NADS-Net: A Nimble Architecture for Driver and Seat Belt Detection via Convolutional Neural Networks

Sehyun Chun[1] Nima Hamidi Ghalehjegh[1] Joseph B. Choi[1] Chris W. Schwarz[2]
John G. Gaspar[2] Daniel V. McGehee[1,2] Stephen S. Baek[1,2*]
[1]Department of Industrial and Systems Engineering, University of Iowa, Iowa City, IA 52242
[2]National Advanced Driving Simulator (NADS), Iowa City, IA 52242

Abstract

A new convolutional neural network (CNN) architecture for 2D driver/passenger pose estimation and seat belt detection is proposed in this paper. The new architecture is more nimble and thus more suitable for in-vehicle monitoring tasks compared to other generic pose estimation algorithms. The new architecture, named NADS-Net, utilizes the feature pyramid network (FPN) backbone with multiple detection heads to achieve the optimal performance for driver/passenger state detection tasks. The new architecture is validated on a new data set containing video clips of 100 drivers in 50 driving sessions that are collected for this study. The detection performance is analyzed under different demographic, appearance, and illumination conditions. The results presented in this paper may provide meaningful insights for the autonomous driving research community and automotive industry for future algorithm development and data collection.

1. Introduction

A vast majority of vehicle accidents reported worldwide are caused by distracted driving behaviors [27]. Examples of distracted driving behaviors include use of smartphones/mobile devices, smoking tobaccos, engaging in a conversation with other passengers, drinking beverages, eating foods, and such, that are irrelevant to the task of driving itself. Different forms of distractions such as drowsiness, fatigues, medication effects, and other medical/physiological issues can also cause life threatening situations [16].

Another significant automotive safety hazard is caused by improper/non-use of seat belt, which can cause a serious injury and fatality. According to the U.S. National Highway Traffic Safety Administration (NHTSA), 10,428 unbuckled drivers and passengers died in 2016 on the U.S. roads [22]. Moreover, even if the drivers and passengers are buckled, improperly positioned seat belt can cause fatal injuries. According to [8], fatal injuries such as intra-abdominal injury are caused by improper positioning of seat belt at the time of crash.

To this end, *in-vehicle monitoring systems* (IVMS) are rapidly becoming a standard technology in consumer vehicles as they can play a critical role in preventing and mitigating traffic accidents by alerting the distracted driver and adaptively adjusting the safety mechanisms. Furthermore, in the upcoming era of autonomous driving, IVMS technologies are expected to be even more critical [33]. For example, an IVMS can provide an alert to the driver when the vehicle system detects an anomaly while in an autonomous driving mode, so that the driver can take over the control prior to the system failure [10]. An IVMS could also provide personalized accommodation to the occupants to maximize the comfort and safety.

For IVMS, vision-based sensing technologies are at the core, as they permit non-invasive, non-obstructive means to monitor and detect in-cabin activities. To this end, visual cues from face, eye-gaze, head-pose, hand gestures, and body poses are detected and tracked in IVMS systems [6, 21, 35, 37, 24]. The goal of vision-based sensing typically includes recognition of a variety of states, activities, and aspects of human automobile users, such as the body posture of the driver and the front row passenger, and correct donning of seat belt, which are the main objectives of this paper. More specifically, this paper proposes a new convolutional neural network (CNN) architecture for 2D driver/passenger pose estimation and seat belt detection that is more nimble and, thus, more suitable for in-vehicle monitoring tasks compared to other state-of-the-art approaches. The new architecture, named *NADS-Net*, utilizes a feature pyramid network (FPN) [18] backbone with multi-branch detection heads, namely, a key point detection head, a part-affinity field [4] detection head, and a seat belt segmentation head. The new architecture shows similar detection accuracy in the body pose estimation task compared to the state-of-the-art algorithm [4], while being more concise and efficient,

*Corresponding author: stephen-baek@uiowa.edu

and capable of doing more (*i.e*. seat belt detection). In addition, we also collected a video data set of 100 drivers in 50 driving sessions to fine-tune the performance of the proposed model pre-trained on the generic human pose estimation in the wild data sets. We analyzed the performance of the new NADS-Net algorithm as well as one of the current state of the art algorithm proposed by Cao *et al*. [4] under different demographic, appearance, and lighting conditions. This may provide insights for future algorithm design and data collection to the academic research community and the automotive industry. The major contribution of this paper is summarized as follows:

- A new architecture for driver/passenger pose estimation and seat belt detection is proposed.
- Insights for CNN algorithm design are distilled by contrasting the new architecture with other typical generic pose detection algorithm.
- Performance of the algorithms are analyzed on different imaging conditions, providing new insights and guidelines for future algorithm development and data collection.

2. Related works

2.1. Human pose estimation

In the automotive industry, human pose estimation algorithms have gained an increasing interest for their enhanced capacity in capturing kinematic posture of people without any sensor instrumentation. The taxonomy of human pose estimation methods in literature can be broadly categorized into *top-down* approaches and *bottom-up* approaches.

Top-down approaches Top-down approaches detect person bounding boxes first and then break each bounding box down into body key points and a skeleton. Papandreou *et al*. [25] employed Faster R-CNN [30] to first predict person bounding boxes and then utilized the residual network (ResNet) [13] to predict both dense heat maps and offset vectors within each person bounding box to localize key points. He *et al*. [12] proposed Mask R-CNN which extends Faster R-CNN to support both person instance segmentation and human key point detection, on top of the Faster R-CNN's bounding box detection head. Moreover, they changed the network backbone to FPN [18], which resulted performance gain in both accuracy and speed. Chen *et al*. [5] proposed cascaded pyramid network (CPN) comprised of two stages: GlobalNet and RefineNet. The CPN first detects the bounding box of a person and the cropped bounding box is passed to GlobalNet where key points are predicted with an FPN backbone. RefineNet then refines the key points predicted by GlobalNet, which, in turn, achieves more precise detection of occluded or invisible key points.

Bottom-up approaches Bottom-up approaches detect all body key points individually, first, and then parse their connections and memberships to construct person instances (*i.e*. skeletons). DeepCut [28] is an example of a bottom-up approach that detects body parts and the relations between each body parts. These outputs are then used to regress the spatial offsets of detected parts and to connect person instances. Later, DeeperCut [14] redesigned the original DeepCut algorithm by utilizing a deeper ResNet architecture to improve the body part detectors and to induce stronger pairwise scores between the body parts. However, both DeepCut and DeeperCut could not achieve a practical inference speed for real-time applications. Newell *et al*. [23] introduced a method that can simultaneously output key point locations and pixel-wise embeddings to automatically group key point detection results into individual poses. Cao *et al*. [4] proposed part affinity fields (PAF) that encompass vector fields indicating how individual key points should be connected. They augmented the convolution pose machine [34] algorithm with the PAF prediction head and employed bipartite graph matching to greedily parse skeleton instances.

In-vehicle human pose estimation Despite the fact that there have been significant breakthroughs in generic pose estimation tasks, there are only few pose estimation models in literature specifically for in-vehicle use. For example, Okuno *et al*. [24] proposed a method that predicts both human posture and face orientation in real-time. Unlike other generic posture estimation models, their model relies on a relatively shallow CNN architecture, comprised of only three convolution layers and two succeeding fully connected layers that directly regress x and y coordinates of eight body parts and face orientation angle. Their model was trained and evaluated on a custom data set comprised of images of twelve subjects, collected for their study. Yuen *et al*. [37] presented a model which modified the PAF model of Cao *et al*. [4] that only focuses on detecting the wrists and elbows of both the driver and the front passenger. They also collected their own data set that consists of real on-road scenes with varying lighting conditions to train and test the model. The method proposed in this paper substantially expands these previous works towards more comprehensive and reliable detection performance.

2.2. Seat belt

There have been ongoing efforts regarding computer vision-based detection of seat belt use. Zhou *et al*. [39] combined an edge detection method, the salient gradient map, and the radial basis functions (RBF) into a unified network architecture to identify whether there is seat belt present in the image or not. Guo *et al*. [11] similarly utilized an edge detection algorithm to detect seat belt from traffic surveillance cameras. Zhou *et al*. [38] used AlexNet [17]

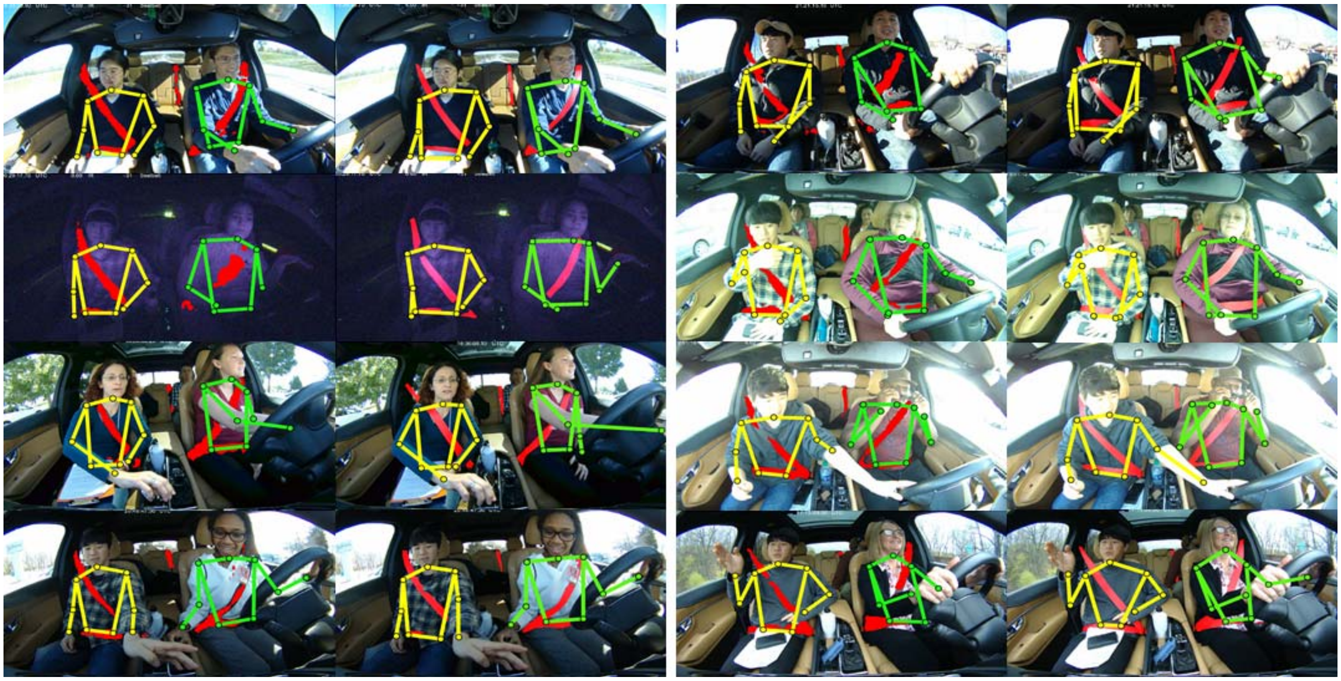

Figure 1: First and third columns in the figure are results produced by NADS-Net and second and fourth columns are the corresponding ground truth annotations.

with batch normalization [15] to identify seat belts. Elihos *et al*. [9] proposed a method that crops passenger regions first using the single shot detector (SSD) [20] and applies a CNN to detect seat belt non-use. The seat belt detection algorithm proposed in this paper attempts to add more granularity in detection results such that, in the future, the detection results can provide information on, not only use and non-use of seat belt, but also proper/improper use cases judged by the relative position of seat belt to the detected body position.

3. Method

In this paper, we propose new NADS-Net architecture for simultaneous pose estimation and seat belt detection. More specifically, the main objectives are (1) to estimate 2D body posture of the driver and (if exists) the front-row passenger; and (2) to segment image pixels that correspond to seat belts.

3.1. Problem overview

The driver and passenger pose estimation problem is similar to the generic 2D human pose estimation in the wild problem, in a sense that we aim to detect body key points and skeletons parsing those key points. However, there are several key differences between the driver/passenger pose estimation problem and the generic pose estimation problems as described below.

Most of the pose estimation models are trained and validated on publicly available data set such as MS COCO [19] and PoseTrack [2] data sets. These data sets are, however, mostly images taken in daytime or bright indoor scenes, whereas the illumination in vehicles can vary drastically. Furthermore, in generic data sets, there is no nighttime infrared (IR) image, hence the performance of a model trained on generic data sets is questionable in IVMS settings. This will be justified later in this paper.

On the other hand, generic data sets contain a variety of human poses whereas poses of drivers and passengers in vehicles are quite limited. Moreover, background texture and the number of people in the generic data sets are more diverse and the pixel-height of the people can also vary largely. In contrast, those quantities vary only narrowly in vehicle environments. From this observation, we hypothesize that a shallower model with lesser parameters would suffice the pose estimation task in IVMS settings. Hence, a higher computational efficiency can be achieved by reducing the neural network architecture without compromising the model performance.

3.2. Data set

One of the main challenges in this study was the lack of appropriate data sets. As noted above, there are many publicly available data sets for more generic human pose estimation problems, but they are not quite suited for in-vehicle monitoring purposes. Especially, we require seat belt annotations, diverse demographics, nighttime IR images, people under dynamic change of illumination as they drive,

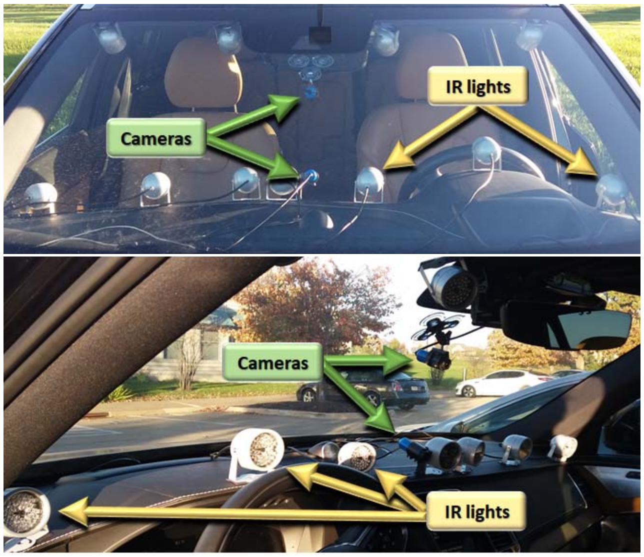

Figure 2: Data collection setup for this study.

and human poses and gestures in the context of driving.

Data collection We collected videos of drivers and passengers in a Volvo XC90 research vehicle through on-road driving studies. Over 7 months ranging from Spring to Winter, the total of 100 subjects consented to participate in the study in compliance to the internal review board (IRB) requirements. The subjects were randomly assigned in one of driving sessions that varied in season, weather, and time of the day. Each driving session was composed of static sessions while vehicle was at park where subjects were instructed to pose a specific set of predefined gestures, and on-road driving sessions. During the static gesture sessions, the subjects were requested to perform certain gestures and motions such as drinking, using smart phones, exercising, yawning, sneezing, leaning on the door, putting hands out the window, searching floor and the center console, adjusting sun visor, and etc. For the safety reasons, no request to perform a gesture or motion was presented to the subjects during the on-road driving sessions.

For data collection, we equipped the research vehicle with IR lights and two cameras. One of the cameras was mounted below the rear view mirror and another was above the center media console. IR lights were installed on the dashboard and behind the sun visors. Figure 2 shows how the vehicle was instrumented.

Statistics In addition to the driving videos, we also collected demographics information of each subject such as age, sex, and race through a survey questionnaire. Additionally, researchers in this study have manually annotated videos to label clothing and accessory types. These are summarized in Table 1.

Sex		Race				
Men	Women	White	Black	Asian	Hisp.	N/A*
53	47	67	11	12	5	5
Age (yr.)						
10-19	20-29	30-39	40-49	50-59	60-69	70-
4	28	17	16	15	17	3

*Not responded.

Eyewear			Facial Hair	
None	Glasses	Sunglasses	None	Beard
47%	33%	20%	87%	13%
Clothing (Top)			Accessories	
Short Sleeves	Long Sleeves	Jacket/Coat	Scarf	Hat
10%	55%	35%	20%	18%

Table 1: Subject statistics.

It should be noted that all driving sessions were accompanied by a research staff as a safety protocol and, thus, the videos contain some repeated appearances of a few research staffs. To minimize the potential bias in the data, the researchers rotated the duty across the driving sessions. By the safety requirement, the researchers had to sit on the front passenger seat when the vehicle was in motion, but while the vehicle was at park, they moved around to different seat positions as much as possible to minimize the data bias. Moreover, researchers were asked to wear different clothing and accessories each time.

Lastly, the route of driving included a good mixture of rural roads, urban areas, and highways to diversify background and illumination.

Data annotation For each session, short video clips were selected manually and diversity in terms of subject demographics, illumination, and pose was promoted. The annotation process was done manually by human annotators. For each image, the annotators were instructed to segment all visible seat belts with a binary mask and to mark x and y pixel coordinates of body key points, as displayed in Figure 1, following the common convention in other publicly available pose estimation data sets such as MS COCO. We did not track the lower extremities as they were not visible in most of the frames. The researchers of this study conducted a final check each time the annotators submitted the job in order to assure the data quality. Annotation errors were fixed by the researchers or sent back to the annotators for rework. Sample annotation results are presented in Figure 1.

3.3. Model

Our algorithm has three heads that generate key point heat maps, PAF heat maps, and a binary seat belt detection mask,

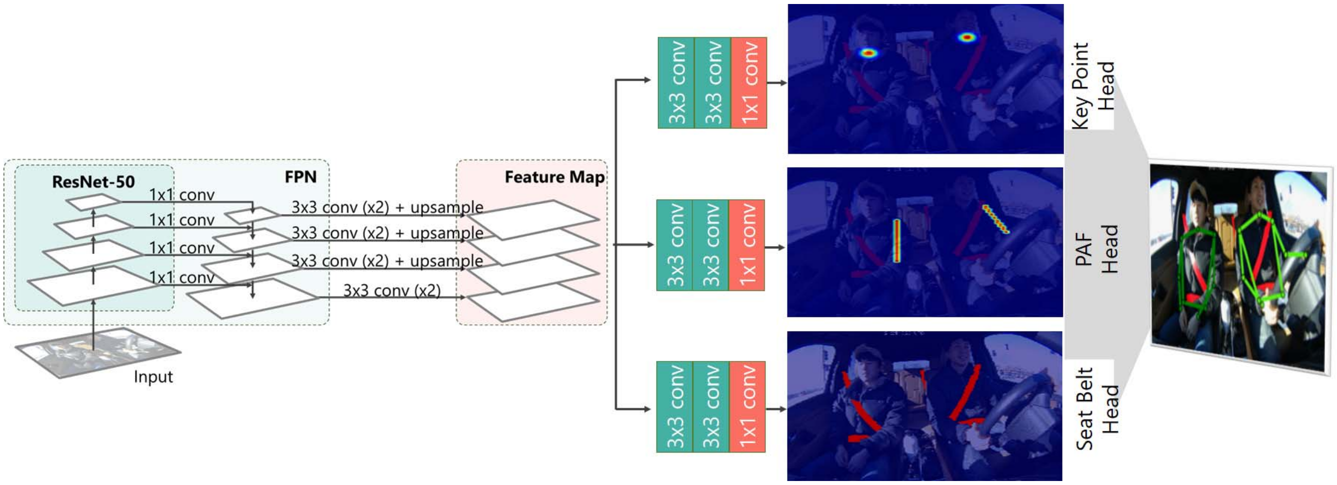

Figure 3: NADS-Net architecture.

sitting on top of the FPN backbone (Figure 3). Outputs from the first two heads are used to parse the key point instances into human skeletons. For the parsing, we employ the same PAF mechanism with the bipartite graph matching as proposed in Cao *et al*. [4]. However, our contribution is the architecture that generates feature maps faster and more efficient. More specifically, we reduced the six-stage architecture of [4] to a single stage architecture. Instead, we replaced the VGG [32] backbone with a strong multi-scale FPN backbone to speed up the inference time and to compensate the reduced staging.

The FPN backbone of NADS-Net is comprised of ResNet-50 and produces a rudimentary feature pyramid for the later detection branches. The inherent structure of ResNet can produce multi-resolution feature maps after each residual blocks, namely C2, C3, C4, and C5, which are sized $1/4$, $1/8$, $1/16$, and $1/32$ of the original input resolution, respectively. For example, for a given 384×384 image input we use in the NADS-Net implementation, the ResNet-50 backbone produces four levels of feature pyramid, each sized 96×96, 48×48, and 24×24 and 12×12. Along such, the number of channels (feature maps) increases from 256 (C2) to 512 (C3), 1,024 (C4), and 2,048 (C5). These are then further convolved with 1×1 convolutions, to compress the number of channels to 256. Lastly, the reduced feature pyramid further undergoes two more 3×3 convolutions and an upsampling to produce a concatenated $96 \times 96 \times 512$ feature map.

Each of the detection branches employs two 3×3 convolutions and a 1×1 convolution to predict a pixel-wise probability distribution. For the key point head, the pixel-wise probability indicates the probability of the corresponding pixel being a certain joint. Since we are interested in detecting joints with background, the key point head produces ten such probability maps of the size 96×96, each of which corresponds to one of the nine joints we are interested in detecting and also background. For the PAF head, similar to [4], we produce vector fields of size 96×96 which encode pairwise relationships between body joints. Lastly, the seat belt head produces a probability distribution map of a size 96×96 indicating the likelihood of each pixel being a seat belt. Each pixel-wise probability is then thresholded to generate a binary seat belt detection mask.

3.4. Implementation

The proposed NADS-Net was implemented in Keras [7] with TensorFlow [1] backend and an NVIDIA GeForce GTX 1080 Ti GPU was used for training and testing. For the training data, 30 driving sessions out of total 50 were used. The rest were reserved for testing. When selecting 30 sessions of the training data, we manually selected half of the nighttime sessions to distribute nighttime IR images equally for both training and testing data. Rest of the training data sets were selected randomly. At the end, 10,550 images were used for training and 7,721 images were used for testing.

We pre-trained the model with MS COCO `train2017` data set and the corresponding human key point annotations. Only the body key point branch and PAF branch were pre-trained. As reported in the result section, the transfer learning strategy provided a significant performance gain. Additionally, random image augmentations were applied to training images, such as rotation, scaling and vertical flipping.

For the final parsing of the skeleton, we strictly followed the implementation of Cao *et al*. [4]. That is, non-maximum suppression was used on the detection confidence maps which allowed the algorithm to obtain a discrete set of part candidate locations. Then, bipartite graph was used to group each person. More details are deferred to [4].

4. Result and discussion

We compare the NADS-Net with the PAF model [4] as a baseline. For the detection accuracy of the body key points, we employ the probability of correct key point (PCK) metric [36] as a criterion. In typical generic human pose estimation applications, the head size of the person being estimated is used as a reference of PCK to determine the tolerance of correctness (PCKh) [3]. This is reasonable for generic applications where the pixel heights of people vary drastically within and across images. However, for the specific in-vehicle monitoring task presented in this paper, we find such a generic way may prevent precise characterization of model performance as the head size can greatly vary depending on the spatial position of the head, while the distance from the camera to the rest of the body (for example, hands on the steering wheel) remains unchanged. To this end, we may benefit by using the headrest size as the reference of the PCK measure instead. The reason can be that, first of all, the distance from the camera to the headrest is almost the same, which allows more stable reference for PCK evaluation. Furthermore, the headrest is about the same size as the human heads, resulting the similar range of PCK values as other human pose estimation literature. This enables more intuitive interpretation of the analysis results. Therefore, we use a modified PCKh metric (mPCKh) where the diagonal length of the headrest is used as the reference (Figure 4). It is worthwhile to note that, although the camera was fixed at the same position and angle across the entire data collection sessions, the pixel size of the headrest might change due to the seat position. However, with respect to the average diagonal length of the headrest (170 pixels), the variation is negligible (less than 10 pixels).

For the seat belt detection task, there is no baseline model available to compare. Instead, we simply report our model's sensitivity, specificity, precision, F1 score, and intersection over union (IOU). As we will discuss below, these are arguably inappropriate ways to characterize the seat belt detection accuracy, but we defer the development of a better metric to our future work.

4.1. Efficiency

In terms of the total number of parameters, the baseline model requires 52,311,446 parameters while NADS-Net uses 39,334,301 parameters, which is about 25% lesser despite the fact that there is an additional seat belt segmentation head. Given that the skeleton parsing method in NADS-Net is exactly the same as in [4], the run-time difference is only a function of the model inference time. The frame rate on an Intel® Core™ i7-7800X 3.50 GHz CPU machine with a 32GB RAM and a NVIDIA GeForce 1080Ti GPU, the benchmark model demonstrated 12 fps while NADS-Net showed 18 fps in average. This shall not be directly interpreted as a definitive speed comparison at their maximal performance but, in fact, further optimization may dramatically change the frame rate. However, the result still serves as a weak but convincing evidence that the NADS-Net performs more efficiently than the baseline model.

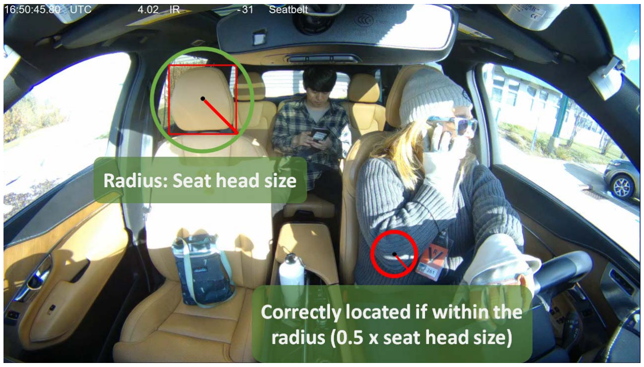
16:50:45.80 UTC
4.02 IR
-31
Seatbelt
Radius: Seat head size
Correctly located if within the radius (0.5 x seat head size)

Figure 4: Modified PCKh metric (mPCKh) used for key point evaluation.

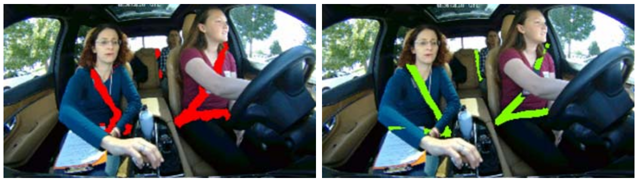

Figure 5: Seat belt detection result (*left*) in comparison with the ground truth (*right*). With human inspection, the prediction result is of good quality as it correctly marks the seat belt area. However, IOU for this particular example was 46%, justifying the need for a better evaluation metric.

4.2. Key point detection

We compared mPCKh scores of NADS-Net model and the baseline model for each individual key point location. As reported at the bottom of Table 2, the baseline model scored the average mPCKh of 82% over all key points while NADS-Net model scored 84%. Unlike the baseline model, NADS-Net does not have refining stages and have fewer parameters, but PCK-wise, shows a similar or slightly better performance. This demonstrates that the multi-resolution feature pyramid produced by the FPN backbone is enough for the driver/passenger pose estimation task and can replace multiple refining stages of generic pose estimation algorithms. Some qualitative results are given in Figure 1.

4.3. Seat belt detection

The seat belt detection head produces a probability density function defined over the image domain, indicating the likelihood of a pixel being a seat belt. The probability distribution is then thresholded to obtain a binary seat belt segmentation mask (see Figure 1). We evaluate the quality of seat

	Cao *et al.* [4]	NADS-Net (Ours)		
	MC*	MC	ND**	All***
Drivers	80%	81%	75%	88%
Front Passengers	85%	89%	78%	90%
Men	84%	87%	79%	90%
Women	79%	81%	73%	88%
White	83%	86%	77%	89%
Black	75%	77%	70%	87%
Asian	85%	87%	80%	91%
Glasses	83%	83%	75%	89%
Sunglasses	78%	82%	72%	85%
Short Sleeves	84%	85%	77%	89%
Long Sleeves	85%	87%	78%	90%
Jacket/Coat	79%	81%	75%	88%
Scarf	82%	84%	78%	91%
Hat	82%	84%	77%	90%
Beard	82%	87%	81%	90%
Daytime	85%	86%	77%	89%
Nighttime	74%	77%	75%	88%
Overall	**82%**	**84%**	**77%**	**89%**

* Trained with MS-COCO data set only.
** Trained with new driving data set only.
*** Trained with both data sets combined.

Table 2: Accuracy evaluation with mPCKh@0.5.

belt segmentation using five metrics: sensitivity, specificity, precision, F1 score, and IOU as reported in Table 3.

The high specificity of the model indicates that the model can correctly classify non-seat belt pixels with a high confidence. However, the other metrics are poor, where the sensitivity, precision, and F1-score were 63.51%, 63.58%, and 63.55%, respectively, and IOU was only 47%. A possible interpretation of this result is that, first of all, NADS-Net model is highly conservative and predicts seat-belt only when there is a high certainty. This can be justified from the strong contrast between the sensitivity and specificity. Furthermore, even if the seat belt detection is correct, just because the predicted seat belt is thinner than the actual ground truth annotation, metrics such as sensitivity and IOU drops significantly as they penalize the thin subset of seat belt that are not detected. Lastly, we also noticed that the manual annotation of seat belt contained a few inconsistencies, which we could not resolve in this study. For example, there was an inconsistency among the annotators where some people discerned the seat belt buckles as a part of the seat belt while the others exclusively labeled the fabric part of the seat belt. These are possible sources of low sensitivity, precision, F1-score, and IOU and need to be addressed in future work.

However, more fundamentally, it is worthwhile to note the lack of suitable evaluation metrics. We inspected the seat belt segmentation results image by image and noticed that most of the error indeed comes from seat belt predicted thinner than the ground truth annotation (*e.g.* Figure 5). A possible solution to this is to skeletonize the seat belt mask and compare the distances between the curves. Another potential solution is a metric such as optimal transport [29]. These will also be the potential venues for future study.

Sensitivity	Specificity	Precision	F1-Score	IOU
63.51%	99.28%	63.58%	63.55%	46.57%

Table 3: Seatbelt Evaluation

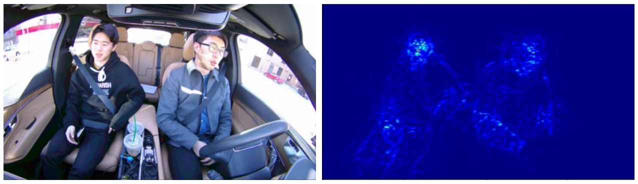

Figure 6: Saliency map visualization [31] for the 'right wrist' class. Note visual cues from the face have significant contribution to the prediction.

4.4. Appearances

Performance on different demographics In Table 2, evaluation of the model performance on different demographic parameters is reported. For women, the overall performance was lower than men for all four experiments—the baseline model trained only on MS COCO data set; NADS-Net model trained only on MS COCO data set; NADS-Net model trained only on the new driving data set; and NADS-Net model pretrained on MS COCO data set and then transferred to the new driving data set. Considering the fact that women-to-men ratio was 1:1, one hypothesis we can derive from this observation is that the appearance variance among women is larger than men, because of larger diversities in hairstyle, accessories, clothing patterns, etc. among female populations. Therefore, it is more advisable to include more female subjects in data collection in the future.

Race-wise, the model performance was slightly better on Asian populations followed by white populations. Performance on people with darker skin tone was noticeably lower, reconfirming the bias of computer vision data sets and algorithms as pointed out by Zou and Schiebinger [40]. We believe the new driving data set collected in this study also can suffer from the same bias. The primary reason was the geographical location where the study was conducted, whose population was predominantly white. Furthermore, coincidentally 60% of black subjects participated in the study during the nighttime, which could also worsen the performance evaluation on this demographic group. The future work, therefore, needs to include more subjects with darker skin tones, to overcome the bias in performance and a more rigorous and controlled analysis.

Clothing/accessories Table 2 reports the model performance on nine different clothing/accessory categories. The numbers reveal that there is no significant influence of clothing/accessories, except for sunglasses. We failed to reach a convincing explanation on this, but a weak hypothesis on this might be that many of the visual cues to a CNN-based pose estimation model come from the facial area. As demonstrated in the saliency map visualization in Figure 6, even for the detection of right wrist, for example, which is relatively far from the face, we can notice the detection relies largely on the visual cues coming from the face, not just the wrist and the arm area. Although this is beyond the scope of this paper, it should be worth investigating this in future work, to deepen our understanding of how pose estimation models perceive and process the visual cues.

Another noticeable fact was that the model performance was poor on people wearing jackets/coats, when the model was trained on MS COCO data set only, but the performance improved significantly when transferred to the new driving data set. This might suggest that MS COCO data set is biased to lighter clothing but confirming this hypothesis by examining the data set image by image should be beyond the objective of this study.

Illumination We could observe a significant drop of performance in nighttime data when the model was trained on MS COCO data set only. This could be easily improved by transferring the model to the new driving data set. We can conclude that there is not much of fundamental differences between daytime images and nighttime IR images in terms of visual cues available for the detection of body parts. Instead, the drop of performance mostly comes from the lack of nighttime data in MS COCO data set, not an inherent deficiency of CNNs.

5. Conclusion and future work

In this paper, we proposed a new CNN architecture called NADS-Net for the purpose of driver and passenger pose estimation and seat belt detection in vehicles. NADS-Net is capable of estimating body pose together with seat belt segmentation with the similar accuracy than the state of the art baseline [4], while requiring fewer parameters and demonstrating a faster inference time. We broke the performance down and provided in-depth analyses in different aspects, including sex, race, clothing, and illumination. These results may provide practical insights to future academic research and to industrial product development.

For the future work, one of the clear challenges is the bias of data set. One trivial solution could be to enlarge the scale of data collection study by including more diverse group of subjects and other imaging parameters. However, practically, this may not be viable in many aspects. To this end, one potential direction we are exploring towards is the creation of a synthetic data set and randomizing imaging conditions virtually. In addition, the current status of our work is limited to a frame-by-frame detection, while, arguably, it is more desirable to take temporal aspects (*e.g.* optical flow) into account as in recent works such as [26].

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] M. Andriluka, U. Iqbal, A. Milan, E. Insafutdinov, L. Pishchulin, J. Gall, and B. Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2017. 3

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 6

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 1, 2, 5, 6, 7, 8

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7103–7112, 2018. 2

[6] I.-H. Choi, T. B. H. Tran, and Y.-G. Kim. Real-time categorization of driver's gaze zone and head pose using the convolutional neural network. In *HCK'16 Proceedings of HCI Korea*, pages 417–422, 2016. 1

[7] F. Chollet et al. Keras. https://keras.io, 2015. Accessed: 2019-08-20. 5

[8] L. Dawson and N. Jenkins. Fatal intra-abdominal injury associated with incorrect use of a seat belt. In *Emergency Medicine Journal, 15*, pages 437–438, 1998. 1

[9] A. Elihos, B. Alkan, B. Balci, and Y. Artan. Comparison of image classification and object detection for passenger seat belt violation detection using NIR & RGB surveillance camera images. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 3

[10] A. Eriksson and N. A. Stanton. Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. In *Human Factors: The Journal of the Human Factors and Ergonomics Society, 59:4*, pages 689–705, 2017. 1

[11] H. Guo, H. Lin, S. Zhang, and S. Li. Image-based seat belt detection. In *Proceedings of 2011 IEEE International Conference on Vehicular Electronics and Safety*, 2011. 2

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[14] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV2016*, pages 34–50, 2016. 2

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015. 3

[16] S. G. Klauer, V. L. Neale, T. A. Dingus, D. Ramsey, and J. Sudweeks. Driver inattention: A contributing factor to crashes and near-crashes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 49, no. 22*, pages 1922–1926, 2005. 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90, 2012. 2

[18] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016. 1, 2

[19] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 3

[21] R. O. Mbouna, S. G. Kong, and M.-G. Chun. Visual analysis of eye state and head pose for driver alertness monitoring. In *IEEE Transactions on Intelligent Transportation Systems, Vol. 14, no. 3*, pages 1462–1469, 2013. 1

[22] National Highway Traffic Safety Administration. USDOT releases 2016 fatal traffic crash data, 2017. 1

[23] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2016. 2

[24] K. Okuno, T. Yamashita, H. Fukui, S. Noridomi, K. Arata, Y. Yamauchi, and H. Fujiyoshi. Body posture and face orientation estimation by convolutional network with heterogeneous learning. In *2018 International Workshop on Advanced Image Technology (IWAIT)*, pages 1–4, 2018. 1, 2

[25] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3711–3719, 2017. 2

[26] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[27] M. Peden. Global collaboration on road traffic injury prevention. In *International Journal of Injury Control and Safety Promotion, 12:2*, pages 85–91, 2005. 1

[28] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2015. 2

[29] J. Rabin, G. Peyré, and L. D. Cohen. Geodesic shape retrieval via optimal mass transport. In *European Conference on Computer Vision*, pages 771–784. Springer, 2010. 7

[30] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[31] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 7

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[33] D. d. Waard, M. v. d. Hulst, M. Hoedemaeker, and K. A. Brookhuis. Driver behavior in an emergency situation in the automated highway system. In *Transportation Human Factors, 1:1*, pages 67–82, 2010. 1

[34] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 2

[35] C. Yan, F. Coenen, and B. Zhang. Driving posture recognition by convolutional neural networks. In *IET Computer Vision, 10:2*, pages 103–114, 2016. 1

[36] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2012. 6

[37] K. Yuen and M. M. Trivedi. Looking at hands in autonomous vehicles: A convnet approach using part affinity fields. *ArXiv*, abs/1804.01176, 2018. 1, 2

[38] B. Zhou, D. Chen, and X. Wang. Seat belt detection using convolutional neural network BN-Alexnet. In *Interlational Converence on Intelligent Computing. ICIC*, pages 384–395, 2017. 2

[39] B. Zhou, L. Chen, J. Tian, and Z. Peng. Learning-based seat belt detection in image using salient gradient. In *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 547–550, 2017. 2

[40] J. Zou and L. Schiebinger. AI can be sexist and racist – it's time to make it fair. *Nature*, 559:324–326, 2018. 7