each difficulty setting. Model aspect ratio was fixed at 2.5, the median of the aspect ratio distribution. In order to include small pedestrians (height, $h < 50$) at a larger template size, experiments were performed with two different values of the parameter, number of octave up (δ).[3] Highest AP was obtained using a model with height $m_h = 50$ and width $m_w = 20$ (with padding $m_h = 64$ and $m_w = 32$). We call this model $M_{50\text{ s}}$ where, 50 is the model height and s indicates single resolution. At $\delta = 0$, under moderate difficulty settings, detector achieved 61.96% AP. However, at $\delta = 1$, AP decreased to 60.82%. This suggests that an increase in recall by detecting pedestrians smaller than 50 pixels is shadowed by the reduction in precision due to increase in false detections. Table III summarizes the detection AP on permutations of difficulty settings on the KITTI dataset. It appears that the detector trained on moderate difficulty settings performs "marginally" better compared to training on other difficulty settings, irrespective of the testing set difficulty. However, the difference is too small to draw any conclusion.

2. Caltech-USA: Nature of this experiment is similar to the one performed on KITTI dataset. Experiments are started with optimal parameters already available from PMT [8]. These are optimal for "reasonable" difficulty settings (pedestrian taller than 50 pixels and have at least 65% visibility). Maximum AP at 46.54% was achieved with $m_h = 50$, $m_w = 20.5$ and $\delta = 1$. Positive samples used for training were changed to include pedestrians taller than 25 pixels (instead of 50) and visibility greater than 65%. We will refer to this model as $M_{50\text{ s}}$.

C. Multi-Resolution ACF Result

1. KITTI: Experimental result from Table III suggests moderate difficulty setting to be well suited for training ACF detector on KITTI datasets. Also, under hard difficulty settings, around 2% of the pedestrians were not recognized by humans [31]. Hence, we analyze the multi-resolution model result under moderate difficulty settings. We trained models with different parameters (model size—$[m_h\ m_w]$, padding, tree depth, shrink) and our experiments suggested that by using just 3 models, most of the detector AP gain could be achieved. Aspect ratio was fixed at 2.5. Each model was tested on a subset of pedestrians taller than the model height (i.e. $\delta = 0$). Other parameters were tuned to maximize AP. AP started decreasing for $m_h > 100$. This is likely due to lack of sufficient positive training samples. We choose $M_{25\text{ m}}$, $M_{50\text{ m}}$, and $M_{100\text{ m}}$ for further experiments. Here, m stands for multi-resolution. Optimal parameters are tabulated in Table IV.

TABLE IV
MULTIRESOLUTION MODEL PARAMETERS USED FOR TRAINING THE ACF DETECTOR ON THE KITTI DATASET

Model	Size	Size + Pad	Tree Depth	w_s	#Trees
M_{25m}	[25 10]	[32 16]	4	1	2624
M_{50m}	[50 20]	[64 32]	4	2	1808
M_{100m}	[100 40]	[128 64]	3	4	1872

TABLE V
MULTIRESOLUTION ACF DETECTOR PERFORMANCE WITH DIFFERENT CONFIDENCE TRANSFORMATIONS ON THE KITTI DATASET. LABEL p_i^1 IS k_i FOR LINEAR, AFFINE, OR SIGMOID TRANSFORMATION AND c_i^{min} FOR THE MIN–MAX TRANSFORMATION. LABEL p_i^2 IS o_i FOR THE AFFINE TRANSFORMATION AND c_i^{max} FOR THE MIN–MAX TRANSFORMATION

		None	Linear	Affine	Min-Max	Sigmoid
AP	All	58.51	63.58	63.72	58.85	63.64
p_i^1	M_{25m}	-	0.5	0.8	-0.95	0.09
	M_{50m}	-	1.1	1.0	-0.92	0.18
	M_{100m}	-	1.0	1.0	-0.66	0.17
p_i^2	M_{25m}	-	-	-30	250.1	-
	M_{50m}	-	-	10	236.1	-
	M_{100m}	-	-	0	261.6	-

To get the overall precision vs recall (PR) curve on moderate difficulty settings, we transform confidence score as discussed in Section III-C. The parameters are grid-optimized to maximize AP. Results along with parameters are tabulated in Table V. Without score transformation the multi-resolution detector perform poorly compared to the single resolution counterpart. Min-max normalization performed poorly due to all the models having similar score limits. As a result, score normalization had no impact on the final scores. Also, linear, affine and sigmoid transformations yield very similar results. Hence, all further experiments will be performed solely with linear transformation.

Fig. 6 plots the PR curves and compares the detector AP between single and multi-resolution detector for different range of pedestrian height. We see that curve for $M_{100\text{ m}}$ is always above the curve for $M_{50\text{ s}}$, evaluated for $25 < h < 50$. Since $M_{50\text{ s}}$ and $M_{50\text{ m}}$ are trained using the same set of annotations and parameters they are interchangeable and hence produce identical result. However, $M_{25\text{ m}}$ did not perform as expected. Initially the curve for $M_{25\text{ m}}$ is above the corresponding single resolution curve ($M_{50\text{ s}}$, tested with $\delta = 1$), but falls off rapidly with increasing recall value, suggesting over-fitting.

[3]Image is upsampled by 2^δ in both spatial dimensions.

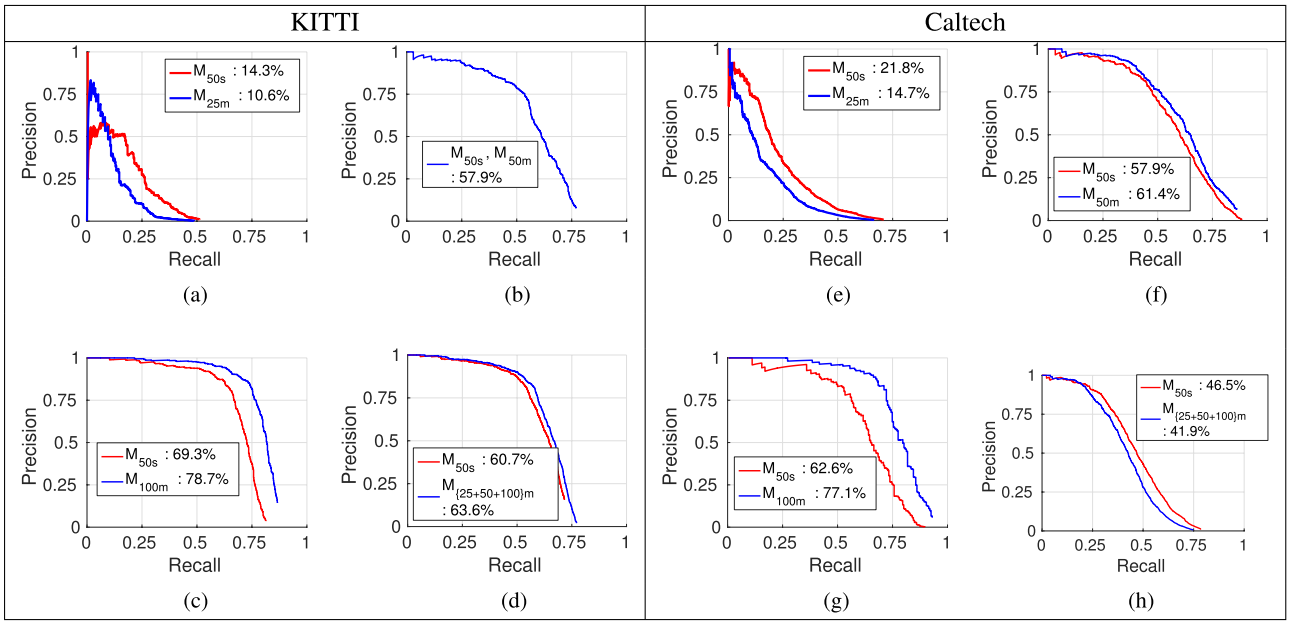


Fig. 6. Comparison of the PR curve for single and multiresolution models using the ACF detector. Curves in red (blue) correspond to a single (multi) resolution detector. See the legend in the plot for the model(s) used and the corresponding AP achieved. Testing height range (h) is mentioned as a subcaption under each plot.

Overall, $M_{50\text{ s}}$ does a better job of detecting smaller pedestrians than $M_{25\text{ m}}$. This observation is likely due to the following reasons.

- Although up-sampling does not add additional information, aggregating channels could still improve detection accuracy.
- HOG channels calculated using a fixed filter i.e., $[-1\ 0\ 1]$ benefits by up-sampling image when detecting small objects.

In Table IV, we see that the model sizes are multiple of 2 and so is aggregation window w_s. Therefore, all the models use same number of features $((h/w_s)(w/w_s)c = (64/2) \times (32/2) \times 10 = 5120)$ but, bigger models are performing better than the smaller ones. This is likely due to the following reasons.

- Pedestrians at lower resolution are inherently difficult to detect due to bias in data collection, i.e., the testing samples are difficult cases.
- Although each model has the same number of features, without aggregation, the HOGv features are sparse i.e., out of 6 bins at-most 2 bins are populated per pixel, but, by aggregating over large w_s more bins are likely to be populated.

2. Caltech-USA: Multi-resolution experiments performed on Caltech dataset is similar to the one performed on KITTI dataset. One difference from single resolution experiment is that the multi-resolution models are trained with all positive samples taller than the corresponding model size. Optimal parameters are tabulated in Table VI.

TABLE VI
MULTIRESOLUTION MODEL PARAMETERS USED FOR TRAINING THE ACF DETECTOR ON THE CALTECH DATASET

Model	Size	Size + Pad	Tree Depth	Shrink	# Trees
M_{50s}	[50 20]	[32 16]	5	2	3440
M_{25m}	[25 10]	[32 16]	5	1	3856
M_{50m}	[50 20]	[64 32]	5	2	1552
M_{100m}	[100 40]	[128 64]	3	4	1408

Fig. 6 plots the PR curves and compares the detector AP between single and multi-resolution detector for different range of pedestrian height. Similar to our experiments on KITTI dataset, $M_{25\text{ m}}$ performed worse than $M_{50\text{ s}}$ (with $\delta = 1$) under $25 < h < 50$, whereas $M_{50\text{ m}}$ and $M_{100\text{ m}}$ models perform better than $M_{50\text{ s}}$ under $50 < h < 100$ and $100 < h < \infty$, respectively. However, due to the bias in distribution of pedestrian height i.e., a significant amount of pedestrians smaller than 50 pixels leads to $M_{50\text{ s}}$ performing better than $M_{25\text{ m}} + M_{50\text{ m}} + M_{100\text{ m}}$. But, if we take a look at "reasonable" difficulty settings, $M_{50\text{ m}} + M_{100\text{ m}}$ performs slightly better than $M_{50\text{ s}}$ $(\delta = 0)$ (67.24% vs 66.91%).

This improvement is smaller than what was achieved on KITTI dataset and is most likely due to the lack of significant amount of pedestrians taller than 100 pixels.

VI. COMPARATIVE PERFORMANCE ANALYSIS

In order to improve any detector, it is crucial to understand where it fails. We propose to perform this analysis using the miss rate vs false positives per image (FPPI) plot. Most likely reasons for detector misses can be visualized in Fig. 7. Colored area under each reason corresponds to the fraction of miss rate

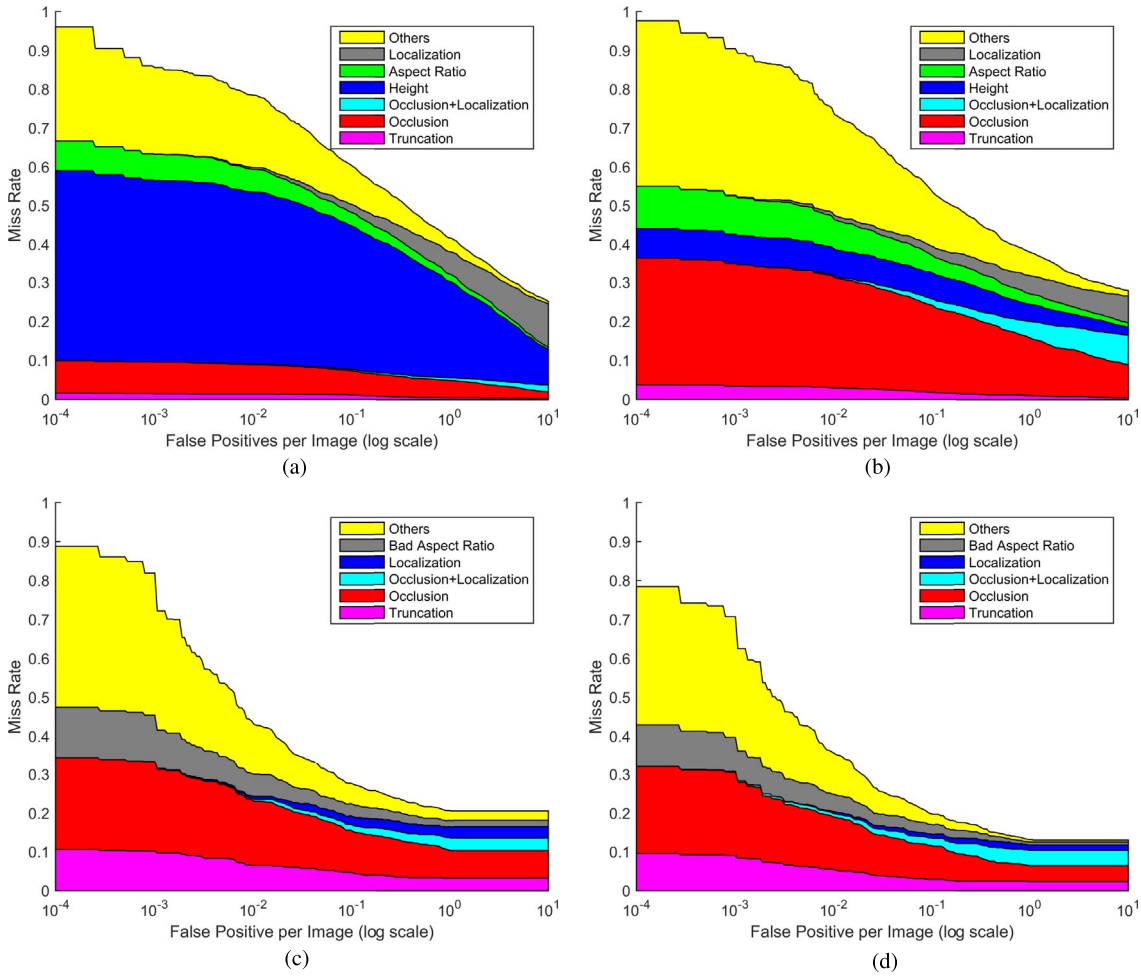


Fig. 7. Failure analysis at various FPPI values. Truncation and occlusion have their usual meaning under moderately difficulty settings. Localization error is defined as the fraction of the missed pedestrian who would have been detected if the minimum overlap threshold criteria were reduced from 50% to 20%. Miss due to height is the fraction of the missed pedestrian who was smaller than 50 pixels in height. Aspect ratio corresponds to the case in which the annotated ground truth box exhibits a large deviation from the model aspect ratio. (a) Caltech: $M_{50\text{ s}},\ 25 < h < \infty$. (b) KITTI: $M_{50\text{ s}},\ 25 < h < \infty$. (c) KITTI: $M_{50\text{ s}},\ 100 < h < \infty$. (d) KITTI: $M_{100\text{ m}},\ 100 < h < \infty$.

most likely contributed by them. The detector failure cases are categorized into several cases as follows. Localization error is defined as the fraction of missed pedestrians that would have been detected if the minimum overlap threshold criteria was reduced from 50% to 20%. Miss due to height of the pedestrian is defined as the fraction of missed pedestrians that were not detected and are smaller than 50 pixels in height. This is relevant to KITTI on account of higher detector AP at $\delta = 0$. Aspect ratio failure corresponds to the cases where the ground truth boxes have aspect ratio that largely deviates from the average aspect ratio. Specifically, if a missed ground truth box has an aspect ratio larger than 3 or smaller than 2, it is considered a failure due to aspect ratio. Some cases which cannot be resolved using the aforementioned attributes are marked as 'others'. Priority when generating the plot is in the order of truncation, occlusion, height, aspect ratio, localization, and others. Occlusion exhibits significant correlation with localization error especially at higher FPPI, and hence a category of occlusion and localization occurring together was added (this was not the case for truncated samples). For each of these cases, Fig. 8 provide examples from both datasets.

A. *Analysis on KITTI Dataset*

It appears that occluded pedestrians are very hard to detect on KITTI dataset. The dataset has a smaller fraction of pedestrians who are small however, a significant fraction of them are occluded. Another interesting aspect to consider is the correlation between occlusion and localization error especially at FPPI higher than 10^0 (light blue colored area). This typically happens when there is a vertical pole or another pedestrian nearby.

It would be interesting to know objects that are occluding pedestrians. Thanks to the rich annotation provided for the KITTI dataset, to some extent, we can analyze the distribution of occludee. With the help of depth information we can find a set of annotated objects that are in front of the occluded pedestrian. For this subset, we calculate the overlap area of the missed pedestrian with the occludes. Occludees causing overlap above 5% are added to the distribution. Fig. 9 plots this distribution.

Although $M_{100\text{ m}}$ has higher detector AP compared to $M_{50\text{ s}}$, from Fig. 7(c) and (d) we see that most of this gain is contributed by fully visible pedestrians (yellow and gray colored regions). This suggests that the $M_{100\text{ m}}$ model does not implicitly learn to detect occluded and truncated pedestrians.

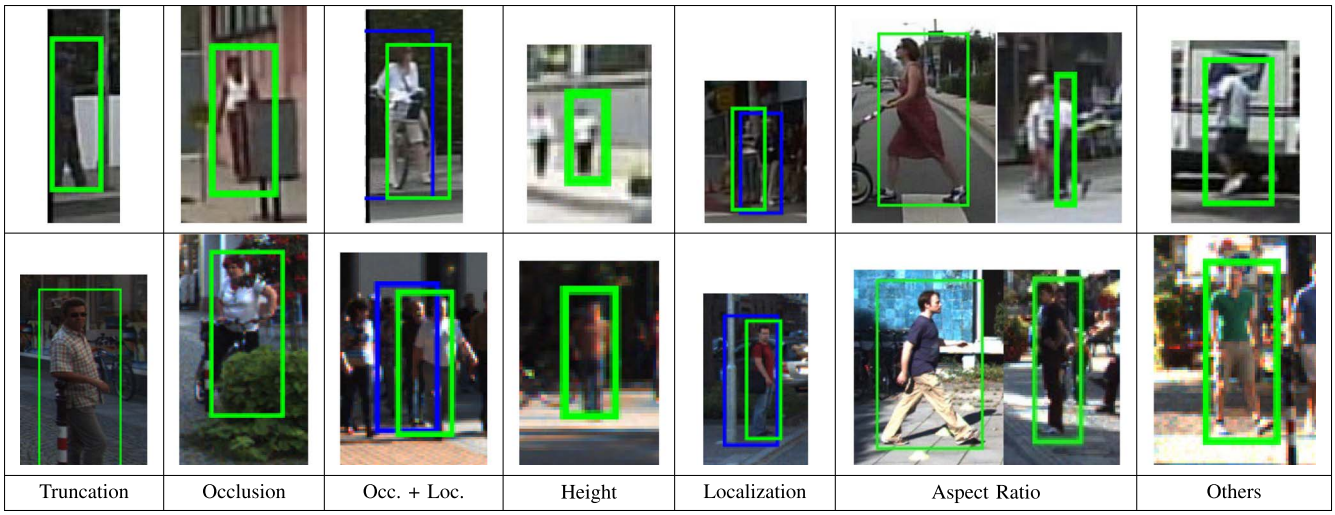


Fig. 8. Examples of different possible reasons for missed detections. Green boxes indicate the ground truth. Blue boxes are the detections, which fail the 50% overlap criteria. These failures are categorized as localization errors. Top row shows the samples taken from Caltech, whereas the bottom row shows samples from KITTI.

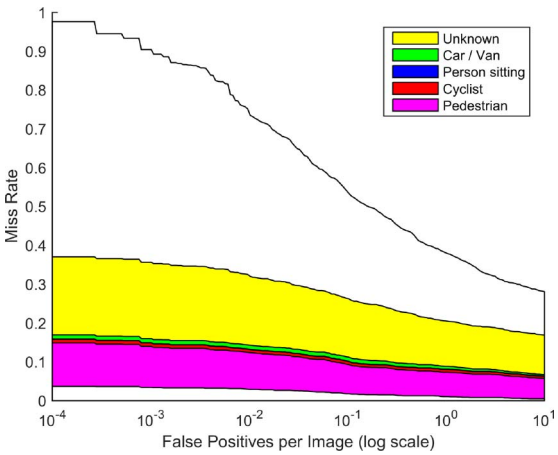


Fig. 9. Failure analysis of occluded pedestrians at various FPPI values on the KITTI dataset. Since occlusion is a major contributor of misses on the KITTI dataset, we further analyze the occluded distribution. Color-shaded regions correspond to the fraction of occluded objects causing occlusion and regions without color correspond to objects that are not occluded.

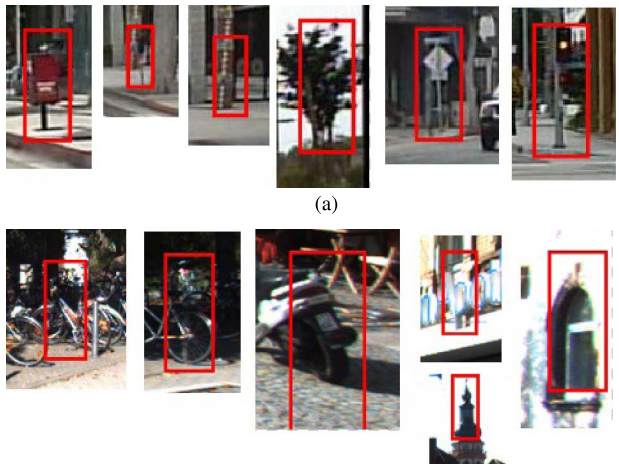


Fig. 10. Most likely false-positive windows generated by the detector at 10^{-2} false positives per image. Notice the similar strong horizontal gradient component in all these samples. (a) High-scoring FP windows from the Caltech dataset. (b) High-scoring FP windows from the KITTI dataset.

B. Analysis on Caltech Dataset

When it comes to pedestrian detection on Caltech dataset, failure can mainly be attributed to the dataset bias with respect to pedestrian size. Since it has more than double the fraction of small pedestrians compared to KITTI dataset, the failure seems to heavily favor this reason. At FPPI higher than 10^0, we see that the detector detects these small pedestrians albeit with lower confidence. Occluded and Truncated pedestrians seem to be almost never detected.

C. Discussion

In order to analyze where the detector fails, let us take a look at the miss rate slice around 10^{-1} FPPI. At this setting, irrespective of the datasets and model size, reasons for failure can be majorly attributed to occlusion and lack of sufficient resolution (i.e. small pedestrians). Although truncated pedestrians are hardly detected, the number of instances is small. This implies that although challenging, addressing other challenges than truncation may provide more significant improvement in performance. Selected high scoring false positives are visualized in Fig. 10. One particular aspect that is common to all these images is the presence of strong horizontal gradient. This suggests that the detector favors gradient information over color information and gets very confident about the presence of pedestrian under the influence of strong horizontal gradient component. Some of the false positive detections do not lie on the ground plane, while some that are far away from camera, are very tall. By incorporating the approach presented in [33], we can enforce geometric constraints to remove some of these false positive detections.

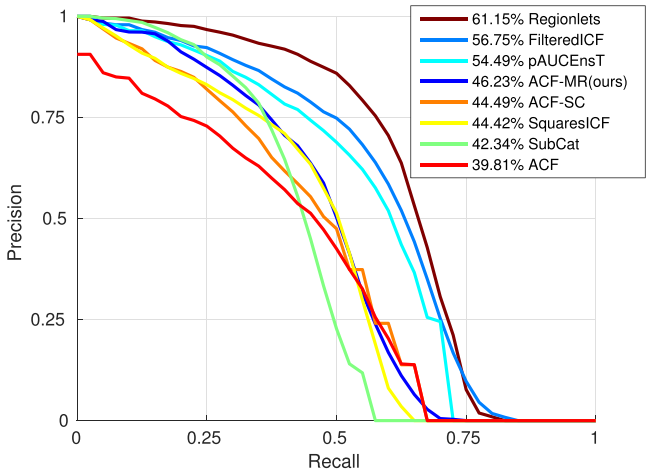


Fig. 11. Comparison of pedestrian detectors on the KITTI evaluation server under moderately difficulty settings. Detector AP is mentioned in the legend.

TABLE VII
COMPARISON OF ACF-BASED APPROACHES, EVALUATED ON THE KITTI EVALUATION BENCHMARK

Name	Addition	Accuracy	Runtime(sec)
ACF	-	39.81%	**0.2**
ACF-SC	Context + scene segmentation	44.49%	0.3
ACF-MR	Multi-resolution	**46.23%**	0.6

VII. MULTI-RESOLUTION ACF RESULT ON EVALUATION BENCHMARK

With the optimal parameters tuned on validation set, another multi-resolution detector was trained on the entire KITTI training set. Using the scaling factor of 0.5 for 25 pixels model detections and 1.1 for 50 pixels model detections, final detections were generated by concatenating detections from 25, 50, and 100 pixels models and then applying NMS. On KITTI Vision Benchmark Suite [31], our detector achieved **46.23**% AP which is about 6% more than 39.81% AP achieved by ACF.

Although methods such as Regionlets [27], FilterdICF [25], pAUCEnsT [34] (refer Fig. 11) achieve better performance, all these methods improve feature selection process and hence is orthogonal to the improvements achieved by multi-resolution approach. Multi-resolution approach can be implemented on top of these methods and further improvements in detector accuracy is expected.

On KITTI Evaluation benchmark,[4] under pedestrian detection, there are three published methods that make use of the aggregated channel features. They are compared in Table VII. Note that the addition of multi-resolution models results in the highest AP gain.

VIII. EVALUATION ON LISA DATASET

While the evaluation metrics associated with a detector trained and tested on the same dataset is a good starting point to evaluate it's performance, it is important to test a detector's performance on data collected from a different camera at a different location. This is especially the case with systems trained offline and deployed in vehicles under different conditions. Such evaluation will give us some insights into robustness of the detector.

In order to perform the above evaluation, we created a testing dataset consisting of 775 frames captured in UCSD campus.[5] The frames are extracted from 2 video sequence captured at a resolution of 1280 $\times$ 420. This data is visually different from the two commonly used datasets i.e. KITTI and Caltech. We annotate all pedestrians taller than 50 pixels, at-most partially occluded and truncated. Other people in the scene are annotated as "ignore". This translated to 848 pedestrians and 548 "ignore" boxes. 32% of the pedestrians are taller than 50 pixels but, shorter than 100 pixels whereas the rest 68% are taller than 100 pixels.

When ACF detector with single resolution approach with $M_{50\ \mathrm{s}}$ trained on KITTI was applied on our dataset, detector AP of 65.15% was achieved. Whereas with multi-resolution approach $(M_{50\ \mathrm{m}} + M_{100\ \mathrm{m}})$, detector AP increased to **68.57**%. Fig. 12 has sample detections from $M_{50\ \mathrm{m}}$ and $M_{100\ \mathrm{m}}$.

This suggests that the ACF detector trained on KITTI dataset seems to generalizes well with data collected from a different source and in a different part of the world.

IX. CONCLUDING REMARKS

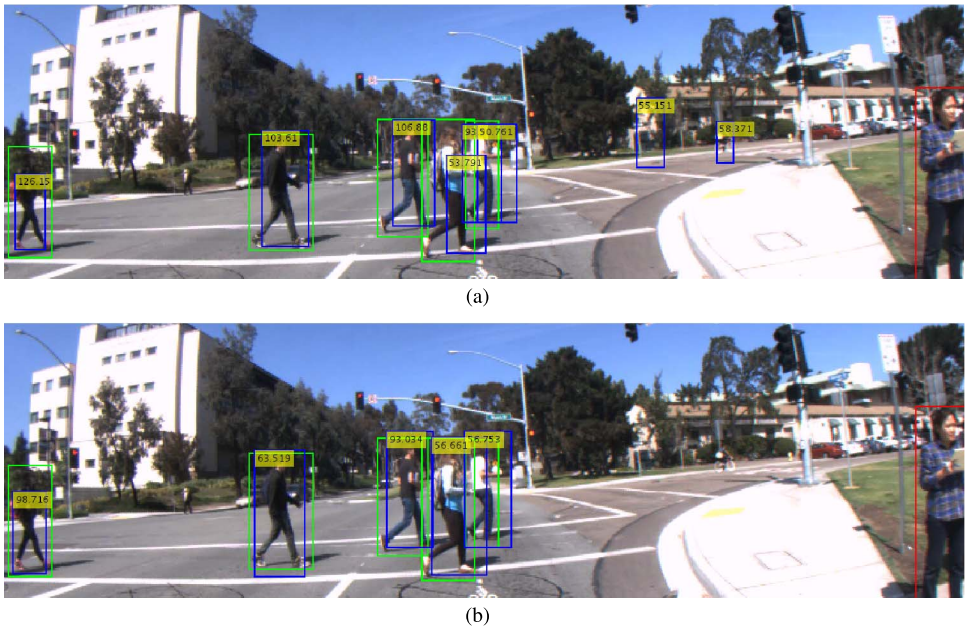
Running detector with multiple models, trained at different resolutions has significant impact on the performance of the detector. On KITTI evaluation server, under moderate difficulty settings, detector AP increased from 39.81% to **46.23**%. Applying the trained models to different dataset (LISA), we achieved similar detector performance and also similar trends with multi-resolution approach. Detailed analysis shows that aggregating channel over a larger window plays a significant role in improving the performance of multi-resolution approach.

In the context of deploying a pedestrian detection system for intelligent vehicles, the following conclusions can be drawn:

1) Pedestrians shorter than 50 pixels are harder to detect but, in the context of intelligent vehicles taller pedestrians are closer to the vehicle and hence needs more attention. This is achieved when using the multi-resolution detector approach by tuning the score calibration function as required.
2) Occluded pedestrians are hard to detect and a higher resolution model does not implicitly learn to detect them. Considerable occlusion occurs due to pedestrians traveling in groups. This is especially true in urban driving scenarios.
3) Current implementation (no feature pyramid sharing) runs under 0.6 seconds per image. ACF-MR achieves highest AP for detector running under 1 sec. This is especially critical in active safety systems for driver assistance.

[4] Available: http://www.cvlibs.net/datasets/kitti/eval_object.php

[5] Available: http://cvrr.ucsd.edu/rnattoji/LISA_PedestrianDataset.zip



(a)

(b)

Fig. 12. Sample annotations with detections using the ACF detector models $M_{50\text{ m}}$ and $M_{100\text{ m}}$, trained on KITTI and tested on the LISA dataset. Green boxes indicate the ground truth. Blue boxes are the detections. Red box shows the "ignore" region. The text inside the blue box indicates the confidence score. The only detection with a score higher than 50 is shown. Notice the variation in scores between the models, verifying the need for score calibration.

Robust and fast pedestrian detection is a critical first step in predicting pedestrian intent and driver attention for surveillance of safety critical events which in turn is essential for building a reliable safety system for self driving and highly automated vehicles.

ACKNOWLEDGMENT

We are grateful to the anonymous reviewers for their careful reading of the earlier version of the manuscript and their constructive suggestions for improving the clarity and quality of this paper. We thank our colleagues at LISA Laboratory, especially Akshay Rangesh for his help in setting up the experiments and Ravi Kumar Satzoda for his helpful comments.

REFERENCES

[1] T. Gandhi and M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[2] E. Ohn-Bar and M. M. Trivedi, "Can appearance patterns improve pedestrian detection?" in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2015, pp. 808–813.

[3] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations," in *Proc. IEEE Intell. Veh. Symp.*, 2015, pp. 808–813.

[4] A. Prioletti, A. Mogelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. B. Moeslund, "Part-based pedestrian detection and feature-based tracking for driver assistance: real-time, robust algorithms, and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1346–1359, Sep. 2013.

[5] T. Gandhi and M. M. Trivedi, "Image based estimation of pedestrian orientation for improving path prediction," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 506–511.

[6] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1292–1304, Dec. 2011.

[7] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 115–120.

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

[9] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "An exploration of why and when pedestrian detection fails," in *Proc. IEEE Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 2335–2340.

[10] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[12] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features" in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Comput. Vis. Pattern Recog.*, 2014, pp. 1–21.

[14] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features for pedestrian, face and edge detection," *CoRR*, vol. abs/1504.07339, 2015.

[15] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 241–254.

[16] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3033–3040.

[17] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intell. Transp. Syst.*, 2015, pp. 2511–2521.

[18] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2903–2910.

[19] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 380–391, Sep. 2009.

[20] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining RGB and dense LIDAR data," in *Proc. Int. Conf. Intell. Robots Syst.*, 2014, pp. 1–6.

[21] A. Gonzàlez, G. Villalonga, J. Xu, D. Vàzquez, J. Amores, and A. M. Lòpez, "Multiview random forest of local experts combining RGB and LIDAR data for pedestrian detection," in *Proc. IEEE Intell. Veh. Symp.*, 2015, pp. 356–361.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, vol. 1, pp. 886–893.

[23] J. H. H. Woonhyun Nam and P. Dollár, "Local decorrelation for improved pedestrian detection," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

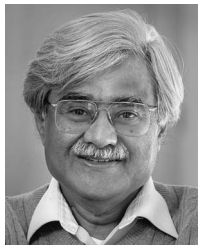
[24] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3666–3673.
[25] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1751–1760.
[26] *Imagenet Large Scale Visual Recognition Challenge 2014*. [Online]. Available: http://image-net.org/challenges/LSVRC/2014/results
[27] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1–8.
[28] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets relocalization," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 260–275.
[29] C. Cadena, A. Dick, and I. Reid, "A fast, modular scene understanding system using context-aware object detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 4859–4866.
[30] P. Xu, F. Davoine, and T. Denœux, "Evidential combination of pedestrian detectors," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–14.
[31] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3354–3361.
[32] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[33] P. Sudowe and B. Leibe, "Efficient use of geometric constraints for sliding-window object detection in video," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2011, pp. 11–20.
[34] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015. [Online]. Available: http://arxiv.org/abs/1409.5209

Rakesh Nattoji Rajaram received the bachelor's degree in electrical engineering from International Institute of Information Technology, Hyderabad, India, in 2014. He is currently working toward the master's degree in intelligent systems, robotics and control at University of California San Diego, La Jolla, CA, USA.

His research interests include computer vision, machine learning, intelligent vehicles, and autonomous robots.

Eshed Ohn-Bar received the M.S. degree in electrical engineering from University of California San Diego, La Jolla, CA, USA, in 2013, where he is currently working toward the Ph.D. degree in signal and image processing.

His research interests include computer vision, object detection, multimodal activity recognition, intelligent vehicles, and driver assistance and safety systems.



Mohan Manubhai Trivedi received the B.E. degree (with honors) from Birla Institute of Technology and Science, Pilani, India, in 1974 and the Ph.D. degree from Utah State University, Logan, UT, USA, in 1979.

He is a Distinguished Professor of electrical and computer engineering with University of California San Diego (UCSD), La Jolla, CA, USA, and the Founding Director of the Computer Vision and Robotics Research Laboratory, UCSD, and Laboratory for Intelligent and Safe Automobiles (LISA), UCSD, which is the winner of the IEEE Intelligent Transportation Systems (ITS) "Lead Institution" award in 2015. Currently, LISA team members are pursuing research in intelligent/highly automated vehicles, machine perception, machine learning, human–robot interactivity, driver assistance, active safety, and ITS. The LISA team has played a key role in several major research collaborative initiatives. These include human-centered vehicle collision avoidance systems, vision-based passenger protection system for "smart" airbags, predictive driver intent analysis, and distributed video arrays for transportation and homeland security applications.

Prof. Trivedi has given over 100 Keynote/Plenary talks. He was a recipient of the IEEE ITS Society's "Outstanding Research Award" and a number of other major awards. He is a Fellow of the International Association of Pattern Recognition and The International Society for Optical Engineers. He serves regularly as a Consultant to industry and government agencies in the USA and abroad.