Carnegie Mellon University
The Robotics Institute

Technologies for Safe and Efficient Transportation
A U.S. DOT UNIVERSITY TRANSPORTATION CENTER

In-Vehicle Vision-based Cell Phone Detection

Final Report

Principal Investigator: Bernardo R. Pires
ORCID ID: https://orcid.org/0000-0003-0591-4250

Collaborators: Anirudh Viswanathan

DISCLAIMER
The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

1. Problem Description

According to the NOPUS survey, at any given daylight moment across America approximately 660,000 drivers are using cell phones or manipulating electronic devices while driving [1]. According to the same survey, this number has held steady since 2010, despite major investments into awareness programs and numerous changes in legislation. Recently, there has been significant interest into automatic detection of driver distraction. Such research often focuses on the driver's eyes in an attempt to detect gaze direction (determine where the driver is looking at.) The difficulty with such approach is that it either requires active infrared illumination, which can be "blinded" by the sun, or requires significant computation to recognize the driver's face, determine pose, and estimate gaze. Furthermore, this approaches often requires high-resolution cameras in order to be able to accurately observe the user's eyes. Instead of focusing on the driver's eyes, this report describes an approach where we obtain an overhead or over-the-shoulder view of the car interior with the objective of determining if the driver is holding or using a cell phone or other electronic device. We expect this to be a superior method because the screen is often illuminated, relatively large and, when in use, turned directly towards the user's head and, consequently, to the over-the-shoulder camera.

2. Approach

The objective of this project is to achieve automatic detection of the use of electronic devices (e.g. cell-phones) by the driver of a motor vehicle. The described approach acknowledges that not all uses of electronic are equally dangerous and attempts to specifically detect the presence of distracting screens. The proposed method uses an over-the-shoulder camera to observe and alert the driver in the event of distracted driving. We use a detection via classification approach using several hand-engineered features. Results on a novel dataset collected with several drivers simulating distracted driving when using a cell-phone are presented to demonstrate the efficacy of the vision-based approach.

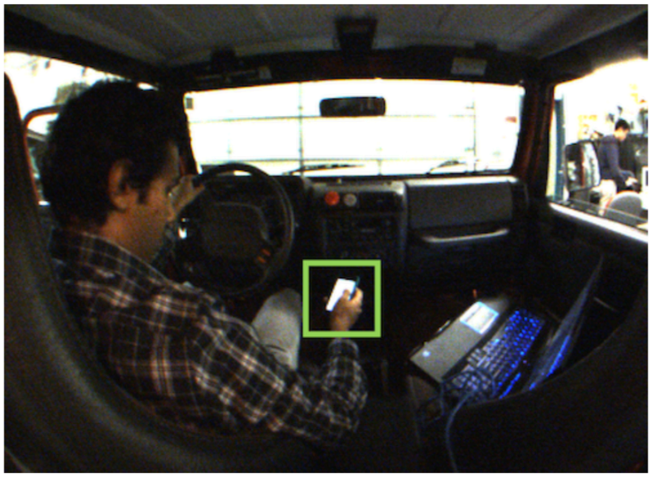

FIGURE 1. View from an over-the-shoulder camera of the interior of a vehicle. The vision-based approach developed in this project detects the use of a cell-phone by the driver of the vehicle (green box).

3. Introduction

Automated solutions to detect driver distraction have the potential to increase road safety and reduce distraction-related fatalities. We developed a novel approach to in-vehicle vision-based detection of cell-phone use by the vehicle driver as shown in Figure 1. Recent solutions to detecting driver distraction typically rely on a combination of face-detection and gaze-tracking. During the detection stage, the face of the vehicle operator is localized. The gaze-tracking stage is able to determine whether the driver is focused on the road or is distracted from the task of driving. The challenges with gaze-tracking are twofold. Firstly, in order to track the eyes, active perception techniques such as infrared cameras are used. The IR cameras can be "blinded" by direct sunlight. Additionally, the face detection stages, pose estimation, and gaze-tracking stages can require significant computation. Second, any eye-tracking mechanism requires high-resolution cameras to focus on the driver's eyes. It is a common practice to wear sunglasses when driving, which causes degraded performance of such eye-tracking systems.

The contributions of this project differ from traditional approaches in the following ways: 1) an over-the-shoulder view is used to determine if the driver is using or holding an electronic device. 2) The approach does not require the use of face-detection and high accuracy eye-tracking. The presented approach relies on the fact that the screen is illuminated when in use by the driver, and consequently visible to the over-the-shoulder camera.

4. Related Work

Advanced Driver Assistance Systems (ADAS) are being adapted to monitor the activity within a vehicle or perform "Cockpit Activity Assessment". In reference [8], the authors design a stereo-camera system to observe the driver's head and gaze direction. The driver's viewing area is divided into four regions, comprising the road ahead, windshield, left, and right mirrors. An attention mapping stage determines whether the driver is distracted. The method in [8], however, requires a specialized stereo-camera setup and the explicit need to map the driver's viewing area, while this project uses a monocular camera.

The authors in [12] propose a novel "Visual Context Capture, Analysis and Televiewing (VCAT)" system. An omnidirectional camera is used to capture the view of the interior of the vehicle, and regions of the road outside the windshield. Gaze detection is performed for the driver of the vehicle, and a synthetic image from the driver's viewpoint is generated. The authors design the system to be able to alert the driver with regards to blind spots and other distractions. Again, although relevant and promising, this work differs from the present project for its need of a highly specialized camera and complex computation.

Recent work in face detection and analysis of facial expressions has led to automated detection of driver distraction. E.g., "IntraFace" is a state-of-the-art facial analysis package [5]. The authors show driver distraction results based on a monocular camera observing the driver's face. The detection stage uses supervised descent [15] and expression analysis is carried out using selective transfer machine [3]. This work does not use infrared cameras and so is not "blinded" by the presence of strong sunlight. However, it is still highly sensitive to the use of eyewear (and, in particular, sunglasses) by the driver.

A "Head mounted Eye-tracking Device (HED)" is proposed in [13]. The authors describe a comprehensive system for eye-tracking and gaze detection using custom designed hardware, and describe the issues related to defining a "distracted gaze." A comprehensive review of sensing technology to characterize driver behavior (e.g., drowsiness) is found in [11]. Head-mounted devices have superior accuracy to almost any other method of driver distraction, but the need for the driver to wear a device greatly limits these applications in real-life scenarios.

5. Data Collection

All data was collected using the NavLab11 platform at Carnegie Mellon University's Robotics Institute. Liability issues restricted data collection to conditions when the vehicle was stationary. Consequently, unmodeled environment dynamics when the vehicle is in motion are beyond the scope of this work.

Data was collected with three different drivers (Figure 2). Each driver had a distinctive vehicle operation characteristic, and pattern of cell-phone use. For example, while one driver was dominantly right handed and used the phone only with the right hand, another driver often swapped hands when using the phone. Another observation was that one driver preferred to use the cell-phone rested against the steering wheel while another driver never brought the phone up near the steering wheel.

Each driver was asked to simulate real driving conditions for the entire data collection period. This included turning on the stereo at the start of the data collection cycle. Going through all car controls and performing typical operations such as adjusting the stereo and mirrors. During the course of simulated driving, the driver was asked to use the cell-phone

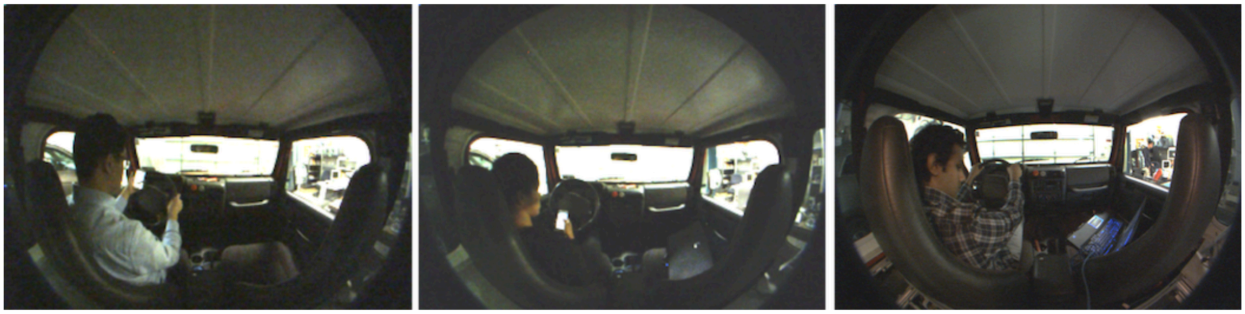

FIGURE 2. Images captured by the Pointgrey Flea2 camera mounted with a Fujinon Fish-eye lens. The three drivers have visually different driving styles and seat adjustment preferences. Additionally, while drivers 1 and 2 (from left) use the cell phone around the steering wheel, driver 3 only uses the phone at waist level.

for typical tasks such as looking at navigation information and replying to a text message. Individual images in each dataset were labeled with ground truth locations of the cell-phone. The bounding box annotation was performed manually after collection of each dataset, and over 125,000 frames were labeled. The focus of the method described in this report is to detect events when the driver was using the cell-phone during the data collection run.

6. Method

An overview of the method is shown in Figure 3. Detection of cell-phone use is performed via classification of individual patches within the image. The images used in the report are 3 channel, 8-bit RGB color images.

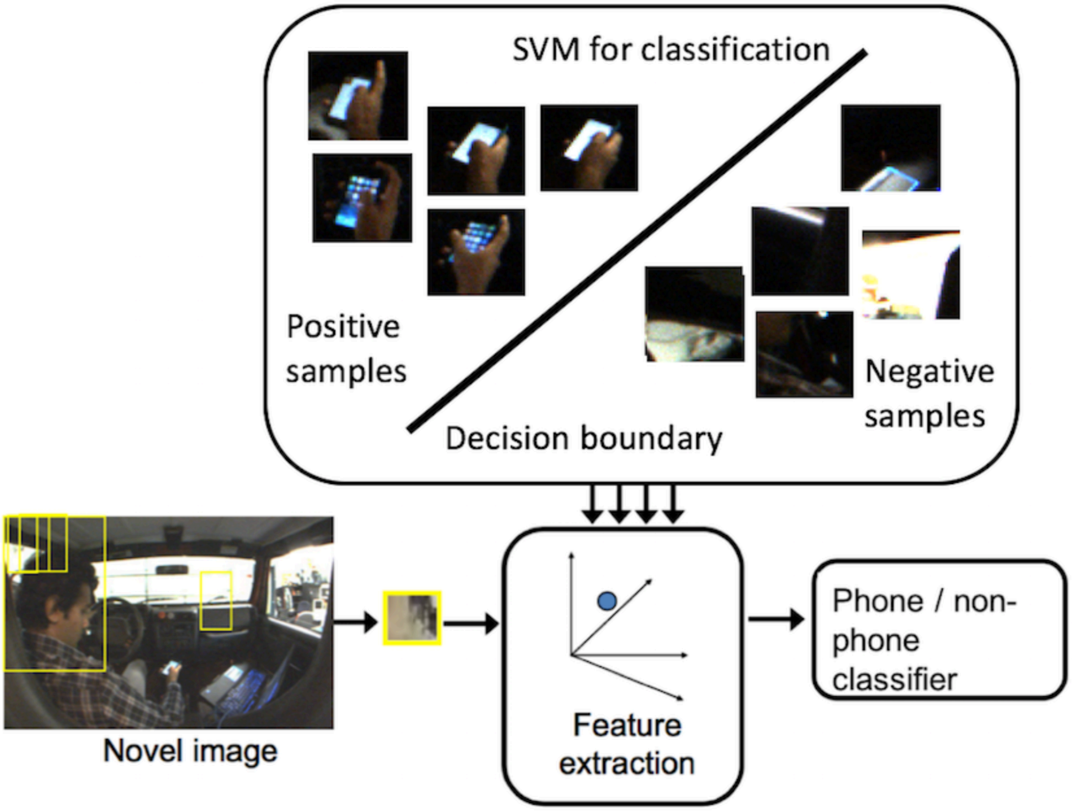

FIGURE 3. An overview of the approach. An SVM is trained using features computed on patches corresponding to positive and negative instances of cell-phones. At test time a sliding window detector performs classification of each patch in the test image [7].

6.1 Terminology

The following notation is adopted:

- I – over-the-shoulder image observing the driver (dimensions m x n)
- M – pixelwise mask of region around driver (dimensions m x n)
- P – image patch (region of interest) (dimensions p x q, p≪m, q≪n)
- yi ∈ {phone, non-phone} – patchwise binary label
- χ = {(P1, y1), . . . , (Pn, yn)} set of patches corresponding to training data with associated labels
- φ(P) – feature computed from patch
- X = {φ(Pi) | yi = phone} – set of features corresponding to positive instances
- N = {φ(Pi) | yi = non-phone} – set of features corresponding to negative instances
- h : φ(P) → y – classifier to predict the label of test patch

6.2 Detection via Classification

The method uses a sliding window-based approach to cell- phone detection. Each window is independently classified as to whether it corresponds to a cell-phone as shown in figure 4. This section describes the training mechanism for the classifier and detection stage.

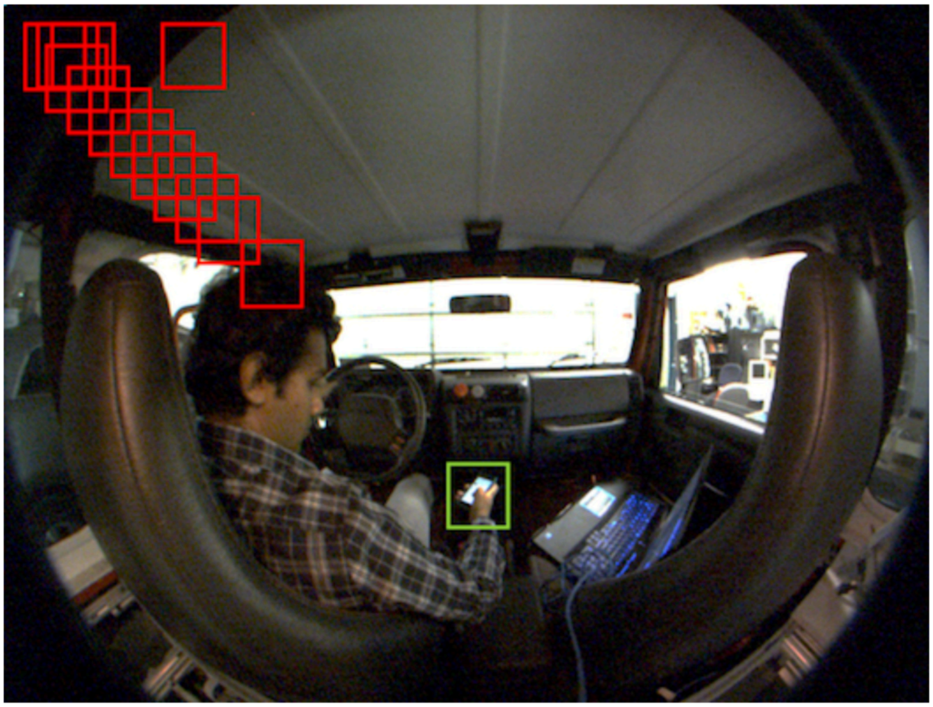

FIGURE 4. The sliding window detector shows several patches corresponding to negative samples (red) and a single positive patch (green). Since the space of all possible negative patches is very large, and the most informative negative patches are obtained using hard-negative mining.

6.3 Classifier Training

The set of training patches χ comprise of regions with cell-phones as positive instances, and randomly sampled patches (without overlap) which do not contain the phone. Each patch is represented in a corresponding feature space, φ(P). A linear Support Vector Machine (SVM) is trained on each φ(P), to learn the corresponding weight vector w and bias b which separates the two classes. The set of positive instances $X \subset \varphi(\chi)$ are separated from all the negatives $N \subset \varphi(\chi)$ by the largest possible margin the corresponding feature space. To learn the weight vector w, the SVM minimizes the convex objective function of the following form:

$$\Omega(\mathrm{w}, \mathrm{b}) = ||\mathrm{w}||^2 \ + \mathrm{C} \sum_{\mathrm{x} \in \mathrm{X} \cup \mathrm{N}} \mathrm{h}(\mathrm{w}^T \mathrm{x} + \mathrm{b})$$

The hinge loss function is of the form h = max(0, 1 − x). The SVM is driven by a small number of examples near the decision boundary. Since the space of all possible negative instances is very large, hard-negative mining is used on a subset of negative patches [10]. The training process alternates between evaluating the model on error-prone patches and adding the corresponding examples to the training set. The current implementation uses the vl-feat library for feature extraction [14]. Initially, positive instances and randomly sampled non-overlapping patches are used to train a model. Subsequently, the method alternates between mining the data for hard negatives and retraining the model. During the negative mining step, data augmentation is performed for the negative samples. An analysis of the patches corresponding to hard-negatives revealed that the regions with strong linear gradients were often selected, e.g., top left corner of the windshield. Intuitively, the top-right corner should also be a hard negative. The hard-negatives are augmented by performing a left-right flip on the corresponding patch and adding the reflected patch to the training set.

6.3 Evaluation on test data

The dataset is collected from three different drivers. Data from a single driver is used during the training stage, and data from the other two drivers are used for testing. Each data collection run lasted twenty minutes per driver, and the overall dataset comprises of over an hour of footage simulating driving conditions and cell phone use. The patches corresponding to positive samples containing the cell-phone are shown in Figure 5. The scale of the detection is known during the training stage. At test time, a sliding window classifier is run at the previously known range of scales and classify each window using the SVM. In order to reduce the false positive rate, classification is performed only for patches which fall within the probabilistic mask M which is learned from data during training time (Figure 6). The mask can also be manually specified by the user around a region of interest encompassing the driver. The mask represents a prior belief on possible locations of phone use. The score map during detection is convolved with the mask, to obtain the posterior locations of the phone. Non-maximal suppression is applied to the convolved score map and the top scoring bounding boxes corresponding to the phone location are returned.

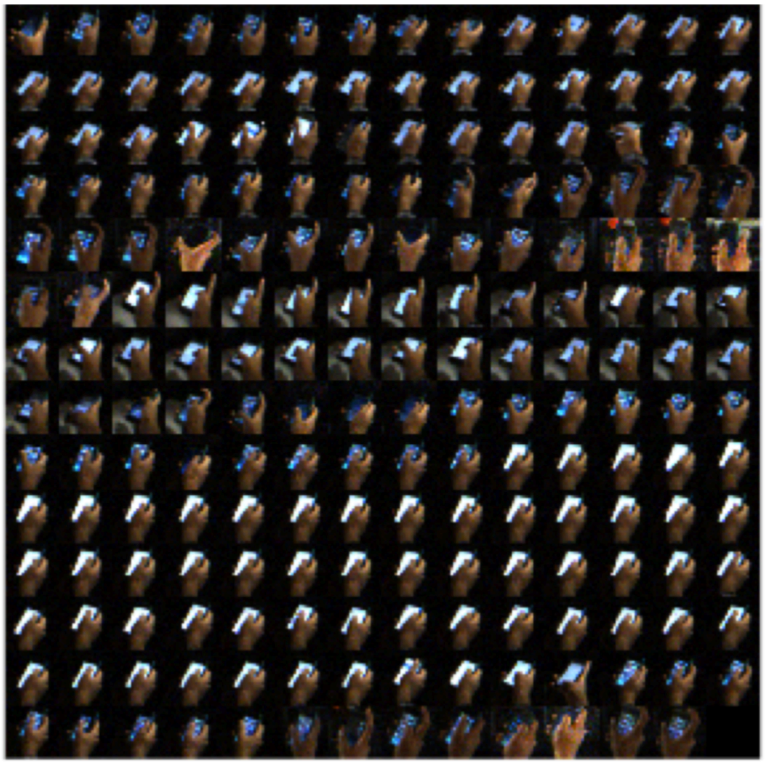

FIGURE 5. Training images corresponding to ground truth locations of the cell-phone. The dominant role of the hand in addition to the brightly lit screen, requires that the approach must be invariant to skin tone. The images captured are from a driver who is right-handed. During training, the dataset is augmented using the left-right flip to be able to adapt to left-handed cell phone users.

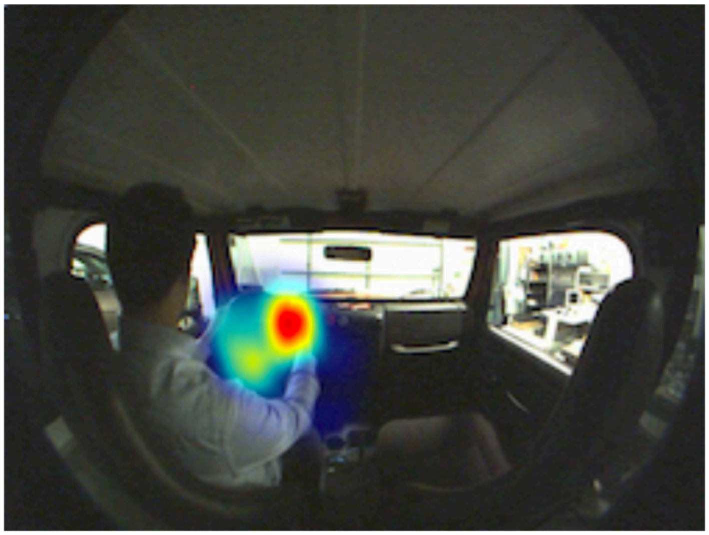

FIGURE 6. The mask within which the detection is performed is overlaid has a heatmap. The current driver is dominantly right-handed. Consequently, the heatmap is relatively focused to capture the region spanned by the driver's right hand.

7. Results

Four hand-engineered features for the task of cell-phone detection are compared. The features evaluated are SIFT [9], color histograms, FREAK [2], and HoG [4], [6]. The precision-recall curves for each feature space are shown in Figure 7. The learned color model and HoG template of the cell-phone is shown in Figure 8.

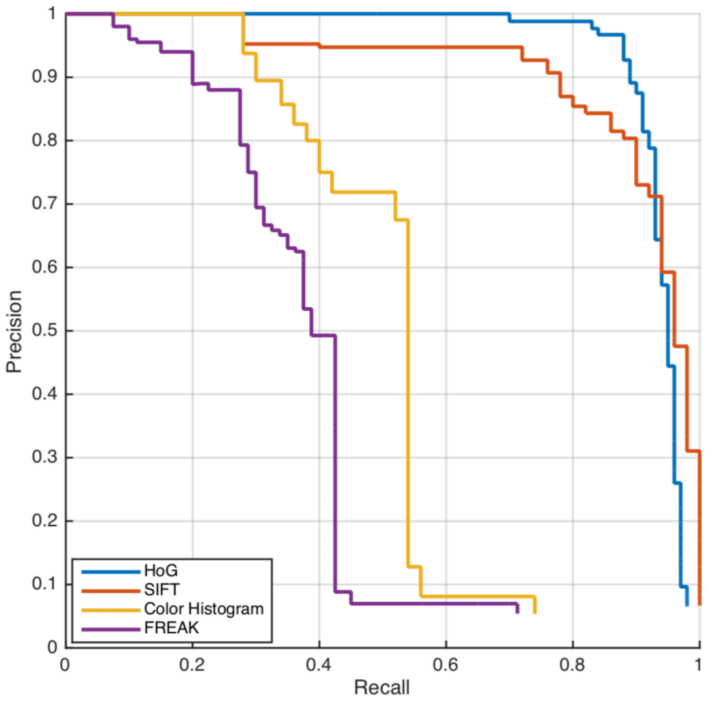

FIGURE 7. Precision-recall curves characterizing the performance of different feature space for cell-phone detection. The gradient-based feature spaces perform the best, since the gradients associated with the brightly lit phone screen are stable. The change in ambient lighting during recording causes intensity-based features to become unstable which results in lower performance.

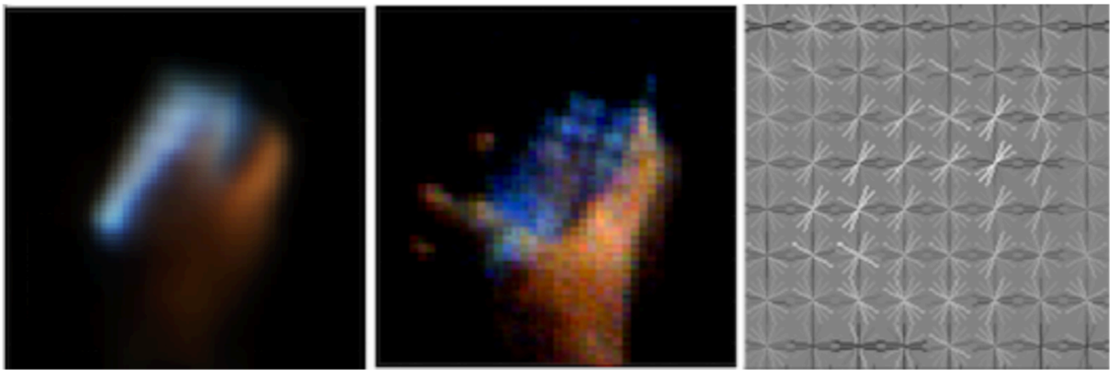

FIGURE 8. From left: Average image of training data, visualization of the weight vector on each color channel, and rendering of the learned HoG template.

The results indicate that HoG features are well suited for the cell-phone detection task since the gradient associated with the boundary of the cell-phone is well characterized when the device is in use as seen in Figure 9.

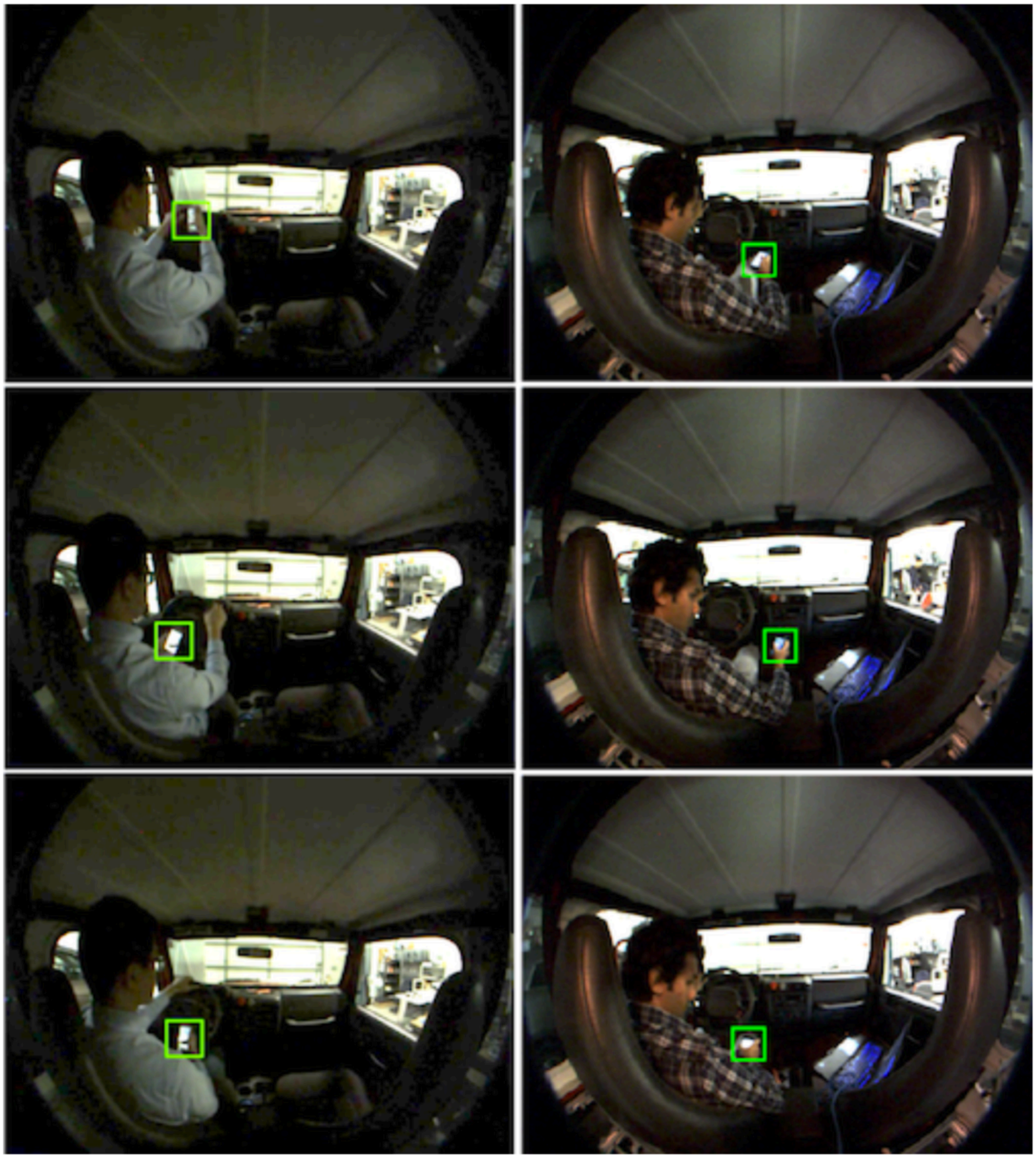

FIGURE 9. Successful detection results on test data. The figures on the left column show the driver switching hands for phone use. The location of the camera was slightly shifted for the figure on the right, and the light at the interior of the vehicle was switched-on resulting in different lighting conditions.

The detection is also successful for cases when the phone is partially occluded by the driver as shown in Figure 10. Since a certain region of the screen is visible to the camera, the detection is successful. The failure cases are when the phone is used with an out of plane rotation by more than 60 degrees as shown in Figure 11. In this case, the only portion of the phone visible to the camera is a narrow region of the screen and the buttons on the side. The template fails to generalize to this case due to the relatively rare occurrence of out of plane rotation.

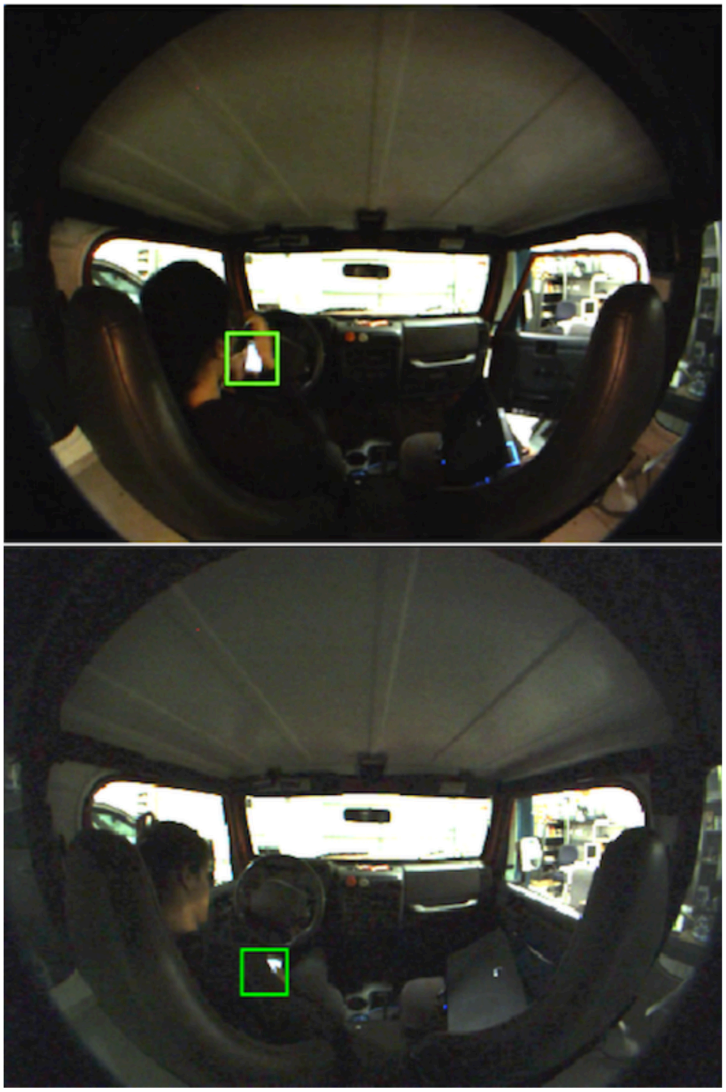

Figure 10. The vision-based method is able to successfully localize the cellphone screen, despite partial occlusion by the driver's hands and shoulders.

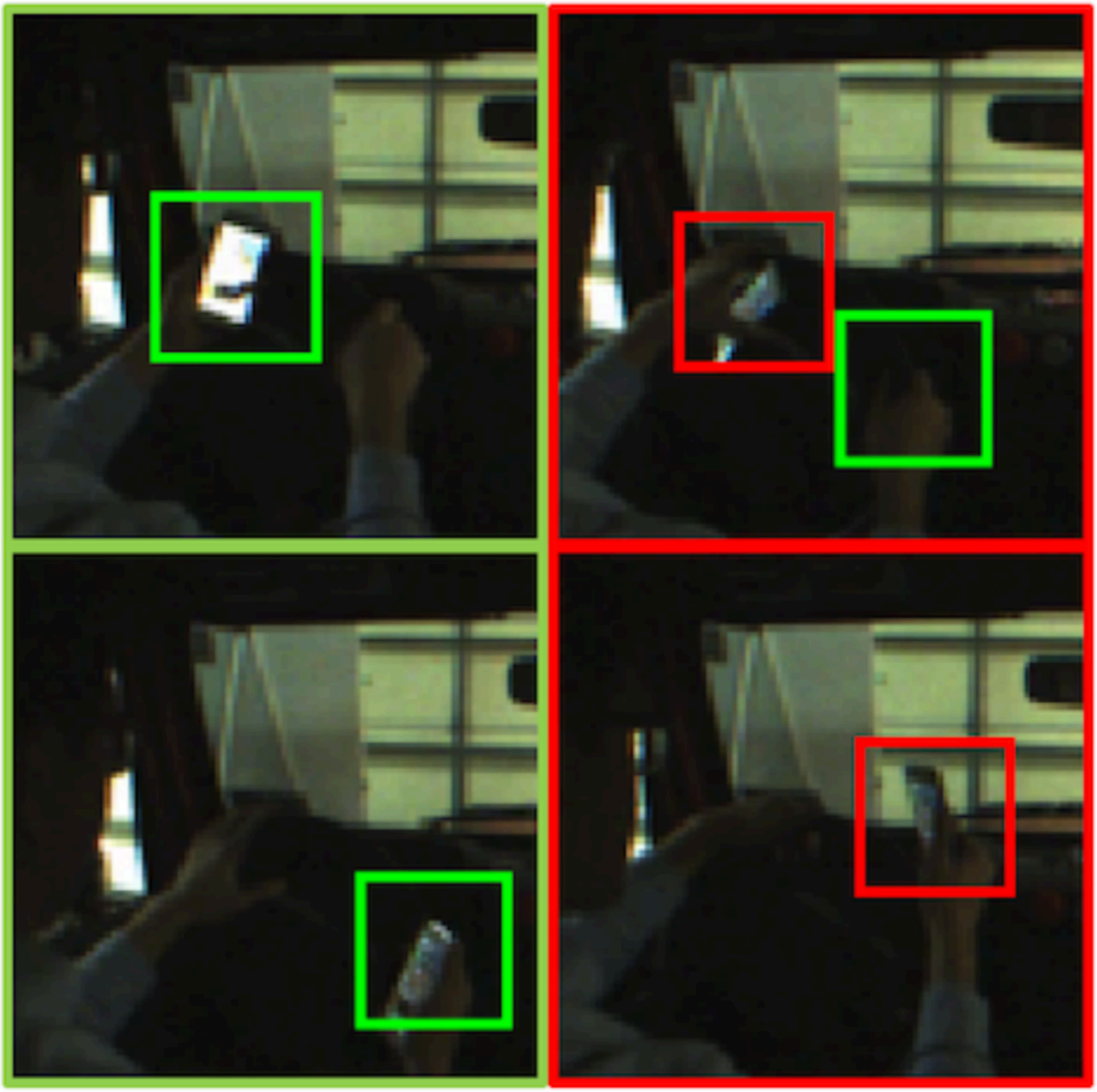

FIGURE 11. The images on the left show correct detections when the driver uses the phone with the left hand, and over the steering wheel. The failure cases are shown in the right column. The large appearance change caused by out of plane rotation causes the detection to fail.

The detection is also tested for domain adaptation on images of cell-phone use collected from the web. The images exhibit significant variation in illumination, scale, and background. The method is able to successfully adapt across different domains, as shown in Figure 12.

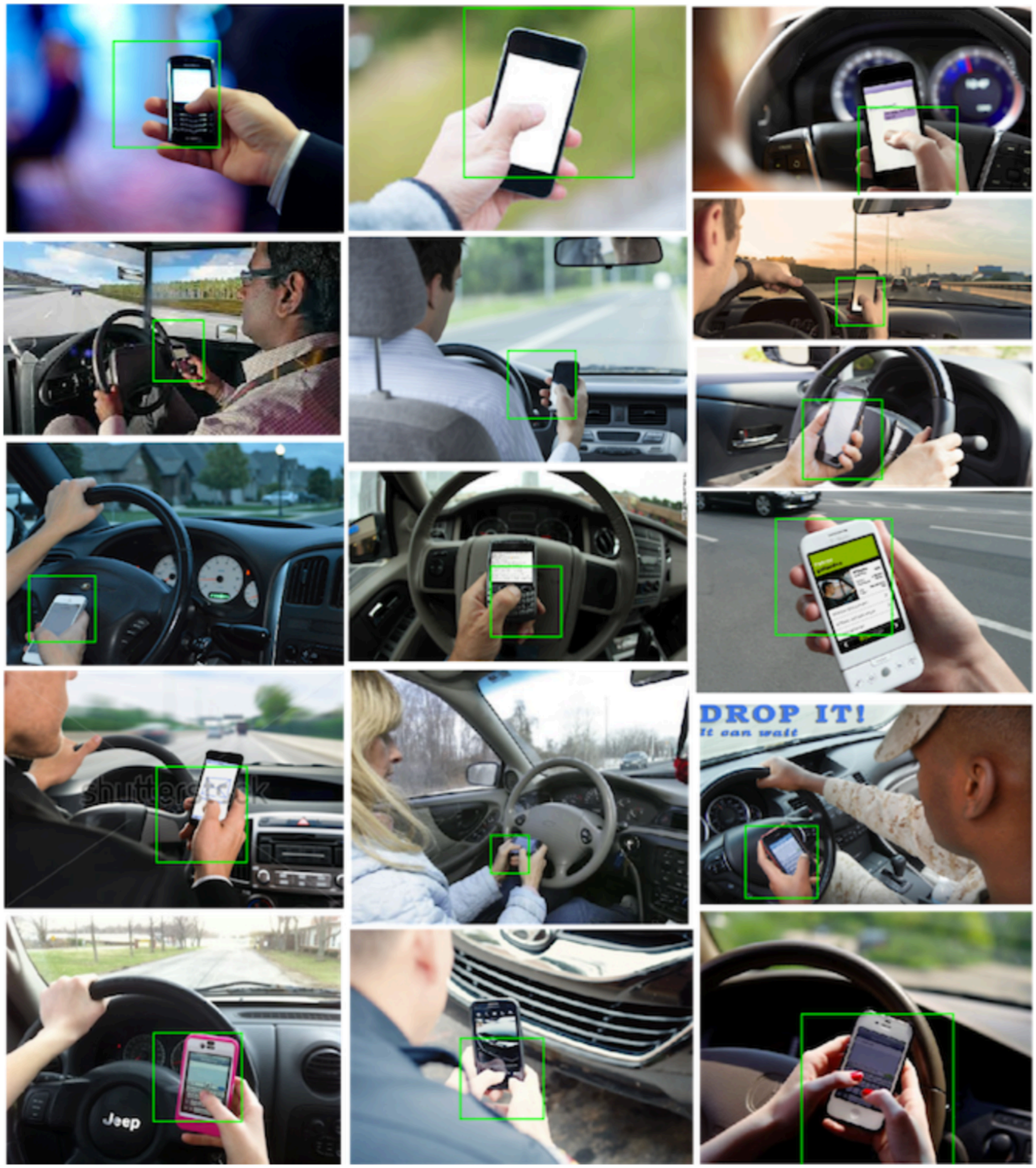

FIGURE 12. Successful detection results on test data. The figures on the left column show the driver switching hands for phone use. The location of the camera was slightly shifted for the figure on the right, and the light at the interior of the vehicle was switched-on resulting in different lighting conditions.

8. Conclusion and Future Work

This report presents a framework for automatic detection driver distraction when using a cell-phone or another handheld electronic device. An over-the-shoulder-camera is used to observe the vehicle operator. A sliding window detector is used to classify individual patches in the image. A quantitative comparison of four different hand-engineered features is presented. The best performance is obtained by using the HoG descriptor. Future extensions to the work include exploration of additional feature spaces and incorporating human body pose estimation into the framework.

Acknowledgements We acknowledge our colleagues at the NavLab group at Carnegie Mellon University. We are especially grateful towards Jian (Michael) Gong and Nikitha Poddatur for help with the data collection and labeling.

References

[1] D. H. . 719, “Driver electronic device use in 2011,” http://www-nrd.nhtsa.dot.gov/Pubs/811719.pdf, 2013.

[2] A. Alahi, R. Ortiz, and P. Vandergheynst, “Freak: Fast retina keypoint,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, June 2012, pp. 510–517.

[3] W.-S. Chu, F. De la Torre, and J. F. Cohn, “Selective transfer machine for personalized facial action unit detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[4] 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, vol. 1, June 2005, pp. 886–893 vol. 1.

[5] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, “Intraface,” in IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015.

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

[7] K. Grauman and B. Leibe, Visual object recognition. Morgan & Claypool Publishers, 2010, no. 11.

[8] M. Kutila, M. Jokela, G. Markkula, and M. Rue, “Driver distraction detection with a camera vision system,” in Image Processing, 2007. ICIP 2007. IEEE International Conference on, vol. 6, Sept 2007, pp. VI – 201–VI – 204.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91– 110, 2004. [Online]. Available: http://dx.doi.org/10.1023/B%3AVISI. 0000029664.99615.94

[10] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar- svms for object detection and beyond," in Proceedings of the 2011 International Conference on Computer Vision, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 89–96. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2011.6126229

[11] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," Sensors, vol. 12, no. 12, p.

16937, 2012. [Online]. Available: http://www.mdpi.com/1424- 8220/ 12/12/16937

[12] K. S.Huang, M. Trivedi, and T. Gandhi, "Driver's view and vehicle surround estimation using omnidirectional video stream," in Intelligent Vehicles Symposium, 2003. Proceedings IEEE, June 2003, pp. 444-449

[13] M. Sodhi, B. Reimer, J. L. Cohen, E. Vastenburg, R. Kaars, and S. Kirschenbaum, "On-road driver eye movement tracking using head-mounted devices," in Proceedings of the 2002 Symposium on Eye Tracking Research & Applications, ser. ETRA '02. New York, NY, USA: ACM, 2002, pp. 61–68. [Online]. Available: http://doi.acm.org/10.1145/507072.507086

[14] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[15] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.