Driver Cell Phone Usage Detection From HOV/HOT NIR Images

Yusuf Artan, Orhan Bulan, Robert P. Loce, and Peter Paul
Xerox Research Center Webster
800 Phillips Rd. Webster NY 14580
yusuf.artan,orhan.bulan,robert.loce,peter.paul@xerox.com

Abstract

Distracted driving due to cell phone usage is an increasingly costly problem in terms of lost lives and damaged property. Motivated by its impact on public safety and property, several state and federal governments have enacted regulations that prohibit driver mobile phone usage while driving. These regulations have created a need for cell phone usage detection for law enforcement. In this paper, we propose a computer vision based method for determining driver cell phone usage using a near infrared (NIR) camera system directed at the vehicle's front windshield. The developed method consists of two stages; first, we localize the driver's face region within the front windshield image using the deformable part model (DPM). Next, we utilize a local aggregation based image classification technique to classify a region of interest (ROI) around the drivers face to detect the cell phone usage. We propose two classification architectures by using full face and half face images for classification and compare their performance in terms of accuracy, specificity, and sensitivity. We also present a comparison of various local aggregation-based image classification methods using bag-of-visual-words (BOW), vector of locally aggregated descriptors (VLAD) and Fisher vectors (FV). A data set of 1500 *images was collected on a public roadway and is used to perform the experiments.*

1. Introduction

Recent statistics show that a high number of accidents have been caused by distracted drivers due to mobile phone usage while driving [2, 9]. According to the study in [2], 995 out of 5474 (18%) who were killed in car accidents in 2009 were considered to be killed by drivers distracted by mobile phones. Due to its impact on public safety and property, several states and countries have enacted regulations to ban mobile phone usage while driving. As of now, 10 states, D.C., Puerto Rico, Guam and the U.S. Virgin Islands prohibit all drivers from using hand-held mobile phones while driving [1]. Except for Maryland and West Virginia (until July 2013), all laws are primary enforcement; an officer may cite a driver for using a hand-held mobile phone without any other traffic offense taking place. To enforce the law, current practice requires dispatching law enforcement officers at the road side to visually examine incoming cars, or employing human operators to manually examine image/video records to identify violators. Both of those practices are expensive, difficult, and ultimately ineffective. Therefore, there exists a strong desire to automatically or semi-automatically detect a violator, i.e., driver using a hand-held mobile phone while driving.

Transportation imaging cameras, widely deployed in highways, toll booths and traffic lights for applications such as red light enforcement, license plate recognition, or parking occupancy estimation, may form a means for detecting driver cell phone violation. Most of these already deployed cameras, however, are either not NIR or do not direct towards the driver and hence, are not useful for detecting cell phone violations. Among these transportation imaging systems, the cameras installed to manage High Occupancy Vehicle (HOV) and High Occupancy Tolling (HOT) lanes provides an opportunity for driver cell phone violation detection as these cameras have NIR capabilities to enable night vision and are directed towards the front windshield of a vehicle to estimate the vehicle occupancy from captured images. Several image-based vehicle occupancy detection systems have been examined in the literature [10, 7, 3] to estimate passenger occupancy from HOV/HOT images.

In this paper, we examine cell phone usage detection as part of an automatic HOV/HOT lane enforcement system. Other modules of the HOV/HOT lane enforcement system would include front/rear seat detection and passenger counting. However, in this work we focus only on cell phone violation detection from front view images captured by a HOV/HOT imaging system. Our experimental setup includes a camera-based imaging system to capture still images in the near-infrared (NIR) band through the windshield of an approaching vehicle. NIR imaging is commonly used in HOV/HOT imaging systems so that the illumination does not distract drivers.

Our algorithm for cell phone violation detection consists of two stages: first, we utilize an elastic deformation model for object detection [13] to first localize the windshield and then the driver face within the captured image. Next, we develop a machine-learning-based image classifier that detects cell phone violations using a global image representation of an ROI around the detected face region.

For image classification, we consider two architectures using full face images and half face images and compare their performance in terms of sensitivity and specificity. For each architecture, we train a linear SVM using locally aggregated descriptors such as BoW [5], VLAD [12] and FV [11], and conduct a comparative study using these descriptors to perform image classification in cell phone usage detection task.

The organization of this article as follows. Section 2 briefly summarizes the image acquisition procedure for HOV/HOT lane management. In Sec. 3, we describe the details of our methodology for driver cell phone usage detection from front view HOT/HOV NIR images using the deformable parts model for face detection and using known descriptors for image classification. Evaluation of the methods using real world road images are presented in Sec. 4. Sec. 5 discusses the key findings of the present study.

2. HOV/HOT Image Acquistion

A large roadway dataset was collected using a HOV/HOT NIR imaging system. The distance from camera to cars is on average 60-feet and the horizontal field of view of the camera is approximately 12-feet at that distance. Figure 1 shows NIR images acquired by the HOV/HOT cameras.

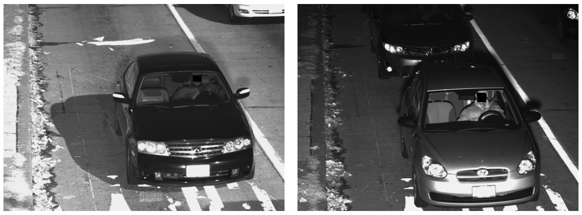

Figure 1. NIR images acquired by a HOV/HOT lane front view camera.

Note that there is a large variation in image intensity due to variations in windshield transmission, illuminator state, sun position, cloud position, and other factors even though images were captured by an NIR camera [4]. The captured vehicle images show a wide variety of human faces, facial poses, occlusions, and other expected 'within class' variations. Our methodology in the next section needs to comprehend these challenges for reliably locating the driver face and performing the classification task for driver cell phone usage.

3. Driver Cell Phone Usage Detection

An overview of our methodology for driver cell phone detection is illustrated in Fig. 2. Our algorithm uses the front view images captured by a HOV/HOT image acquisition system, where the first step is detection of the driver's face in the captured image. Detecting the driver's face from the entire HOV/HOT NIR image, however, is challenging as the driver's face is often partially occluded by the sun visors, windshield mirror, or car roof. We therefore, constrain the region of interest by first detecting the windshield of the vehicle. The driver's face is later searched in the right half of the detected windshield using a state-of-the-art face detection algorithm [13]. Following face detection, we extract a patch around the detected face and adopt an image classification approach to identify driver cell phone usage while driving.

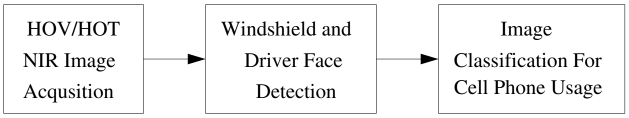

Figure 2. An overview of the methodology for detecting driver cell phone usage from HOV/HOT NIR images.

3.1. Windshield and Driver Face Detection

For detecting the driver face in the captured HOV/HOT images, we adopted the DPM based face detector method [13], which considers pose estimation and face detection problems jointly using a deformable parts model. This model forms a mixture of trees with a set of shared parts V, where each facial part is considered as a node in the tree. The model then forms an optimization problem by jointly optimizing appearance and shape of the facial parts. For a given image, this optimization problem is solved using dynamic programming to find the best configuration of facial parts. While this method works well for detecting faces with different poses, detecting faces with partial occlusions still remains to a challenge for this method.

Figure 3 illustrates several instances of partial occlusions in HOV/HOT images due to sun visors, windshield mirror, or car roof. In order to detect faces in these cases while not causing false positives/wrong detections, we restrict the search space by first detecting the windshield of a vehicle in the captured HOV/HOT image. The reason for detecting the windshield prior to driver face detection is to constrain the search space in the captured images. The driver face is later searched in the right half of the detected windshield. Restricting the region of interest for face detection provides flexibility for adjusting the detection threshold. Instead of setting a fixed threshold for face detection, we pick the best candidate with the highest score calculated by the face detector as the driver's face, which enables detection of faces with partial occlusions whose score otherwise would not pass the detection threshold.

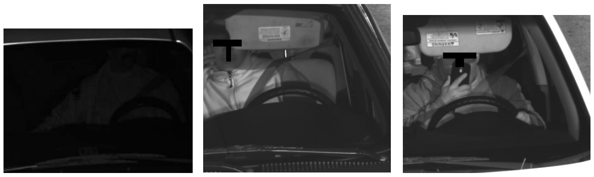

Figure 3. Examples for driver face images with partial occlusions.

Using the framework described in [13], we generated a windshield model by positioning a set of landmarks and their relative locations around the windshield, the 13 landmark points are labeled manually in a set of training images similar to Fig. 4 in the same sequence. In our implementation, we used 15 positive and 20 negative images for training the windshield model. Negative samples were selected from images that do not contain windshields in the scene. In our model, rather than generating a mixture model, we consider a single topological view as the images are captured with a fixed camera and vehicles are always driving towards the camera along the same direction and angle with respect to the image plane. We generated a linearly parametrized tree-structured model $T = (V, E)$, where V is the set of parts and E is the set of edges between parts. A windshield score for a particular configuration of landmarks $L = \{l_i; i \in V\}$ in a given image I is defined as

$$\begin{aligned} S(I, L) &= \mathit{App}(I, L) + \mathit{Shape}(L) \\ &= \sum_{i \in V} w_i \cdot \phi(I, l_i) \\ &+ \sum_{ij \in E} a_{ij} dx^2 + b_{ij} dy^2 + c_{ij} dx + d_{ij} dy \end{aligned} \tag{1}$$

Note that this score function is a abridged version of the general score function defined for mixture of trees [13] for a single tree-structure. In this function, $\phi(I, l_i)$ represents the histogram of gradients (HoG) features extracted at pixel location l_i [6], and $l_i = (x_i, y_i)$ stands for the pixel location of part i. The appearance evidence of each of the landmarks is included in the App and the evidences for the spatial locations of the landmarks with respect to each other is included in the Shape term. In [8], this model was viewed as a linear classifier with unknown parameters w_i and $\{a_{ij}, b_{ij}, c_{ij}, d_{ij}\}$, which is learned during training using a latent SVM. The training constructs a model by learning the appearance at each landmark point and the relationship between points as shown in Figure 5.

For an incoming image I, we identify a list of candidate windshield areas by maximizing the score function Eq. (2) over L using dynamic programming to find the best configuration of parts [13, 8].

$$S^*(I) = \max_L S(I, L) \tag{2}$$

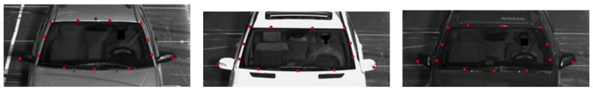

Figure 4. Training landmarks (red points) overlaid on top of the cropped windshield image.

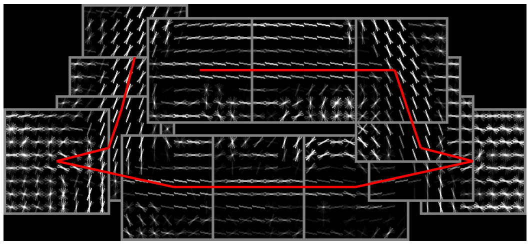

Figure 5. Windshield model learned using SVM based method presented in [13].

3.2. Image Classification For Driver Cell Phone Usage Detection

After detecting the driver's face in the captured image, a region of interest is defined around the detected face and it is extracted from the image. The extracted image patch is analyzed through image classification to detect and identify a cell phone violation. For performing the image classification task, we consider local invariant descriptors that are aggregated into an image level vector signature, which is subsequently passed into a classifier. We use three locally aggregated descriptors BoW, VLAD, and FV due to their extensive usage and success in image classification and categorization tasks.

Among these descriptors, BoW has the oldest history and was the initiative for other locally aggregated descriptors [5]. From a set of training images, BoW first calculates dense-SIFT features and constructs a codebook of visual words that consist of K-centroids calculated by either k-means or Gaussian mixture model clustering algorithms applied on the calculated SIFT features. The dimension d for the SIFT feature vectors was 128. For an image in query I, local image descriptors $I = (x_1, x_2, ..., x_N)$ are calculated and assigned to the closest cluster centers. Following the assignment step, a histogram of local descriptors is calculated and aggregated to generate the final image signature vector.

Similar to BoW, VLAD is a feature encoding method for image classification tasks that first constructs a vocabulary of visual words by clustering dense-SIFT features of a set of training images. The visual vocabulary is generated by $K-$clusters calculated by either k-means or Gaussian mixture model clustering algorithms. Each cluster is represented by the cluster mean μ_k. For a query image, VLAD first calculates local image descriptors and assigns them to the closest cluster centroids. For the descriptors assigned to the same cluster, it calculates a total distance vector from the cluster mean as

$$v_k = \sum_{x_i:NN(x_i)=k} (x_i - \mu_k) \tag{3}$$

The final VLAD descriptor is formed by the concatenation of the d-dimensional distance vectors v_k for each cluster as $\phi(I) = [v_1, v_2, ..., v_K]$.

Fisher vectors have been recently flourished as the probabilistic version of the VLAD and have been reported to achieve the best performance in several image classification and categorization tasks [12]. FVs incorporate generative models into discriminative classifiers [11] by first fitting a $K-$dimensional Gaussian mixture model to local image descriptors calculated from a set of training images. Assuming the mixture weight, mean vector and variance matrix (assumed diagonal) for the k^{th} Gaussian is represented by w_k, μ_k, and σ_k, for a query image $I = (x_1, x_2, ..., x_N)$, FVs first calculate the d-dimensional gradient vector u_k with respect to the mean μ_k of the k^{th} Gaussian as:

$$u_k = \frac{1}{K\sqrt{w_k}} \sum_{i=1}^{N} \alpha_k(i)(\frac{x_i - \mu_k}{\sigma_k}) \tag{4}$$

where $\alpha_k(i)$ is the assignment of the descriptor x_i to the k^{th} Gaussian. The final FV $\phi(I) = [u_1, u_2, ..., u_K]$ is the concatenation of the u_k vectors for $k = \{1...K\}$ and has a dimension of $K \times d$.

After calculating locally aggregated image descriptors using either BoW, VLAD, or FV, we utilize a linear SVM to construct the classification model to perform the image classification for cell phone usage detection.

4. Experiments

In this section, we evaluated the performance of the proposed algorithm for cell phone violation detection. The algorithm is implemented in Matlab and tested on a set of images acquired by the HOV/HOT image acquisition system as described in Sec. 2.

4.1. Database

Using the HOV/HOT image acquisition system, we captured 1500 front-view vehicle images as shown in Fig. 1. We manually label the captured images for ground truth for driver cell phone usage. In the collected image set, we had 378 drivers using cell phone and 1122 drivers without using cell phone. The resolution of the captured images was 2352×1728 pixels. Windshield region of the these images were on average 1000×360 pixels, and resolution of face region was larger than 100×100 pixels.

The HOV/HOT NIR camera captured and saved the images in $16-$bit tif format. Prior to processing the images in the pipeline of the proposed algorithm, we converted to 16-bit images to 8-bit format and applied a Wiener filtering with 3×3 windows to mitigate the impact of image acquisition noise due to low light. We also performed adaptive histogram equalization to improve the contrast of the captured images.

4.2. Windshield and Driver Face Detection

We performed windshield detection on the captured images using the model described in Sec. 3.1 and applied DPM-based face detector on the right hand side (i.e., driver side) of the detected windshield. For windshield detection, we set the detection threshold to -0.65. After detecting the windshield, we selected the highest score window as the driver's face returned by the face detector in the right side of the windshield. Our sequential methodology successfully detected faces in 1445 images out of 1500 images in our database, yielding 97% accuracy. Figure 6 shows some of the challenging faces with partial occlusions that were detected by the proposed sequential algorithm. Note that the extracted patches after face detection is 80 pixels wider and 20 pixels higher than the detected face to capture the cell phone usage around the detected face region.

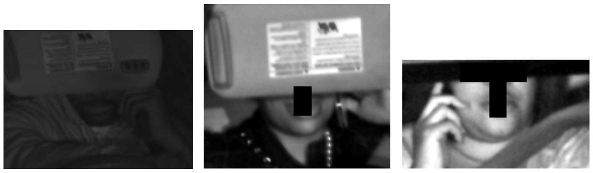

Figure 6. Detected faces with partial occlusions.

4.3. Image Classification For Cell Phone Usage

Following driver face detection, we evaluated the performance of classification on the extracted image patches using local aggregation methods (i.e., BoW, VLAD, and FV). For classification task, we consider two different architectures. In the first architecture, we train a classifier using the full face images as shown in Fig. 6. In the second architecture, we divided face images into two separate parts; drivers holding phone with left hand and right hand. Figure 7, for example, shows a set of left half face images using a cell phone. We then trained a classifier for each side using locally aggregated descriptors and fuse the results of each classifier to make a final decision on driver cell phone usage.

For locally aggregated descriptors, we extract features from 32×32 pixel patches on regular grids (every 4 pix-

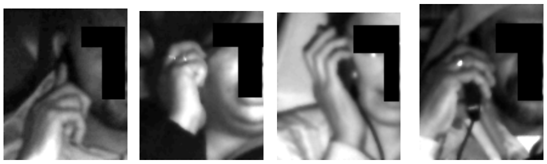

Figure 7. Left half faces used to train a classifier for drivers using a cell on the right hand side.

els) at 5 scales. We only extract 128-D SIFT descriptors for these image patches. For all descriptors (i.e., BoW, VLAD, FV), we used gaussian mixture models (GMM) with $K = 256$ clusters to compute the descriptors. The GMM's are trained using the maximum likelihood (ML) criterion and a standard expectation maximization (EM) algorithm. Similar to [12], we apply the power and L2 normalization to descriptors to improve the classification performance. In the case of bag of visual words, we follow a spatial pyramid based BoW representation in which we create histograms for the original image and spatial grid partitioning of the original image. We concatenate histograms from the original image, 2×2 and 4×4 regular spatial grid.

Once the features are extracted (for positive and negative training images) using one of the local aggregation methods presented above, we train a linear SVM classifier to perform the classification task.

4.4. Performance Across Different Training Sizes and Descriptors

We first evaluated the performance of the local aggregation based methods' across different training set sizes. In our experiments, we evaluated the performance by performing 5-fold cross validation for various number of training samples for each of the classification architecture.

4.4.1 Classification From Full Face Images

In this architecture, we used full face images to train the classifier for cell phone usage. Figure 8 illustrates the performance across different training set sizes for all three descriptors. As expected, as the number of training images increases the classification performance tends to improve but the improvement diminished after a certain number of training samples. For FV, for example, the best performance is obtained with 180 training images.

Setting the number of training images to 180, we plotted ROC curve for all three descriptors as shown in Fig. 9. In line with earlier studies [3, 12], FV typically performs better in terms of accuracy than VLAD and BoW features.

Figure 10 presents the cell phone usage detection results for several sample images in HOV/HOT lanes using FV descriptors. FV has correctly determined the presence of cell-phone usage in Row 1 images. Parts (e)-(f) and (g)-(h) shows the misses and false alarms obtained with this approach, respectively. Note that false alarms arises in cases where the driver has an object occluding part of driver's face.

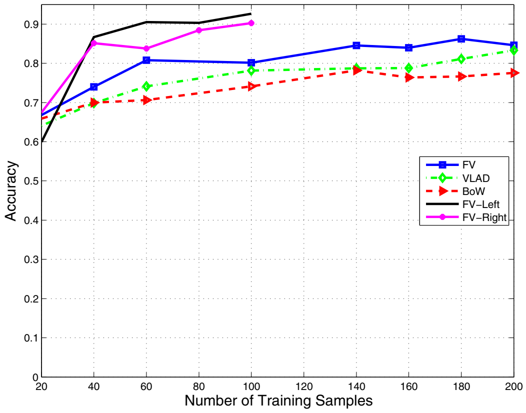

Figure 8. Mean accuracy obtained with three local aggregation methods for various number of training samples.

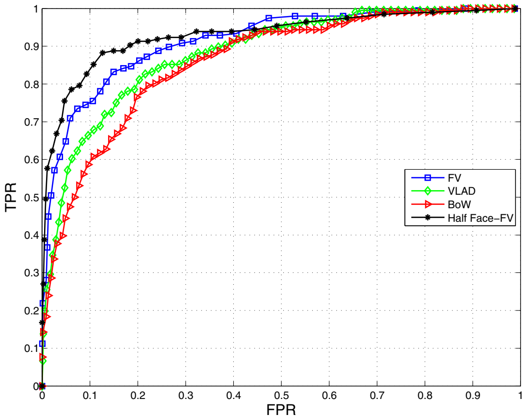

Figure 9. ROC curve for FV, VLAD, BoW and half-face FV features.

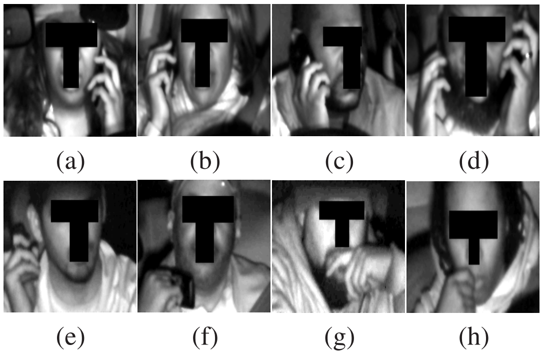

Figure 10. Performance of the FV based cell phone usage detection on test images. Part (a)-(d) presents the correctly detected cell phone usage violations, part (e)-(f) and (g)-(h) present the misses and false alarms obtained with this approach, respectively.

Table 1. Classification accuracy, sensitivity and specificity for full and half face classification architectures using Fisher Vectors (FV).

	Full Face	Half Face
Accuracy	86.19	85.85
Sensitivity	76.19	86.84
Specificity	88.18	84.44

4.4.2 Classification From Right and Left Side Face Images

Given the fact that FV outperformed VLAD and BOW in Sec. 4.4.1, we evaluated the performance for right and left side face images only using FV descriptors. Figure 8 illustrates performance for left and right side driver face images for different number of training samples. Note that number of training images for individual left-face and right-face classifiers are half of the original cell phone users. Similar to Fig. 8, the classification performance shows a significant improvement as the number of training images increases. Mean accuracy for individually trained right-face and left face FV classifiers are better than full-face FV classifier (92% vs 86%). However, this does not reflect the overall cell-phone detection accuracy since individual classifiers only consider one side of the face region.

4.5. Performance Across Different Classification Architectures

We finally compare the performance of two different classification architectures after fusing the results of the classifiers for right and left side face images. In the fusion, we declare a violation if any of the left/right side classifier detects a cell phone usage. In the experiment, we set the number of training images to 180 and used FV as the descriptor as suggested in Sec. 4.4.1. We performed a $5-$fold cross-validation to evaluate the performance. Table 1 shows side-by-side performance for both of the full and half face classification architectures in terms of classification accuracy, specificity, and sensitivity. Recall rate is higher for half-face FV architecture, but full-face FV has better specificity and overall accuracy. Since the performances of two architectures are similar to each other, we have also generated an ROC curve for half face FV classification algorithm as shown in Fig. 9. It can be easily deduced that half face FV architecture generates a better cell phone usage detector since it has a larger area under the ROC value compared to that of full-face architecture.

5. Conclusion

In this paper, we propose a method for detecting driver cell phone usage from an NIR camera system directed at the vehicle's front windshield. Proposed system utilizes a local aggregation based image classification technique to classify an ROI around the drivers face to detect the cell phone usage. The ROI around the drivers' face is found using landmark based DPM models. Experiments are presented using 1500 real-world images captured on a city roadway to evaluate the performance of the proposed system. Proposed system yields an accuracy rate above 86% for the cell phone usage detection task.

References

[1] Regulation for driver phone usage while driving. http://www.ghsa.org/html/stateinfo/laws/cellphone_laws.html. 1

[2] U.S. DOT national highway traffic safety administration distracted drive report. http://www-nrd.nhtsa.dot.gov/Pubs/811379.pdf. 1

[3] Y. Artan, A. Burry, F. Perronnin, and P. Paul. Comparison of face detection and image classification for detecting front seat passengers in vehicles. In *IEEE Winter Conference in Applications of Computer Vision*, 2014. 1, 5

[4] N. Bouguila, R. I. Hammoud, and D. Ziou. Advanced offline statistical modeling of eye and non-eye patterns. In *Passive Eye Monitoring*, pages 55–81. Springer, 2008. 2

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–22, 2004. 2, 3

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 3

[7] W. Daley, O. Arif, J. Stewart, J. Wood, C. Usher, E. Hanson, J. Turgeson, and D. Britton. Sensing system development for hov/hot (high occupancy vehicle) lane monitoring. Technical report, Georgia Department of Transportation, 2011. 1

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 3

[9] R. Hammoud, M. Smith, R. Dufour, D. Bakowski, et al. Driver distraction monitoring and adaptive safety warning systems. In *Proceedings of SAE Commercial Vehicles Engineering Congress and Exhibition, Chicago, IL*, 2008. 1

[10] X. Hao, H. Chen, and J. Li. An automatic vehicle occupant counting algorithm based on face detection. In *Signal Processing, 2006 8th International Conference on*, volume 3. IEEE, 2006. 1

[11] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999. 2, 4

[12] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, 2012. 2, 4, 5

[13] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 2, 3