2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Pedestrians and their Phones - Detecting Phone-based Activities of Pedestrians for Autonomous Vehicles

Akshay Rangesh, Eshed Ohn-Bar, Kevan Yuen and Mohan M. Trivedi

***Abstract*— Over the last decade, there have been many studies that focus on modeling driver behavior, and in particular detecting and overcoming driver distraction in an effort to reduce accidents caused by driver negligence. Such studies assume that the entire onus of avoiding accidents are on the driver alone. In this study, we adopt a different stance and study the behavior of pedestrians instead. In particular, we focus on detecting pedestrians who are engaged in secondary activities involving their cellphones and similar hand-held multimedia devices from a purely vision-based standpoint. To achieve this objective, we propose a pipeline incorporating articulated human pose estimation, and the use gradient based image features to detect the presence/absence of a device in either hand of a pedestrian. Information from different streams and their dependencies on one another are encoded by a belief network. This network is then used to predict a probability score suggesting the involvement of a subject with his/her device.**

***Index Terms*—Pedestrian activity recognition, human pose estimation, panoramic surround behavior analysis, highly autonomous vehicles, deep learning, computer vision, panoramic camera arrays.**

I. INTRODUCTION

With the explosion of hand-held device usage globally, smart phones have made their way into most hands. This trend is expected to continue as devices get cheaper and find more utility in our day to day lives. As of 2011, there were more phones than people in the USA, and internationally, the number of mobile phone subscriptions is an estimated 5.9 billion. Though such devices are extremely useful and even indispensable for many, it is this very dependence that is a major cause of pedestrian distraction, and possible injury. From here on-wards, we shall make use of the term *cellphone* as a placeholder for any hand-held multimedia device that a pedestrian may interact with.

Distracted walking, like distracted driving, is likely to increase in parallel with the penetration of electronic devices into the consumer market. Although driver distraction has received more attention since the turn of the century, distraction among pedestrians is a relatively nascent area of research. This is surprising given that pedestrians are in fact prone to acting less cautiously when distracted. Furthermore, a recent report by the Governors Highway Safety Association (GHSA) reveals a disturbing trend - between the mid-1970s and early 2000s, pedestrian deaths steadily declined, eventually dipping to around 11 percent of all motor vehicle fatalities. But since 2009, pedestrian fatalities have actually increased by 15 percent, climbing to 4,735 in 2013. Meanwhile, the percentage of pedestrians killed while using cell phones has risen, from less than 1 percent in 2004 to more than 3.5 percent in 2010, according to [1]. Also, the number of pedestrians injured while on their cells has more than doubled since 2005, the study shows.

The severity of this phenomenon is further reflected by the number of studies conducted over the last few years, each of which arrive at similar conclusions. In a recent study conducted by Thompson et al. [2], they conclude that nearly one-third (29.8%) of all pedestrians performed a distracting activity while crossing, with text messaging associated with the highest risk among different technological and social factors. Meanwhile, Nasar et al. [1] found that mobile-phone related injuries among pedestrians increased relative to total pedestrian injuries, and paralleled the increase in injuries for drivers, and in 2010 exceeded those for drivers. The study by Byington et al. [3] confirms this by a virtual street based simulation, stating that - while distracted, participants waited longer to cross the street, missed more safe opportunities to cross, took longer to initiate crossing when a safe gap was available, looked left and right less often, spent more time looking away from the road, and were more likely to be hit or almost hit by an oncoming vehicle. Moreover, it is noted that the demographic of individuals between ages 18-29 is more susceptible to exhibit such behavior. For a detailed report on the global nature of the pedestrian safety problem and the inadequacy of current systems in ensuring it, we refer the reader to [4].

It is also interesting to note that as the emphasis of automobile manufacturers gradually shifts towards more automated vehicles, so must the emphasis placed on preventing pedestrian distraction related injuries. In such scenarios, the intelligent vehicle must be able to gauge the risk associated with each pedestrian, and demonstrate more caution in avoiding those with large risks.

In this study, we focus only on distraction due to technological factors, particularly the use of cellphones for different tasks, and ignore social impacts such as talking or walking in a group. The contributions of this study are listed below:

- A high resolution dataset of images for the study of pedestrian distraction is presented. This enables fine-grained analysis of each pedestrian and the objects they interact with.
- A novel framework that fuses different image-based information streams to predict a probability score suggesting the engagement of a pedestrian with his/her cellphone.

The authors are with the Laboratory for Intelligent and Safe Automobiles, University of California, San Diego, La Jolla, CA 92092, USA
arangesh, eohnbar, kcyuen, mtrivedi@ucsd.edu

Fig. 1: A sample of pedestrians present in the newly proposed dataset. The dataset is naturalistic and unrestricted in the scale, size, and pose of the pedestrians.

II. Related Work

There is an abundance of work related to pedestrian detection, human activity recognition, and their classification from the last decade. However, these studies pertain to generic human activities and are not of much use in studying pedestrian distraction. Most studies that claim to recognize pedestrian activity [5]–[7] do so indoors with the help of wearable sensors. This offers little utility from an intelligent vehicle stand-point. Although not directly related to pedestrian distraction, studies like [8], [9] propose systems to predict avoid collisions with generic pedestrians thereby improving road safety. To the best of our knowledge, there have been no vision-based studies for detecting pedestrian distraction, or more specifically, cellphone engagement.

As we mentioned earlier, driver behavior and activity has been analyzed in detail over the years. Some representative studies in this domain include [10], [11]. A more general analysis of humans in the age of highly automated vehicles is provided in [12].

As far as pedestrian datasets are concerned, we have noticed a rise in datasets that provide fine-grained categorization of pedestrians. Examples of these include [13], [14] where attributes such as age, clothing, sex and weight of the pedestrian are annotated in addition to the bounding box and articulated pose. However, none of these datasets annotate or even include a considerable representation of distracted pedestrians.

III. Dataset Description

Since pedestrian distraction due to cellphone usage is more common among the college-age population, we mounted 4 GoPro cameras, each facing a different direction, on an intelligent vehicle testbed parked at an intersection in the UC San Diego campus on a busy afternoon. By capturing different viewpoints on each camera, we ensure that pedestrians are not predisposed to appear in a particular location or facing a certain direction. Furthermore, pedestrians are captured holding a variety of objects in addition to cellphones, such as bags, drinks, food and other miscellaneous items. To facilitate the finer analysis of each pedestrian, videos were captured at 2.7k resolution, resulting in pedestrians as large as 400 pixels in height in few cases. Some exemplar pedestrians are shown in Figure 1 to depict the diversity and quality of the dataset.

IV. Proposed Framework

In this framework, we suggest incorporating information on a finer scale along with holistic information from the full body pose of a pedestrian. Figure 2 provides a high-level view of our proposal. In the subsections that follow, we detail the inner workings of each module.

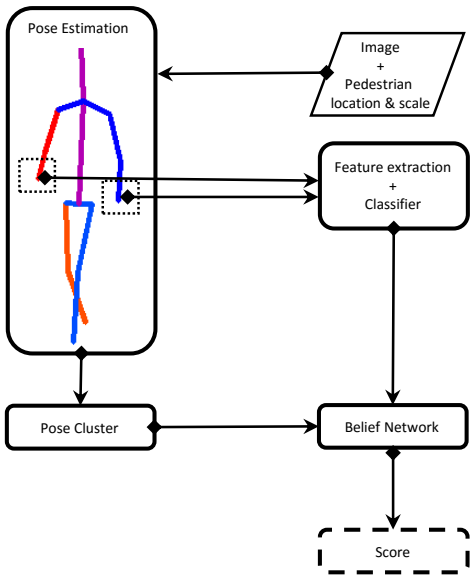

Fig. 2: Block diagram of proposed framework.

A. Pedestrian Pose Estimation and Cluster Formation

The articulated pose of a pedestrian can be an invaluable cue in estimating the activity he/she is involved in. Recent advances in pose estimation using deep convolutional neural networks (ConvNets) have led to state of the art results on challenging benchmarks. We make use of one such architecture, called the Stacked Hourglass Networks [15] proposed

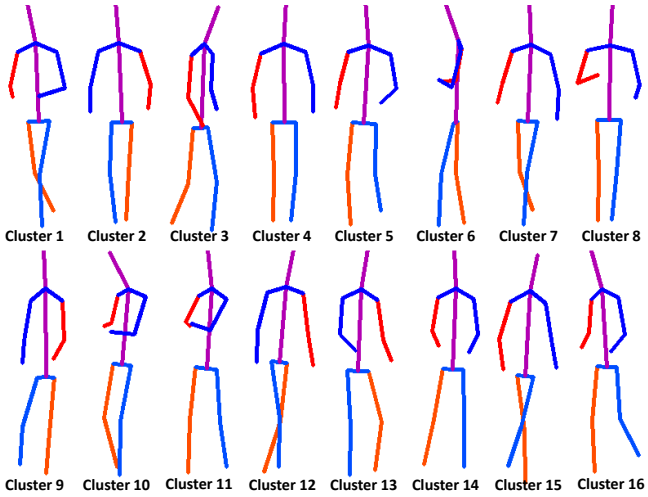

Fig. 3: Representatives obtained by clustering the articulated pose of pedestrians. Joints corresponding to the right arm are colored in red.

by Newell et al. This network has been trained on the MPII dataset [16] comprising of 25K images containing over 40K people, involved in 410 different activities, and outputs the locations of 16 joints corresponding to the articulated pose of a human body. We use this pre-trained network and fine-tune it on our own dataset. This gives us marginal improvement in performance compared to an out-of-the-box implementation. The reader is requested to refer to Figure 3 for a typical visualization of the predicted joints.

Most human pose estimation algorithms require the rough location and scale of the human in the image plane. In this study, we assume that such information is available beforehand, and focus our attention on analyzing each pedestrian in greater detail. However, if desired, the location and scale of pedestrians may be obtained easily from any generic pedestrian detector.

Consider a pedestrian bounding box parametrized as (x, y, w, h). Here, x and y correspond to image coordinates of the top left corner of the bounding box, and w and h describe the dimensions of the box. For the pedestrian under consideration, the pose estimation network outputs a set of image locations $\{(x_i, y_i)\}_{i=1,\cdots,16}$ corresponding to each joint. The set of normalized joint locations $\{(\bar{x_i}, \bar{y_i})\}_{i=1,\cdots,16}$ are then found as follows:

$$\bar{x_i} = \frac{x_i - x}{w}, \bar{y_i} = \frac{y_i - y}{h}, \forall i = 1, \cdots, 16. \quad (1)$$

Over a training dataset of 1019 pedestrians, the above process is repeated to generate a vector of normalized articulated pose for each pedestrian as described below:

$$\mathbf{x} = (\bar{x_1}, \bar{y_1}, \cdots, \bar{x_{16}}, \bar{y_{16}}) \in \mathcal{R}^{32}. \quad (2)$$

Given a set of all such pose vectors, we carry out a soft clustering of poses using a Gaussian Mixture Model (GMM) with 16 mixture components. We prefer the use of a GMM over hard clustering techniques like K-means as it tends to suppress smaller (yet considerably different) clusters which ultimately reduces the variety among cluster representatives. The choice of 16 clusters is made after observing a plot of the Bayesian Information Criterion (BIC) versus the number of components in the GMM. We scale and plot the representatives (mean values) for each cluster in Figure 3. It is interesting to note that some clusters (e.g. 6, 8 and 10) are inherently prone to higher chances of the pedestrian being engaged in cellphone activity. We make use of this observation in later sections.

(a) (b)

Fig. 4: Image patches obtained when the local window is centered around the (a)wrist versus the (b)hand.

B. Cellphone presence/absence classification

Next, we aim to determine the presence or absence of a cellphone in either hand of a pedestrian. To do so, we first regress to the approximate location of the hands of a pedestrian, assuming that it is *collinear* with the joints corresponding to the elbow and wrist. Let (x_e, y_e) and (x_w, y_w) denote the image plane coordinates of the elbow and wrist respectively. With the assumption above, the approximate location of the hand (x_h, y_h) is obtained as follows:

$$x_h = x_e + \frac{x_w - x_e}{r}, y_h = y_e + \frac{y_w - y_e}{r}, \quad (3)$$

where r is a a parameter that depends on the ratio of distances of the elbow from the wrist and hand respectively. In our experiments, $r = 5/6$ seemed to generate the best results.

Once we have the rough locations of both hands in the image plane, we crop out a local image window around these locations. The window size is chosen to be αh for a pedestrian parametrized by (x, y, w, h). Here α is a hyper-parameter that ensures that the local window scales with the size of the pedestrian. In our experiments, α is set to 0.3. The window around each hand is resized to a 64×64 image patch and HOG [17] features are extracted from it. Examples of such local patches for windows centered around both the wrist and the hand can be found in Figure 4. It is obvious that inferring the hand location, even if approximate, helps in centering the object of interest with respect to the window.

We use the same training data as before and train an SVM classifier [18] over HOG features extracted from all such image patches. We additionally augment the data by flipping and scaling the windows. A stratified 3-fold cross validation

is carried out for tuning the hyper-parameters of the SVM. An SVM with RBF kernel and parameters $C = 10$, $\gamma = 0.01$ resulted in the best cross-validation accuracy of 87.7%.

C. Belief Network

It must be noted that the presence of a cellphone alone does not imply pedestrian distraction. In fact, it is habitual for many pedestrians to simply hold their phones in one hand while walking, without engaging themselves in its use. Similarly, cluster membership for a particular pose alone cannot guarantee cellphone usage. This calls for a fusion of information from both these sources, while maintaining their dependencies on one another.

To achieve this, we propose a *belief network* over the following random variables. Let $E, C \in \{0, 1\}$ be binary random variables indicating the engagement of a pedestrian (with his/her cellphone) and the presence of a cellphone in either hand of the pedestrian respectively. A pedestrian is considered to be *engaged* if he/she is involved in any cellphone activity that places a visual or cognitive load. Examples of such activities include texting, browsing the phone, and attending a call. Next, let $K \in \{1, \cdots, 16\}$ be a random variable that denotes the cluster membership of the articulated pose of a pedestrian. Finally, let $I \in \mathcal{R}^{64\times64\times3}$ represent the local image patch around the hands of the pedestrian. With the above notation in place, our final goal is to estimate the probability of pedestrian engagement with his/her cellphone given the articulated pose and extracted image patches i.e. $\mathsf{P}(E|K, I)$.

The desired probability, after marginalizing over C and E may be written as:

$$\mathsf{P}(E|I, K) = \frac{\sum_{C\in\{0,1\}} \mathsf{P}(C, E, I, K)}{\sum_{E\in\{0,1\}} \sum_{C\in\{0,1\}} \mathsf{P}(C, E, I, K)}. \quad (4)$$

The numerator and denominator may be decomposed further to yield:

$$\mathsf{P}(E|I, K) = \frac{\mathsf{P}(E|K) \sum_C \mathsf{P}(C|E, K)\mathsf{P}(I|C, K)}{\sum_E \mathsf{P}(E|K) \sum_C \mathsf{P}(C|E, K)\mathsf{P}(I|C, K)}. \quad (5)$$

Assuming conditional independence of I and K given C, and using Bayes' rule, we may replace $\mathsf{P}(I|C, K)$ in equation 5, resulting in

$$\mathsf{P}(E|I, K) = \frac{\mathsf{P}(E|K) \sum_C \mathsf{P}(C|E, K)\mathsf{P}(C|I)\mathsf{P}(C)^{-1}}{\sum_E \mathsf{P}(E|K) \sum_C \mathsf{P}(C|E, K)\mathsf{P}(C|I)\mathsf{P}(C)^{-1}}. \quad (6)$$

In equation 6, the terms $\mathsf{P}(E|K)$ and $\mathsf{P}(C|E, K)$ are replaced by their maximum likelihood (ML) estimates. This is done by manually counting the occurrences of each random variable, as well as co-occurrences among different variables in the training dataset. It must also be noted that $\mathsf{P}(C = 1|E = 1, K) = 1$. The terms $\mathsf{P}(E|K)$ and $\mathsf{P}(C|E, K)$ encode the inclination of certain clusters to be composed of mostly engaged pedestrians, and of some clusters that generally have a higher presence of cellphones. The probability $\mathsf{P}(C|I)$ is obtained as a class confidence score from the SVM classifier trained in subsection IV-B based on the following rule:

$$\mathsf{P}(C = 1|I) = max(\mathsf{P}(C = 1|I_{left}), \mathsf{P}(C = 1|I_{right})), \quad (7)$$

where I_{left} and I_{right} are image patches corresponding to the left and right hands of the pedestrian.

Finally, $w = \mathsf{P}(C = 1)^{-1}$ can be thought of as a weighting parameter that expresses the importance assigned to pedestrian pose relative to the presence or absence of a cellphone. Increasing w makes the system more indifferent to the presence of a cellphone and assigns probabilities based purely on the pedestrian pose.

V. Experimental Evaluation

To evaluate the performance of our framework, we create a test dataset of 150 images (separate from the training dataset used earlier) amounting to a total of 182 pedestrians. Each pedestrian is manually labeled to be either engaged in using a cellphone ($E = 1$) or otherwise.

As the framework outputs a single probability score $\mathsf{P}(E|I, K)$, it needs to be thresholded to predict a final class E. Since the scores are dependent on the ML estimates and the weight parameter w, there is no necessity to restrict the threshold t to be 0.5. It is possible take a conservative approach and fix the threshold to a much lower value if necessary. In Table I, we list the classification accuracy over the test dataset for different values of the threshold t and w. We also use different values for the weight parameter $w \in \{5, 8, 10, 12, 15\}$ to study its effect on the result. Setting $w = 8$ and $t = 0.6$ results in the highest accuracy of 0.9120, and seems to agree the most with class labels assigned by human annotators. Moreover, the accuracy for extreme values of t indicates that there is a good separation between the scores assigned to the two classes.

Example results for cellphone engagement are shown in Figures 5 and 6. For the ease of discussion, we refer to the panels in Figure 5 from here on. The proposed system is seen to work reliably on different pedestrians irrespective of their orientation with respect to the camera. Even in cases where the pedestrian is walking away from the camera, the predicted scores are seen to be reasonable (panel A). Pedestrians on calls are assigned higher scores almost always, even if certain joints may be localized incorrectly (panel D). The cellphone presence classifier is seen to be extremely useful in cases similar to the one shown in panel B. Here, the pose alone assigns a high probability for cellphone engagement. However, the classifier manages to suppress this high value as no cellphone is present in either hand of a pedestrian. In a similar fashion, the pose information of the pedestrian in panel C suppresses his score for cellphone engagement, despite the presence of a cellphone. This shows that our system leverages information from both sources to overcome individual constraints and predict a coherent final score.

Panels E and F highlight some common scenarios where incorrect scores are predicted. The pedestrian in panel E is

TABLE I: Results on the test dataset for different values of weight w and threshold t. Classification accuracy is seen to peak at $w = 8, t = 0.6$.

w	t	**Accuracy**	w	t	**Accuracy**	w	t	**Accuracy**	w	t	**Accuracy**	w	t	**Accuracy**
5	0.1	0.7417	8	0.1	0.6593	10	0.1	0.6043	12	0.1	0.5659	15	0.1	0.5329
	0.2	0.8516		0.2	0.7967		0.2	0.7417		0.2	0.7362		0.2	0.6813
	0.3	0.8901		0.3	0.8516		0.3	0.8241		0.3	0.8021		0.3	0.7857
	0.4	0.9065		0.4	0.8846		0.4	0.8626		0.4	0.8571		0.4	0.8406
	0.5	0.9065		0.5	0.8956		0.5	0.8956		0.5	0.8901		0.5	0.8626
	0.6	0.8956		0.6	**0.9120**		0.6	0.9113		0.6	0.9010		0.6	0.9010
	0.7	0.8626		0.7	0.8956		0.7	0.9010		0.7	0.9010		0.7	0.9065
	0.8	0.8351		0.8	0.8406		0.8	0.8626		0.8	0.8626		0.8	0.8791
	0.9	0.8131		0.9	0.8131		0.9	0.8186		0.9	0.8241		0.9	0.8406

Fig. 5: Results of our proposal on the test dataset. Pedestrians are cropped out for better visualization. The articulated pose is overlaid on the pedestrian and the corresponding cellphone engagement score $\mathsf{P}(E|I,K)$ is written on top (in yellow). Note that each black box is referred to as a *panel* in the text.

assigned a relatively lower score despite being engaged in cellphone activity. This is because the cellphone is barely visible in the hands of the pedestrian, resulting in a low score from the classifier. The examples in panel F indicate that higher scores are assigned to pedestrians who are holding phones, irrespective of where they are actually looking. This may lead to incorrect results in rare cases where the pedestrian is momentarily looking elsewhere.

VI. Concluding Remarks

In this paper, we investigated the need for pedestrian distraction monitoring in an effort to reduce the growing number of pedestrian fatalities. To this end, a pipeline based on belief networks is proposed to fuse image information from both coarse and fine scales, and predict a final score for cellphone engagement. To train and test the validity of our proposal, a real world dataset of distracted pedestrians is presented. Such a system goes beyond pedestrian detection by assigning confidence scores indicating phone-based distraction. Pedestrians with higher scores are at a higher risk of being disengaged with their surroundings, and hence must be handled with extra care by the driver/intelligent vehicle.

Future work encompasses studying other sources of pedestrian distraction (e.g. talking, walking in a group, listening

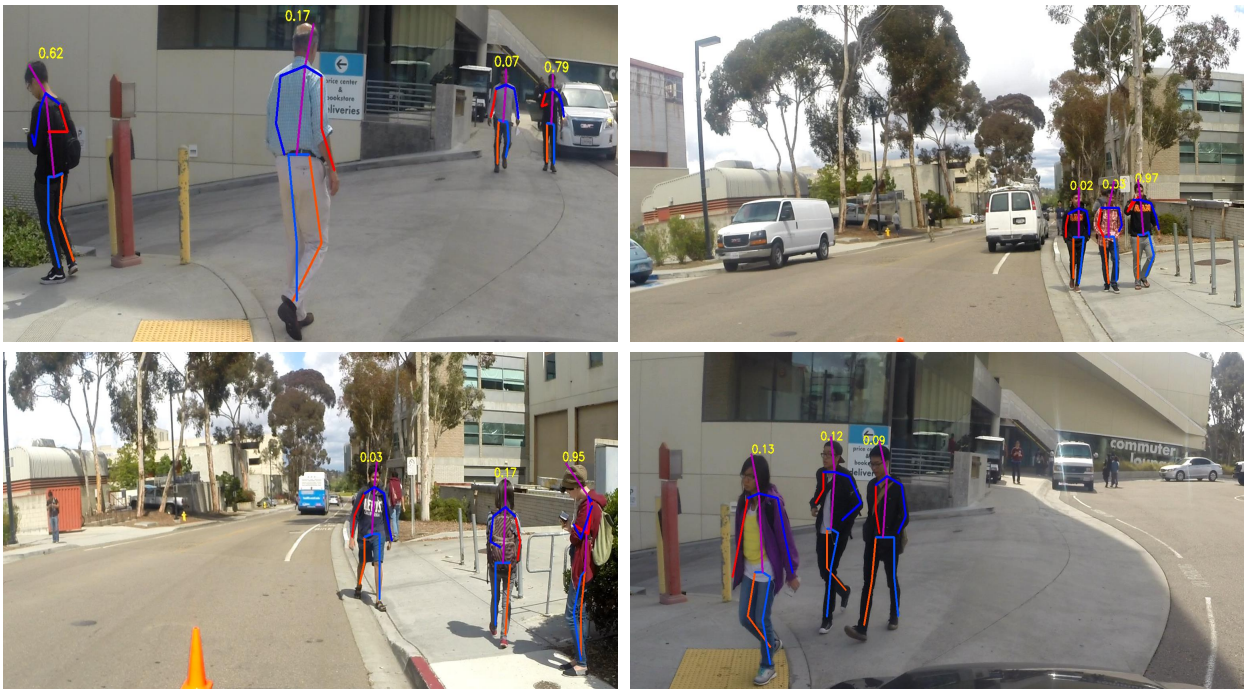

Fig. 6: More results on the test dataset. The proposed systems outputs an engagement score (in yellow) corresponding to each pedestrian detected in the image.

to music etc.), and integrating all such factors to predict a combined distraction score for each pedestrian. Additionally, the value in including other cues like head pose and gaze for pedestrian distraction needs to be examined.

VII. Acknowledgments

We thank our sponsors for their generous and continued support. We also express our gratitude to all colleagues at LISA lab, UCSD for their assistance in collecting and annotating the dataset.

References

[1] J. L. Nasar and D. Troyer, "Pedestrian injuries due to mobile phone use in public places," *Accident Analysis & Prevention*, vol. 57, pp. 91–95, 2013.

[2] L. L. Thompson, F. P. Rivara, R. C. Ayyagari, and B. E. Ebel, "Impact of social and technological distraction on pedestrian crossing behaviour: an observational study," *Injury prevention*, vol. 19, no. 4, pp. 232–237, 2013.

[3] K. W. Byington and D. C. Schwebel, "Effects of mobile internet use on college student pedestrian injury risk," *Accident Analysis & Prevention*, vol. 51, pp. 78–83, 2013.

[4] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 413–430, 2007.

[5] Z. Sun, X. Mao, W. Tian, and X. Zhang, "Activity classification and dead reckoning for pedestrian navigation with wearable sensors," *Measurement science and technology*, vol. 20, no. 1, p. 015203, 2008.

[6] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Handel, "Continuous hidden markov model for pedestrian activity classification and gait analysis," *Instrumentation and Measurement, IEEE Transactions on*, vol. 62, no. 5, pp. 1073–1083, 2013.

[7] S. Khalifa, M. Hassan, and A. Seneviratne, "Adaptive pedestrian activity classification for indoor dead reckoning systems," in *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on*. IEEE, 2013, pp. 1–7.

[8] T. Gandhi and M. M. Trivedi, "Image based estimation of pedestrian orientation for improving path prediction," in *Intelligent Vehicles Symposium, 2008 IEEE*. IEEE, 2008, pp. 506–511.

[9] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 330–335.

[10] S.Martin, A.Rangesh, E.Ohn-Bar and M. MTrivedi, "The rhythms of head, eyes and hands at intersections," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2016.

[11] E. Ohn-Bar and M. M. Trivedi, "A comparative study of color and depth features for hand gesture recognition in naturalistic driving settings," in *Intelligent Vehicles Symposium (IV), 2015 IEEE*. IEEE, 2015, pp. 845–850.

[12] E. Ohn-Bar and M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," in *IEEE Transactions on Intelligent Vehicles (to appear)*, 2016.

[13] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5482–5491.

[14] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *arXiv preprint arXiv:1603.06937*, 2016.

[16] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3686–3693.

[17] 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. IEEE, 2005, pp. 886–893.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.