IEEE Intelligent Vehicles Symposium, June 2015 (to appear)

Can appearance patterns improve pedestrian detection?

Eshed Ohn-Bar and Mohan M. Trivedi[1]

Abstract— **This paper studies the usefulness of appearance patterns for the challenging task of pedestrian detection. Despite appearance specific models being common in rigid object detection, the technique is still little understood for pedestrians. Three main approaches for reasoning over orientation, occlusion, and visual cues in obtaining the appearance patterns are compared. This work demonstrates that large gains in detection performance (up to 17 AP points on the challenging KITTI dataset) can be made using a state-of-the-art pedestrian detector.**

I. Introduction

In recent years, pedestrian detection has attracted tremendous interest in the research community. Although challenging, accurate detection of people has many potential applications both in the intelligent vehicles domain [1], [2] and other human-observing application domains. A motivating force propelling the field forward has been improving the quality of extracted image features [3]–[5]. For instance, a combination of gradient, color, local binary patterns, region covariance, and spatial pooling produced state-of-the-art results [6] on the Caltech pedestrian detection benchmark [7]. KITTI pedestrians results [8] have also shown a similar trend [9]. Local de-correlation of gradient and color features was proposed in [10] with considerable improvements due to better generalization of the model.

This paper studies an alternative method for improving detection results to the aforementioned. The general idea is to produce specific appearance patterns from the data which allows for training models for varying aspect-ratio, orientation, occlusion, or other visually challenging settings. The method builds upon the fast pedestrian detection framework of aggregate channel features (ACF) [11], [12]. In ACF, a set of 10 channel features are computed efficiently and classified in a sliding window manner. Approximation of features at some of the scales of the feature pyramid allows for further speedups. We demonstrate that learning multiple models for different types of pedestrians, produces significant improvement in detection performance when compared to training one detector (referred to as the 'monolithic' case) over the entire pedestrian dataset.

The main cost is in speed reduction due to evaluation of multiple models over the feature pyramid. Nonetheless, incorporation of richer features also generally comes with a large speed reduction. For instance, the method in [13], which studies different haar-like patterns on the gradient+LUV feature channels takes about 1.6 seconds on 640×480 images. On the other hands, ACF reaches over 30 frames per second (fps) on the same image size. Evaluation of multiple models only gracefully reduces run-time speed, while also providing improved detection performance. For instance, the baseline ACF model runs at a little under 9 fps for 1242×375 with similar settings to [11] on a desktop CPU machine. Adding 8 component models reduces run-time to 7 fps with the same environment.

In this work we study performance improvement due to increasing modeling capacity. This is done by partitioning the data into smaller, better-handled clusters. The resulting specialized models improve the overall modeling capacity and provide semantic information [14]. Improving modeling capacity have been then in literature in different ways. For instance, the notion of parts, in particular head, upper, and lower body, contain useful cues for pedestrian detection [15], [16]. The deformable parts model (DPM) is another well known example [17], where parts are latent and learned by employing a latent SVM. We note that generally DPM and ACF has been compared side-by-side, yet ACF employed a single rigid model and DPM employs a multiple components model. Hence a careful examination of a multi-component ACF is a natural extension.

Rigid object detection has shown significant gains in performance by learning aspect-ratio, orientation, and/or occlusion specific models [18]–[22]. Nonetheless, the extension to pedestrians is not straightforward. This work studies three methodologies of extracting appearance patterns for training the detectors. Since each detector is now specialized towards a specific appearance pattern, evaluation in test time is fast with non-pedestrian windows being rejected early in the cascade. The method is suited for parallelized systems as well as each appearance cluster component model is is evaluated independently.

For the experiments, the challenging KITTI dataset [8] is employed which contains 7481 images and over 3000 pedestrians. The dataset is split as is into two, a training set and a validation set of the same number of images. All types of pedestrians at all occlusion levels and sizes (greater than 25 pixels in height) are used, which is quite challenging ('hard' settings as defined in [8]). In training, images are flipped to double the number of pedestrians available.

II. Baseline Detector

AdaBoost [11] is learned using depth-2 decision trees as weak classifiers. Detection at multiple scales is handled using approximation of features at nearby scales with a power law [23]. Given an image, a set of image channels are computed. Six gradient orientation channels and three LUV (color space

[1] Laboratory for Intelligent and Safe Automobiles, University of California San Diego, La Jolla CA 92092, USA {eohnbar, mtrivedi}@ucsd.edu

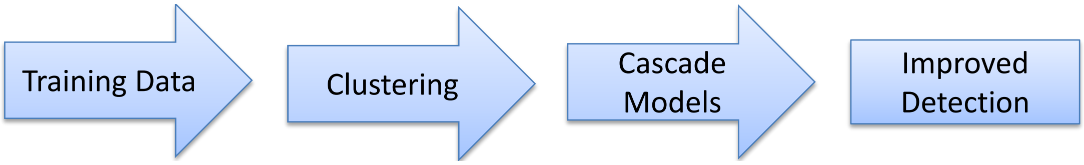


Fig. 1: The proposed approach learns multiple specialized AdaBoost models from the pedestrian dataset as opposed to a train on all approach.

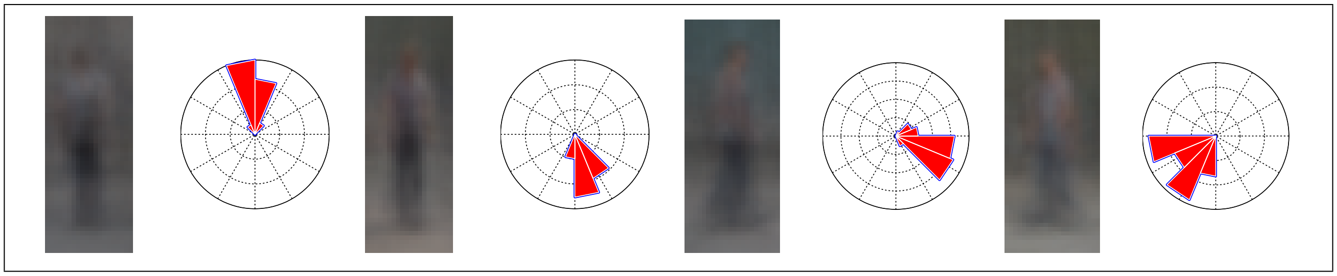
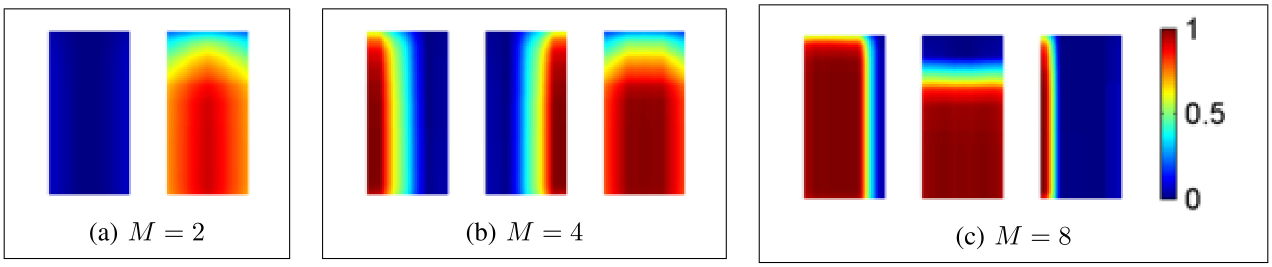


Fig. 2: Example orientation clusters obtained for pedestrian detection on KITTI. The images in each clusters are averaged to produce the visualized mean image, and a rose plot shows a histogram count of orientation values within the cluster. Best viewed in color on a computer screen.



(a) $M = 2$

(b) $M = 4$

(c) $M = 8$

Fig. 3: For each occlusion type cluster, an average occlusion map is visualized, where M is the number of occlusion clusters (for $M = 8$ only selected clusters are visualized). Note that although only shown for $M = 2$, the non-occluded instances cluster is always present.

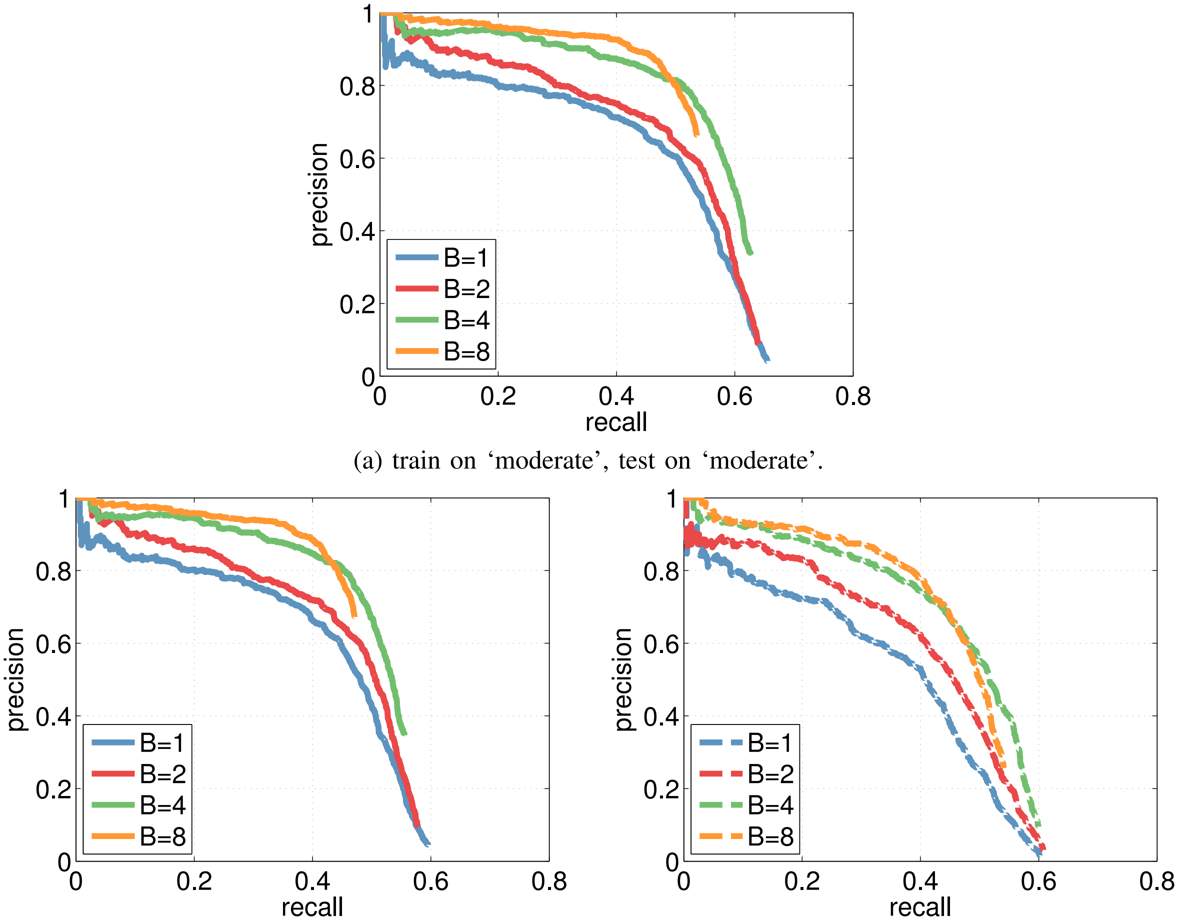
transformation) occur. The images are pre-smoothed with a Gaussian filer. Next, the color and gradient image features are averaged in 4×4 blocks in order to produce fast pixel lookup features (as opposed to the Viola-Jones haar-like type features). There are four stages of training. In the first stage, 5000 random negatives are collected and a detector is learned using 32 weak classifiers. Next, three more rounds of hard negative mining follow, where the number of weak classifiers are quadrupled in each round up to 2048. Generally, the last round results in very few hard negatives mined which shows convergence. This provides a fast training and testing pipeline, although increasing the number of weak classifiers, number of allowed negatives, and tree depth can result in improved detection performance [6].

III. How to Obtain the Appearance Patterns?

The training set can be clustered into groups. Generally, the method in [18], [24] is followed, but it is adapted to the pedestrian domain. Below are the details of the three types of features used for for clustering in order to obtain the appearance models. Throughout the experiments, the model's height is kept fixed at 62 pixels, and the width is obtained from the median in-cluster aspect ratio. Although models at additional resolutions (as in [18]) were experimentally shown to significantly improve detection performance (partly due to better handling of small pedestrians), for simplicity sake this work concentrates on a best performing single resolution model.

Orientation (B): As shown in Fig. 2, one possible clustering can be induced by orientation. The parameter for orientation bins in the experiments is referred to as B. KITTI provides 3D bounding boxes for pedestrians, with a known orientation in 3D. Samples are binned according to their 3D orientation, and the template aspect ratio is determined using the mean aspect ratios in the cluster. Although orientation is important for object such as cars, the same needs to be shown for pedestrians.

Occlusion (M): As mentioned in [7], generally most pedestrian instances fall into a small set of possible occlusion configurations. Although KITTI does provide a coarse occlusion metric (little, partial, and heavy occlusion), we hypothesize that this is not sufficient for good disambiguation of occlusion types. For instance, as heavy occlusion can occur in several ways over the pedestrian (left, right, bottom, etc.),



(a) train on ‘moderate’, test on ‘moderate’.

(b) train on ‘moderate’, test on ‘hard’.

(c) train on ‘hard’, test on ‘hard’.

Fig. 4: Impact of adding orientation components (B) over the baseline model which trains a single model over the entire dataset ($B = 1$). In each figure, we vary the difficulty in training and testing. Note how incorporation of ‘hard’ samples is results in lower precision at small recall, but possibly longer curves. Note how learning per-orientation models significantly improves detection performance.

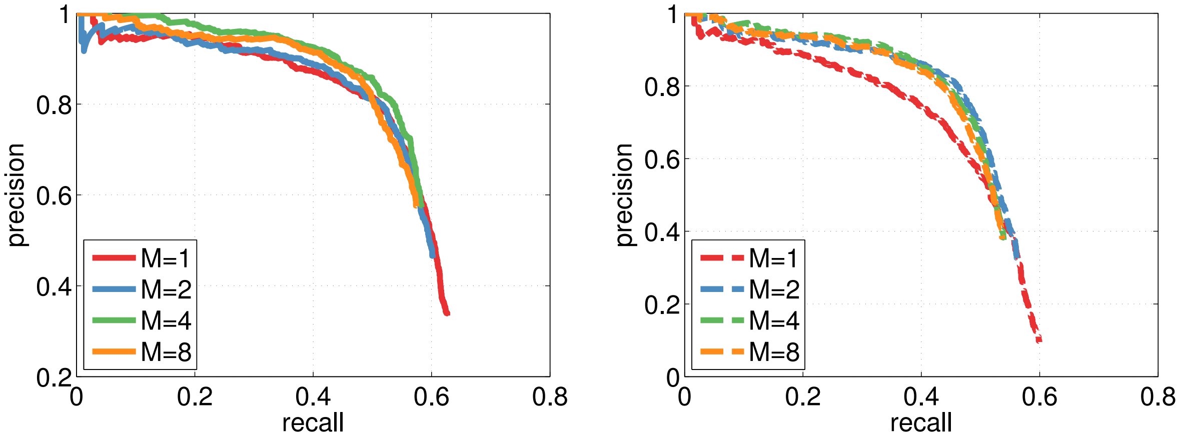
a rigid model could benefit from further granularity. First, all the samples are re-sized to a fixed size and processed to obtain an occlusion map. The process is automatic, as follows. Using the LIDAR information, for each pedestrian, we can find all the annotated objects that overlap it while being closer to the camera. The intersection of these occluders and the occludee box provide an occlusion value for the mask. Next, the occlusion masks are clustered using k-means and are separated by orientation. The clustering process aims to quantize together similarly occluded pedestrians. Hence, each binary mask can be measured in similarity to another by simply checking at each pixel whether both are 1 (occluded) or 0. Therefore, it is natural to use the average number of pixels that agree over the two occlusion masks as the similarity metric in clustering. This simple algorithm produces well aligned clusters, as shown in Fig. 3.

Visual: We also experiment with clustering based on visual features directly, which may be informative to researchers working on datasets with no 3D orientation available. Furthermore, in principle rich visual features can capture many more data-driven appearance variations (e.g. not just due to orientation or occlusion), thereby providing an interesting comparison. For features, we employ R-CNN [25] which is fine tuned on the PASCAL VOC dataset. Each sample in the entire dataset is re-sized to the expected dimension for the network in [25]. The dimensionality of the final feature set for each sample is 4096, and k-means provides the final clustering assignment.

An AdaBoost model is learned for each cluster, producing a set of detection models. In test time, each window is scored independently using the detection models. Detections are merged using a greedy non-maximum suppression (NMS) procedure; once a bounding box is suppressed by an overlap criterion, it can no longer suppress weaker detections.

IV. Experimental Evaluation

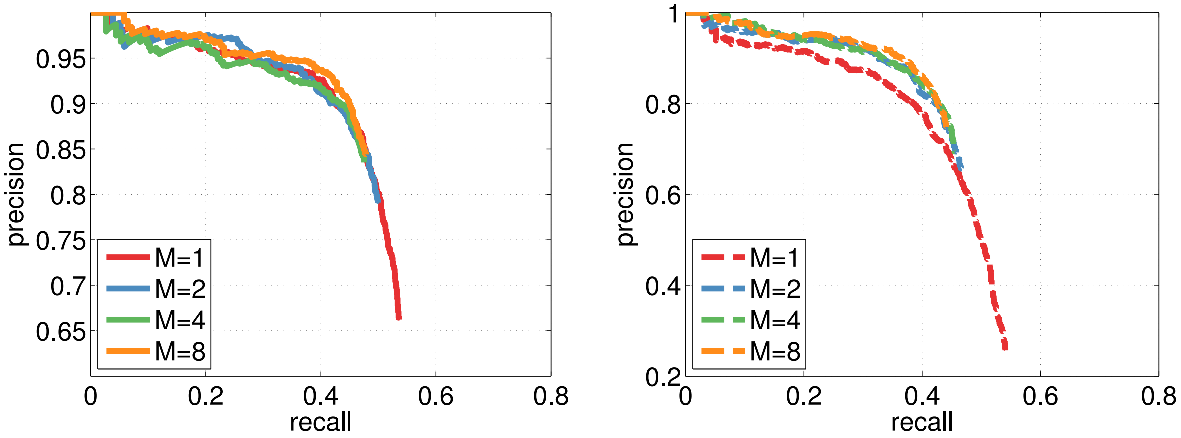
In order to gain further insight into the role of the appearance patterns in obtaining better modeling capacity, the approach is evaluated in two ways. First, only ‘moderate’ difficulty samples are used (25 pixels in height and above,



(a) train on ‘moderate’, test on ‘moderate’.

(b) train on ‘hard’, test on ‘hard’.

Fig. 5: For a fixed $B = 4$ orientation bins, what is the impact of adding occlusion clusters? ‘hard’ settings contain a large number of heavily occluded samples.



(a) train on ‘moderate’, test on ‘moderate’.

(b) train on ‘hard’, test on ‘hard’.

Fig. 6: For a fixed $B = 8$ orientation bins, what is the impact of adding occlusion clusters?

partial occlusion and truncation up to 30%). Second, training is done on ‘hard’ difficulty which allows for heavy occlusion and up to 50% truncation.

As shown in Fig. 4, incorporation of orientation clusters significantly impacts performance in both of the training methodologies. Increasing from 1 bin (the baseline which uses the entire training set) to 4 and 8 bins results in a noticeable AP improvement. Over 4-8 bins produced little to no gain in performance. Furthermore, it is observed how training on ‘hard’ samples actually hinders detection performance at low recall [26], [27]. Occluded samples provide noisy and difficult cases which are not well resolved by existing state-of-the-art on KITTI.

Figs. 5 and 6 show the improvement due to incorporation of occlusion clusters in addition to the orientation clusters of $B = 4$ and $B = 8$, respectively. Here, the occlusion scheme is shown to better handle ‘hard’ samples in training, translating to improved performance when such samples are present. Finally, all of the methods are directly compared in Fig. 7 on ‘hard’ settings. All are shown to greatly improve over the monolithic, $B = 1$ model. Some, such as a four orientation and two occlusion clusters model (total of 8 clusters, $B = 4, M = 2$) result in more graceful decline as detection score threshold decreases and recall increases, but lower precision at low recall rates. On the contrary, $B = 8$ (moderate) which was trained on ‘moderate’ difficulty is shown to produce the more precise at low recall curves shown before on ‘medium’ settings. Interestingly, the clusters do contain some complementary information, as shown by combining an 8-cluster model over visual CNN features (CNN-8) and an orientation+occlusion model ($B = 4, M = 2$). Under these settings, a monolithic classifier produces an 33.63 AP value, while the 16 cluster combined model produces 50.92 AP on the ‘hard’ settings. Another interesting fact is that unlike in car detection [18], the performance is less sensitive to the clustering method used, and both CNN-based and geometry-based features work well.

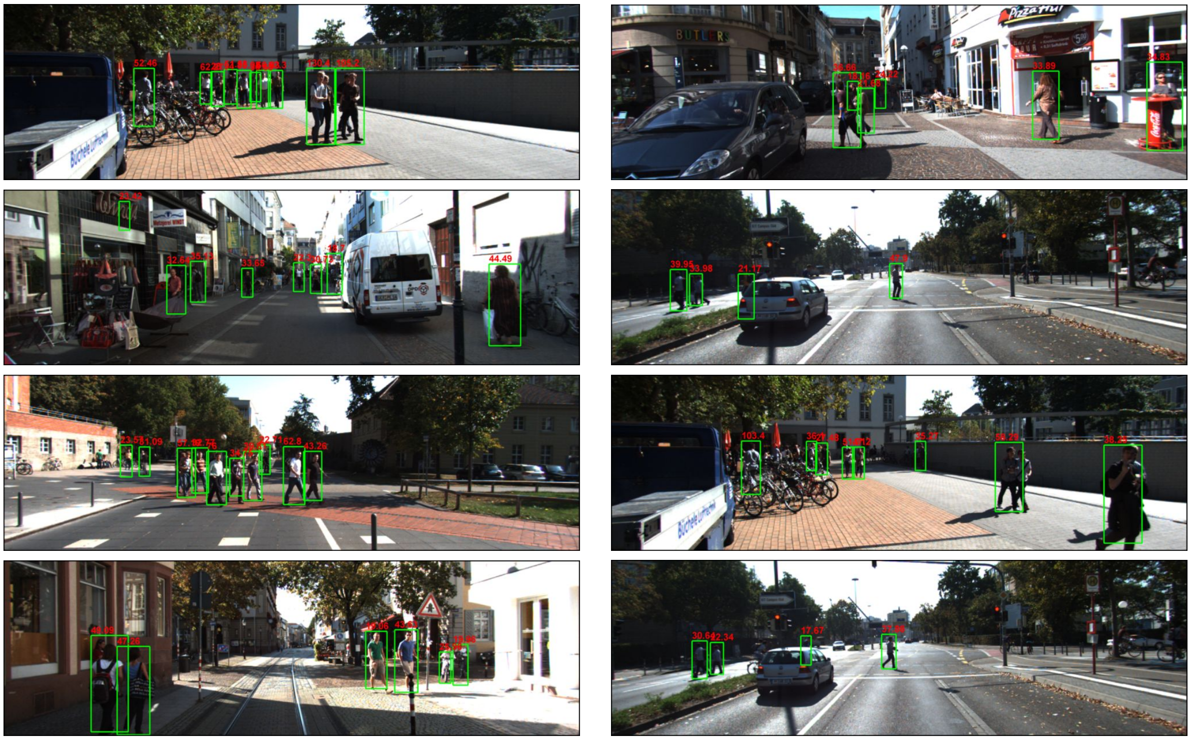


Fig. 8: Detection results using the proposed approach on KITTI. Best viewed on screen.

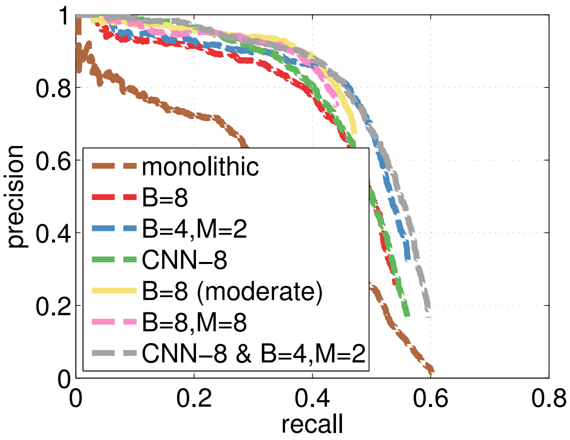


Fig. 7: A comparison of all the different clustering techniques while testing on 'hard' settings. All are shown to improve over the monolithic baseline, and visual clustering is shown to produce complementary appearance patterns to the orientation and occlusion clustering.

V. Conclusion

This paper studied the effectiveness of multi-component AdaBoost models on the task of pedestrian detection. The approach showed promise over the most challenging settings of the KITTI dataset. The analysis demonstrates that the detection task itself benefits from orientation models. In the future, the impact of the the appearance patterns on orientation estimation and tracking will be studied [28], [29]. Integrative approaches may also provide improved detection performance [30]. We would also like to study the benefit for effective pedestrian-based active safety systems [31].

References

[1] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation," in *IV*, 2014.

[2] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Vision on wheels: Looking at driver, vehicle, and surround for on-road maneuver analysis," in *Computer Vision and Pattern Recognition Workshops-Mobile Vision*, 2014.

[3] R. Benenson, M. Omran, J. Hosang, , and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV, CVRSUAD workshop*, 2014.

[4] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *CVPR*, 2015.

[5] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, 2009.

[6] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," in *arXiv*, 2014.

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, 2012.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets relocalization," in *ACCV*, 2014.

[10] W. Nam, P. Dollár, and J. Han, "Local decorrelation for improved pedestrian detection," in *NIPS*, 2014.

[11] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *PAMI*, 2014.

[12] A. Mogelmose, D. Liu, and M. M. Trivedi, "Traffic sign detection for u.s. roads: Remaining challenges and a case for tracking," in *ITSC*, 2014.

[13] S. Zhang, C. Bauckhage, and A. B. Cremers, "Efficient pedestrian detection via rectangular features based on a statistical shape model," *TITS*, 2014.

[14] M. Enzweiler and D. Gavrila, “Integrated pedestrian classification and orientation estimation,” in *CVPR*, 2010.

[15] A. Mgelmose, A. Prioletti, M. M. Trivedi, A. Broggi, and T. B. Moeslund, “Two-stage part-based pedestrian detection,” in *ITSC*, 2012.

[16] A. Prioletti, A. Mgelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. Moeslund, “Part-based pedestrian detection and feature-based tracking for driver assistance: Real-time, robust algorithms and evaluation,” *TITS*, vol. 14, no. 3, pp. 1346–1359, Sep. 2013.

[17] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[18] E. Ohn-Bar and M. M. Trivedi, “Learning to detect vehicles by clustering appearance patterns,” *TITS*, 2015.

[19] B. Pepik, M. Stark, P. Gehler, and B. Schiele, “Occlusion patterns for object class detection,” in *CVPR*, 2013.

[20] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Data-driven 3D voxel patterns for object category recognition,” in *CVPR*, 2015.

[21] Y. Xiang and S. Savarese, “Object detection by 3D aspectlets and occlusion reasoning,” in *3dRR*, 2013.

[22] R. Mottaghi, Y. Xiang, and S. Savarese, “A coarse-to-fine model for 3D pose estimation and sub-category recognition,” in *CVPR*, 2015.

[23] P. Dollár, S. Belongie, and P. Perona, “The fastest pedestrian detector in the west,” in *BMVC*, 2010.

[24] E. Ohn-Bar and M. M. Trivedi, “Fast and robust object detection using visual subcategories,” in *CVPRW-Mobile Vision*, 2014.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *CVPR*, 2014.

[26] J. Yebes, L. Bergasa, and M. Garca-Garrido, “Visual object recognition with 3D-aware features in kitti urban scenes,” *Sensors*, 2015.

[27] J. J. Yebes, L. M. Bergasa, R. Arroyo, and A. Lázaro, “Supervised learning and evaluation of KITTIs cars detector with dpm,” *IV*, 2014.

[28] T. Gandhi and M. M. Trived, “Image based estimation of pedestrian orientation for improving path prediction,” in *IV*, 2008.

[29] A. Mogelmose, M. Trivedi, and T. Moeslund, “Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations,” in *IV*, 2015.

[30] S. Sivaraman and M. M. Trivedi, “Integrated lane and vehicle detection, localization, and tracking: A synergistic approach,” *TITS*, vol. 14, no. 2, pp. 906–917, 2013.

[31] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, “On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems,” *CVIU*, vol. 134, no. 0, pp. 130–140, 2015.