On Diversity and Complementarity of Pedestrian Detection Models

T Toprak¹, B Belenlioglu², S Dogan², B Aydın², M A Selver³

¹Institute of Natural and Applied Sciences, Dokuz Eylul University, Kaynaklar Campus, 35160, Buca, Izmir, TURKEY

²Kentkart, Kahramanlar Mahallesi Mürselpaşa Bulvarı – İzmir, TURKEY

³Department of Electrical and Electronics Engineering, Dokuz Eylul University, Kaynaklar Campus, 35390, Buca, Izmir, TURKEY

Email tugce.toprak@ogr.deu.edu.tr, burak.belenlioglu@kentkart.com.tr, sinan.dogan@kentkart.com.tr, burak.aydin@kentkart.com.tr, alper.selver@deu.edu.tr

Abstract. Pedestrian Detection (PD) is now almost a mature field of signal processing in which numerous application areas exists. Emerging one of these applications is driving autonomy that requires advanced driver assistance systems (ADAS) for PD. Although various benchmark data sets are collected and several methods are already developed, the reported results show that current systems lack generalization capacity due to strong dependence of system performance on PD training set. In other words, the systems trained with a pedestrian data set usually perform considerably well on a test set taken from the same benchmark, but their performance drop drastically when tested on a different benchmark data set. In order to overcome this limitation, one of the approaches might be fusing outcomes of different systems. Here, “different systems” can refer either to train the same model with different data sets or using entirely different models. In both cases, the outcomes of different systems should be diverse enough to cover the entire solution space. Up to our knowledge, the performance of existing well-established models for PD is frequently compared to each other but their diversity is not analysed in detail yet. In this study, the feasibility of PD systems, whose effectiveness on PD is shown by earlier studies, was examined by systematically training and testing them with different benchmark data sets and their combinations. For this purpose, pairwise and non-pairwise measures of diversity have been employed and several conclusions are driven about the complementarity of existing PD models.

1. Introduction

Intelligent Transportation Systems (ITS) are designed to provide all kinds of communication and information exchange in vehicles, between vehicles, and between the vehicle and control center for providing safety, efficiency, and quality. Traffic safety and prevention of possible accidents have a crucial place among the purposes of ITS [1]. Within the context of this, camera-based pedestrian detection (PD) constitutes an important role. When the related literature is analyzed, it can be observed that a tremendous effort is already given to create adequate data sets and to develop advanced systems [2]. However, recent surveys and comparisons to human PD performances suggest that still a significant improvement in performance is required in order to reach flawless "advanced driver assistance systems" (ADAS) for ITS [3].
Accordingly, many studies are still being conducted to determine and quantify the most impactful strategies on PD quality. The main purpose of using all the datasets prepared for PD and developed models (i.e. feature-classifier systems, deep networks or their combinations) is to create a system that can represent the pedestrians in the most unique and effective way. Comparisons between advanced and/or state-of-the-art systems are continuously being performed by evaluating the results of application performances on the same benchmark data sets [4]. Although the results are discussed according to the strong and weak aspects of the systems, their complementarity properties have not been analyzed quantitatively in detail yet [5]. Existing survey papers mainly focus on two directions:

1) The parametric (adaptive learning rate), architectural (number and type of layers), experimental (i.e. pre-training with larger data sets such as ImageNet and fine tuning to pedestrian data sets via transfer learning) and design (i.e. multi-tasking, semantic labeling, multi-scale training etc.) strategies for deep architectures such as convolutional neural networks (CNN) and its derivatives including Regional-CNN (RCNN), Faster RCNN etc.

2) Extracting more informative features for Feature-Classifier Systems (FCS) and combining multiple feature sets to achieve higher success rates. In the light of extensive number of simulations and experiments, so far, no conclusive evidence could be found to conclude that a particular type of classifier is better suited for PD among others (e.g. linear/non-linear SVM, Adaboost cascades or decision forests).

Moreover, INRIA has images with high resolution.

This paper mainly focuses on the second issue since recent studies suggest that more informative features are still necessary to resolve difficulties in PD. This conclusion points the importance of FCS, in which features are extracted explicitly (i.e. hand-crafted) and fed to the classifiers after adequate preprocessing (i.e. dimensionality reduction, feature selection, normalization etc.). Despite implicit representation of features automatically arranged by many layers of deep architectures, FCS allows full control over feature space and information extraction.

Moreover, CNNs internally include feature pooling operations that make them vulnerable to alignment and localization challenges. Therefore, most of the deep systems prune their results using hand-crafted features (such as Integrate Channel Features – ICF) or use external FCS are used to generate detection proposals. Such two-stage strategies are reported to save computation by reducing the number of bounding box (BB) candidates for CNNs to process with a risk for making more mistakes.

Similar to deep architecture, it has been observed and shown that different FCS provide variable performance under different operating conditions, which are demonstrated with examples in this proposal. Thus in this study, the diversity and complementarity of FCS is analyzed in detail.

2. Pedestrian Data Sets and Their Characteristic Properties

In recent years, several datasets have been developed for training and evaluation parts of pedestrian detection such as ETH [6], Daimler [7], TudBrussels [8], KITTI [9], Caltech-USA [10] and INRIA [11]. Caltech-USA and INRIA are the two complementary datasets. Caltech-USA was recorded by driving through large cities and provide annotated frames on video sequences. However, INRIA includes images from personal digital image collections taken over a long time period and a few images from web. The analysis show that INRIA is more challenging in terms of aspect ratio and posing while Caltech has more diversity for representing the real time traffic situations.

The total number of frames in Caltech-USA, which are used in this paper, is nearly 114K and the total number of annotated bounding boxes is nearly 245K. On the other hand, in INRIA, these numbers are pale by comparison with Caltech-USA. The number of frames is nearly 1K and the number of annotated bounding boxes is nearly 2K. Although the number of frames is high in Caltech-USA, they include only ~1300 unique pedestrians. Although the number of frames is relatively small in INRIA, almost all of the frames include unique pedestrians. Moreover, INRIA has images with high resolution.
Despite the large number of frames in Caltech-USA, the dataset suffers from low-density. Another weakness of Caltech-USA is that dataset was recorded in a single city. Hence, the diversity in pedestrian and background appearances is restricted. Conversely, the INRIA dataset includes many several appearance of pedestrians. However, almost whole annotated bounding boxes in both datasets are upright pedestrians. The average aspect ratio is representing variations of appearance of pedestrians in dataset and can be determined with dividing width by height of bounding box of pedestrian. The ratios of the datasets are 0.41 and 0.33 respectively.

The main idea of collecting a mixed dataset from them is to achieve a dataset which is more diverse than individual ones. The created mixed dataset has more variety in terms of appearances and also has higher aspect ratio than individual ones as shown in Figure 1. The complementarity analyzes of feature-classifier models are observed with this mixed dataset.

Figure 1. Aspect ratio distribution of mixed dataset.

3. Employed Feature-Classifier Models

Aggregate Channel Features (ACF) + AdaBoost Cascades:
The structure of the ACF system is based mainly on the extension of channels and has been applied since digital images began to be used. ACF combines gradient histograms and gradient magnitudes with color features and corresponds to an important benchmark in pedestrian detection. The most common color channels are red-green-blue (RGB), hue-saturation-value (HSV) and luminance-chroma-hue (LUV). In addition, different channels can be created using linear or non-linear transformations of the view. The gradient magnitude and gradient histogram extraction from these channels are also generalized as HOG features. Although, many variants are proposed mainly by utilizing different filter types (i.e. square averaging, checkerboards, eigen-vectors from linear discriminant analysis and rotated filters), ACF constitutes a baseline to all. Its computational efficiency is increased by approximations or multiple model use across different scales and neighboring windows.

After extension of channels by appropriate methods, they are transferred to the sub-sample space depending on a predetermined coefficient. All the pixels of the ACF channels that are transmitted to the sub-sample space are transformed into a vector on a look-up table. In this way, the feature vector to be trained with the AdaBoost cascades is obtained.

AdaBoost is one of the popular boosting technique that helps to combine multiple weak classifiers into a strong classifier. Weak classifier performs poorly, but also performs better than random guessing. Usage of cascade version of these classifiers is very practical method for classification problems, but weight of each classifier is important to determine for getting efficient combination. In AdaBoost cascades, AdaBoost determines how much weight should be given to each classifier’s proposed answer when combining their result.

In this paper, ACF + AdaBoost cascades model is implemented with use of “detectPeopleACF” function in Matlab. There is two different kind of trained detector as ACF model. One of them is trained with INRIA pedestrian dataset and the classification model of the function have to be set as “inria-100x41”. The other one is trained with Caltech pedestrian dataset and classification model have to be set as “caltech-50x21”. The function have many options which are changable according to applications,
but the most important option is “SelectStrongest”. SelectStrongest option can be selected as false for cancelling selection of strongest box from a group, so more detected bounding boxes can be located in image. The results of detections with ACF, the difference between “SelectStrongest” is true or false and also comparison with other detectors will be discussed in last chapter.

**Histogram of Oriented Gradients (HOG) + Support Vector Machines (SVM):**

Feature extraction using the HOG method is one of the well-known methods that is often used to express the characteristics of an image and training features with SVM in pedestrian detection systems achieve high performance [12].

According to this method, a mask, such as (-1, 0, 1), is applied to image in order to find gradients. Subsequently, the image is divided into regions called cell which consist of independent pixels. One dimensional gradient histograms are calculated for each cell region. For each pixel in the cell, the gradient is assigned to the appropriate one of 9 different directions (0-40, ..., 321-360) and each pixel votes in a direction that is proportional to the magnitude of the gradient value. The groups of cells formed in this way are combined to generate blocks. Thus, the histogram of each cell region is normalized by looking at the gradient energies of the other cell regions in the block. The gradient histogram vectors from each block in the image are accumulated to obtain the final feature vector to be used in the classification.

Support vector machines perform classification by determination of hyperplane that maximizes the margin between two classes. In our case, support vector machine algorithm learns person/non-person classification according to HOG feature vector. Linear SVM is used for the classification procedure, because of there is only two different class and it provides computational efficiency.

In this paper, HOG + SVM model is implemented with use of “vision.PeopleDetector” function in Matlab. The classification model of the function which is more suitable for problem, is “UprightPeople_96x48”. The function have many options which are changable according to applications, but the important options are “MergeDetections” and “ClassificationThreshold”. MergeDetections option is used for cancelling the merged similar detections, so more detected bounding boxes can be located in image. In case, ClassificationThreshold is the threshold value for decision of person or not. If the threshold is smaller, the more box is labeled as person. The results of detections with HOG, the difference between “MergeDetections” is true or false and also comparison with other detectors will be discussed in last chapter.

**Deformable Part Models + Latent SVM:**

Deformable Part Models (DPM) have recently emerged as a useful and popular tool for conflict with the diversity problem on object detection systems. When objects detect with whole feature map of training model, some kind of position changing situations might be a problem for algorithm. For our problem, standing person and running person has different aspects and according to test model, one of them might be missed.

Basically, deformable part models include histogram of gradient features and the SVM algorithm is utilized for generation of the model. However, differently from HOG + SVM model detection system, DPM has model parts for object to be detected. The parts of model are determined by voting the magnitude and orientation of gradients.

The test part of detection algorithm has two feature maps. Feature map with low resolution is to convolve with whole body model and the other one for parts of model. After convolving the whole body and parts, responses are obtained. The combined score of root locations are determined with adding all responses each other.

In this thesis, DPM + LSVM model is implemented with use of a function that can be run in Matlab. The DPM function is released by [14] gitHub and that includes open source code. The function is trained with four different data sets which are INRIA, Pascal VOC 2006, Pascal VOC 2007 and Pascal VOC 2008. The classification model of the function is chosen as “INRIA” in this paper. The results of detections with DPM and comparison with other detectors will be discussed in last chapter.
4. Application & Results

Both complementarity and inter-usability of the models mentioned in Section 3 are important factors for covering the whole solution space without false negatives. Moreover, it provides an insight about the generalization capability of the systems. The previous studies indicate that the detectors, which complete their training with a dataset that is different than the test set, cannot perform enough accuracy. This lack of generalization is observed to be related with diversity of the pedestrians in a dataset rather than number of samples.

In this paper, we assume that detection algorithms complement each other and number of false negatives are decreased with respect to their individual versions. For proving this hypothesis, first of all, detection algorithms have to be examined separately. Table 1 shows detection results for each model. As presented below, the LAMR values, which correspond to failure ratio, of all models are very high and that means all of them are inadequate for PD.

| Detector | Miss | Hit | True Pos. | False Pos. | LAMR |
|----------|------|-----|-----------|------------|------|
| DPM      | 236945 | 10207 | 10081 | 7144 | 0.964 |
| ACF(w/Caltech) SelectStrongest True | 168066 | 79086 | 79567 | 67956 | 0.822 |
| ACF(w/Caltech) SelectStrongest False | 161906 | 85246 | 903733 | 646588 | 0.957 |
| ACF(w/INRIA) SelectStrongest True | 229549 | 17603 | 17054 | 41029 | 0.948 |
| ACF(w/INRIA) SelectStrongest False | 227112 | 20040 | 442478 | 323367 | 0.988 |
| HOG(Upright 96x48) MergeDetections True Class. Thresh. = 1 | 227549 | 19603 | 19538 | 398389 | 0.988 |
| HOG(Upright 96x48) MergeDetections False Class. Thresh. = 1 | 220234 | 26918 | 213432 | 1366450 | 0.977 |
| HOG(Upright 96x48) MergeDetections False Class. Thresh. = 0 | 210875 | 36277 | 637498 | 16247008 | 0.977 |

At this point, an important experiment was performed for ACF+AdaBoost systems to see the effect of the dataset on which the model is trained. Both datasets are divided by the performance of the algorithm instead of the increasing difficulty level. For example, the score values, in the application result of the Caltech-trained model, are sorted by values and first quarter with high score is formed as first set. Four subsets of datasets were reorganized in this manner. Then, both the Caltech-trained and INRIA-trained model were applied to these groups (Figure 2.a-b). Similarly, they were applied to the sets that were reorganized according to the performance of the INRIA-trained model (Figure 2.c-d). Pursuant to obtained results, a model has inability to achieve same performance on the set where the performance of another model is high, and even this performance remains at a very low level. Different methods, different models and even similar models trained with different datasets provides good performance for different parts of the same dataset. This ensures high diversity results and shows that the results, that will occur when different systems are used as an ensemble, are potentially complementary.
Measurement of the diversity of the individual methods can be analyzed in two subsection which are pairwise and non-pairwise diversity measurements. Both of them are determined with same technique for better comparison. The measurement of interrater agreement $\kappa$ is chosen as technique for diversity. Non-pairwise version of $\kappa$ is determined with equations 1-3.

$$\kappa = 1 - \frac{1}{L} \sum_{j=1}^{N} l(z_j) (L - l(z_j)) \frac{N(L - 1) \bar{\rho} (1 - \bar{\rho})}{N(L - 1) \bar{\rho} (1 - \bar{\rho})}$$

$$l(z_j) = \sum_{i=1}^{L} y_{j,i}$$

$$\bar{\rho} = \frac{1}{NL} \sum_{j=1}^{N} l(z_j)$$

The $y_{j,i}$ term means that ith detector detects jth ground truth or not. If detection result is labelled as true positive $y_{j,i}$ becomes 1 otherwise $y_{j,i}$ becomes 0. This y values are summed for all detectors and $l(z_j)$ term is observed as equation 2. The $\bar{\rho}$ term means the average of individual classification accuracy for all detectors and ground truths as shown in equation 3. N is equal to number of total ground truths and L is equal to number of detectors. However, all these equations can be defined according to [13] and [15] for pairwise diversity measurements as shown in Table 2 and equation 4.
Table 2. 2-by-2 table for relationship between two detectors.

| Detector1 Hit | Detector1 Miss |
|---------------|---------------|
| N^{11}        | N^{10}        |
| N^{01}        | N^{00}        |

\[ \kappa = \frac{2(N^{11}N^{00} - N^{01}N^{10})}{(N^{11} + N^{10})(N^{01} + N^{00}) + (N^{11} + N^{01})(N^{10} + N^{00})} \] (4)

According to equation 4 and Table 2, the all combinations of the detectors, which are mentioned in Section 3, are observed and their pairwise \( \kappa \) values are determined. The results can be interpreted in light of this information Table 3 to 7 gives the details.

Table 3. The diversity between DPM and ACF trained with Caltech and INRIA datasets respectively.

|              | DPM & ACF(w/Caltech) – SS True | DPM & ACF(w/Caltech) – SS False |
|--------------|--------------------------------|---------------------------------|
|              | DPM Hit | DPM Miss | DPM Hit | DPM Miss |
| ACF Hit      | 8660    | 70426    | 9275    | 75971    |
| ACF Miss     | 1547    | 166519   | 932     | 160974   |
| \( \kappa \) | 0.1697  |          | 0.1753  |          |

|              | DPM & ACF(w/INRIA) – SS True | DPM & ACF(w/INRIA) – SS False |
|--------------|-------------------------------|-------------------------------|
|              | DPM Hit | DPM Miss | DPM Hit | DPM Miss |
| ACF Hit      | 8987    | 8616     | 9516    | 10524    |
| ACF Miss     | 1220    | 228329   | 691     | 226421   |
| \( \kappa \) | 0.6321  |          | 0.6162  |          |

Table 4. The diversity between DPM and HOG trained with “UprightPeople_96x48”.

|              | MergeDetections True – Class. Thresh. = 1 | MergeDetections False – Class. Thresh. = 0 |
|--------------|------------------------------------------|--------------------------------------------|
|              | DPM Hit | DPM Miss | DPM Hit | DPM Miss | DPM Hit | DPM Miss |
| HOG Hit      | 7597    | 12006    | 10052   | 26225    |
| HOG Miss     | 2610    | 224939   | 155     | 210720   |
| \( \kappa \) | 0.4877  |          | 0.4199  |          |

|              | MergeDetections False – Class. Thresh. = 1 |
|--------------|------------------------------------------|
|              | DPM Hit | DPM Miss |
| HOG Hit      | 9260    | 17658    |
| HOG Miss     | 947     | 219287   |
| \( \kappa \) | 0.4825  |          |
Table 5. The diversity between ACF trained with Caltech dataset and ACF trained with INRIA dataset.

|                | ACF (w/ Caltech) – SS True & | ACF (w/ Inria) – SS False |
|----------------|-------------------------------|---------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| ACF (w/Inria) Hit | 16845 | 3195 |
| ACF (w/Inria) Miss | 62241 | 164871 |
| $\kappa$       | 0.2890                       |

|                | ACF (w/ Caltech) – SS False & | ACF (w/ Inria) – SS True |
|----------------|-------------------------------|--------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| ACF (w/Inria) Hit | 16439 | 1164 |
| ACF (w/Inria) Miss | 68807 | 160472 |
| $\kappa$       | 0.2872                       |

Table 6. The diversity between ACF trained with Caltech dataset and HOG trained with “UprightPeople_96x48”.

|                | SS True & MergeDetections True – Class. Thresh. = 1 |
|----------------|-----------------------------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| HOG Hit        | 17296 | 2307 |
| HOG Miss       | 61790 | 165759 |
| $\kappa$       | 0.3069 |

|                | SS True & MergeDetections False – Class. Thresh. = 0 |
|----------------|-----------------------------------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| HOG Hit        | 28498 | 7779 |
| HOG Miss       | 50588 | 160287 |
| $\kappa$       | 0.3987 |

|                | SS False & MergeDetections True – Class. Thresh. = 1 |
|----------------|------------------------------------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| HOG Hit        | 18141 | 1462 |
| HOG Miss       | 67105 | 165976 |
| $\kappa$       | 0.3081 |

|                | SS False & MergeDetections False – Class. Thresh. = 0 |
|----------------|-------------------------------------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| HOG Hit        | 30840 | 5437 |
| HOG Miss       | 54406 | 156469 |
| $\kappa$       | 0.4223 |

|                | SS False & MergeDetections False – Class. Thresh. = 1 |
|----------------|-------------------------------------------------------|
|                | ACF (w/Caltech)Hit | ACF (w/Caltech)Miss |
| HOG Hit        | 24955 | 1963 |
| HOG Miss       | 60291 | 159943 |
| $\kappa$       | 0.3926 |
Table 7. The diversity between ACF trained with INRIA dataset and HOG trained with “UprightPeople_96x48”.

| SS True & MergeDetectors True – Class. Thresh. = 1 | SS True & MergeDetectors False – Class. Thresh. = 1 |
|-------------------------------------------------|-------------------------------------------------|
| ![Table content](image1.png)                     | ![Table content](image2.png)                     |

| SS False & MergeDetectors False – Class. Thresh. = 0 | SS False & MergeDetectors False – Class. Thresh. = 1 |
|---------------------------------------------------|---------------------------------------------------|
| ![Table content](image3.png)                      | ![Table content](image4.png)                      |

On the other hand, all these feature-classifier models can be combined as non-pairwise, for example the diversity of quartet models can be determined and analysed according to their total accuracy. Tables 8-13 give the detailed results.

Table 8. The diversity between DPM, ACF trained with Caltech – SS True, ACF trained with INRIA – SS True and HOG “UprightPeople_96x48” – MergeDetectors True – Class. Thresh. = 1.

| ACF Caltech SS True & ACF INRIA SS True & MergeDetectors True – Class. Thresh. = 1 |
|-----------------------------------------------|-----------------------------------------------|
| DPM Miss                                      | DPM Miss                                      |
| ![Table content](image5.png)                  | ![Table content](image6.png)                  |

\( \kappa = 0.2757 \)

Number of total true positives = 126240

Number of total false positives = 138745
Table 9. The diversity between DPM, ACF trained with Caltech – SS True, ACF trained with INRIA – SS True and HOG “UprightPeople_96x48” – MergeDetections False – Class. Thresh = 1.

| ACF Caltech SS True & ACF INRIA SS True & MergeDetections False – Class. Thresh. = 1 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       |
|                  | (Miss)          | (Miss)          | (Miss)          | (Hit)           | (Hit)           | (Miss)          | (Hit)           | (Hit)           |
| DPM Miss         | 163691          | 511             | 53936           | 1149            |                 |                 |                 |                 |
| DPM Hit          | 1868            | 449             | 8834            | 6507            |                 |                 |                 |                 |
| DPM Hit          | 204             | 285             | 73              | 385             |                 |                 |                 |                 |
| DPM Hit          | 213             | 845             | 730             | 7472            |                 |                 |                 |                 |

κ = 0.3115

Number of total true positives = 320134
Number of total false positives = 423144

Table 10. The diversity between DPM, ACF trained with Caltech – SS False, ACF trained with INRIA – SS True and HOG “UprightPeople_96x48” – MergeDetections True – Class. Thresh = 1.

| ACF Caltech SS False & ACF INRIA SS True & MergeDetections True – Class. Thresh. = 1 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       |
|                  | (Miss)          | (Miss)          | (Miss)          | (Hit)           | (Hit)           | (Miss)          |
| DPM Miss         | 159649          | 329             | 61614           | 3347            |                 |                 |
| DPM Hit          | 884             | 112             | 6182            | 4828            |                 |                 |
| DPM Hit          | 147             | 319             | 380             | 1764            |                 |                 |
| DPM Hit          | 62              | 404             | 631             | 6500            |                 |                 |

κ = 0.2646

Number of total true positives = 950406
Number of total false positives = 717377

Table 11. The diversity between DPM, ACF trained with Caltech – SS True, ACF trained with INRIA – SS False and HOG “UprightPeople_96x48” – MergeDetections True – Class. Thresh = 1.

| ACF Caltech SS True & ACF INRIA SS False & MergeDetections True – Class. Thresh. = 1 |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       | ACF Caltech     | ACF Inria       |
|                  | (Miss)          | (Miss)          | (Miss)          | (Hit)           | (Hit)           | (Miss)          |
| DPM Miss         | 163502          | 1454            | 56451           | 3532            |                 |                 |
| DPM Hit          | 1146            | 415             | 5320            | 5123            |                 |                 |
| DPM Hit          | 143             | 660             | 142             | 1665            |                 |                 |
| DPM Hit          | 78              | 666             | 328             | 6525            |                 |                 |

κ = 0.2819

Number of total true positives = 551664
Number of total false positives = 421083
### Table 12. The diversity between DPM, ACF trained with Caltech – SS False, ACF trained with INRIA – SS False and HOG “UprightPeople_96x48” – MergeDetections True – Class.Thresh = 1.

|                  | ACF (Caltech) Hit | ACF (Inria) Miss | ACF (Caltech) Miss | ACF (Inria) Hit | ACF (Caltech) Miss | ACF (Inria) Miss | ACF (Caltech) Hit | ACF (Inria) Miss |
|------------------|-------------------|------------------|-------------------|----------------|-------------------|-----------------|-------------------|-----------------|
| DPM Miss         |                  |                  |                   |                |                   |                 |                   |                 |
|                  |                   |                  |                   |                | 159359            | 619             | 60594             | 4367            |
| DPM Hit          |                  |                  |                   |                |                   |                 |                   |                 |
|                  | 860              | 136              | 5608              | 5402           |
| DPM Hit          |                  |                  |                   |                | 112              | 354             | 173               | 1971            |
|                  | 48               | 418              | 358               | 6773           |

\[ \kappa = 0.2748 \]

Number of total true positives = 1375830

Number of total false positives = 999715

### Table 13. The diversity between DPM, ACF trained with Caltech – SS False, ACF trained with INRIA – SS False and HOG “UprightPeople_96x48” – MergeDetections False – Class.Thresh = 1.

|                  | ACF (Caltech) Hit | ACF (Inria) Miss | ACF (Caltech) Miss | ACF (Inria) Hit | ACF (Caltech) Miss | ACF (Inria) Miss | ACF (Caltech) Hit | ACF (Inria) Miss |
|------------------|-------------------|------------------|-------------------|----------------|-------------------|-----------------|-------------------|-----------------|
| DPM Miss         |                  |                  |                   |                |                   |                 |                   |                 |
|                  | 159096            | 530              | 57645             | 2016           |
| DPM Hit          |                  |                  |                   |                | 1123              | 225             | 8557              | 7753            |
|                  | 95               | 222              | 50               | 580            |
| DPM Hit          |                  |                  |                   |                | 65               | 550             | 481               | 8164            |

\[ \kappa = 0.3141 \]

Number of total true positives = 1569724

Number of total false positives = 1284114

### Table 14. The diversity between DPM, ACF trained with Caltech – SS False, ACF trained with INRIA – SS False and HOG “UprightPeople_96x48” – MergeDetections False – Class.Thresh = 0.

|                  | ACF (Caltech) Hit | ACF (Inria) Miss | ACF (Caltech) Miss | ACF (Inria) Hit | ACF (Caltech) Miss | ACF (Inria) Miss | ACF (Caltech) Hit | ACF (Inria) Miss |
|------------------|-------------------|------------------|-------------------|----------------|-------------------|-----------------|-------------------|-----------------|
| DPM Miss         |                  |                  |                   |                |                   |                 |                   |                 |
|                  | 156165            | 244              | 53744             | 567            |
| DPM Hit          |                  |                  |                   |                | 4054              | 511             | 12458             | 9202            |
|                  | 17               | 43               | 4                 | 91             |
| DPM Hit          |                  |                  |                   |                | 143              | 729             | 527               | 8653            |

\[ \kappa = 0.3235 \]

Number of total true positives = 1993790

Number of total false positives = 4334429
According to detection results for all feature-classifier models as shown in Table 1, the complementarity of these models is observed with measurement of interrater agreement (i.e. κ) for pairwise and non-pairwise analyzes which are shown in Tables 3 to 13.

The complementarity of DPM model in pairwise analyzes is insufficient for models except ACF trained with Caltech. The κ between Caltech trained ACF and DPM models has the smallest value in all analyzes. This result is probably due to the differences between features and training sets. Hence, κ between DPM and INRIA trained ACF models, and κ between DPM and HOG models are high since, the training sets or features are same. The same situation is expected about the complementarity between HOG and Caltech trained ACF. However, κ is not smaller as DPM, because HOG model generates much more bounding boxes and number of targets which are hit by both of them.

Rather than elimination of DPM model, analyzes of all feature-classifiers are obtained in non-pairwise part. However, the results show that, the smallest value of missed targets can be observed with Caltech trained ACF and HOG models as Table 6 when SS is false, MergeDetections is false and ClassificationThreshold is zero. Other two models reduce that number a little, but this is insignificant for ensemble models.

As a conclusion, the features and training sets are important for complementarity of feature-classifier models. However, it is also observed that the ensemble version of two or more models always detect more target compared to their individual versions.

Acknowledgements
This study is conducted as a part of EUREKA-EURIPIDES project “ADORAS [Advanced On-board Data Recording and Analysis System]” and supported by TUBITAK TEYDEB under grant 9150121, which is conducted by Kent Kart.

References
[1] C. Bila, F. Sivrikaya, M. A. Khan and S. Albayrak, "Vehicles of the Future: A Survey of Research on Safety Issues," in IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 5, pp. 1046-1065, May 2017.
[2] T. Gandhi and M. M. Trivedi, "Pedestrian Protection Systems: Issues, Survey, and Challenges," in IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 3, pp. 413-430, Sept. 2007.
[3] D. Geronimo, A. M. Lopez, A. D. Sappa and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 7, pp. 1239-1258, July 2010.
[4] R. Benenson, M. Omran, J. Hosang, B. Schiele, "Ten years of pedestrian detection what have we learned?", Proc. Eur. Conf. Comput. Vis., pp. 1-8, 2014.
[5] S. Zhang, R. Benenson, M. Omran, J. Hosang and B. Schiele, "Towards Reaching Human Performance in Pedestrian Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, no. 99, pp. 1-1, 2017. doi: 10.1109/TPAMI.2017.2700460
[6] Ess, A., Leibe, B., Schindler, K., & Van Gool, L. (2008, June). A mobile vision system for robust multi-person tracking. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8). IEEE.
[7] Enzweiler, M., & Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and experiments. IEEE transactions on pattern analysis and machine intelligence, 31(12), 2179-2195.
[8] Wojek, C., Walk, S., & Schiele, B. (2009, June). Multi-cue onboard pedestrian detection. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 794-801). IEEE.
[9] Geiger, A., Lenz, P., & Urtasun, R. (2012, June). Are we ready for autonomous driving? the kitti vision benchmark suite. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE.
Conference on (pp. 3354-3361). IEEE.

[10] Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence, 34*(4), 743-761.

[11] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 886-893). IEEE.

[12] Li, K., Wang, X., Xu, Y., & Wang, J. (2016). Density enhancement-based long-range pedestrian detection using 3-d range data. *IEEE Transactions on Intelligent Transportation Systems, 17*(5), 1368-1380.

[13] Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5*(1), 105-112.

[14] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence, 32*(9), 1627-1645.

[15] Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning, 51*(2), 181-207.