Driver's postural monitoring using different data-driven approaches

Mingming Zhao

N°d'ordre NNT : 2021LYSE1241

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

Ecole Doctorale NED162

(Mécanique, Énergétique, Génie Civil, Acoustique (MEGA))

et en cotutelle avec

l'Université Tongji - Chine

Spécialité de doctorat : Biomécanique et ingénierie automobile

Discipline : Génie Mécanique

Soutenue publiquement le 10/11/2021, par :

Mingming Zhao

Monitoring de la posture du conducteur par des approches basées sur des données

Devant le jury composé de :

Nom, prénom	grade/qualité, établissement/entreprise	
Président. M. Zhou, Su		
M. Multon, Franck	Professeur, Université Renne2	Rapporteur
M. Niu, Jianwei	Professeur, USTB (Chine)	Rapporteur
Mme. Cheze, Laurence	Professeure, UCBL	Examinatrice
Mme. Nie, Bingbing	Prof. Associée, U. Tsinghua (Chine)	Examinatrice
M. Zhou, Su	Professeur, Université Tongji	Président
M. Wang, Xuguang	Directeur de Recherche, Univ-Eiffel	Directeur
Mme. Wang, Hongyan	Professeure, Université Tongji	Co-Directrice
M. Beurier, Georges	Chargé de Recherche, Univ-Eiffel	Co-Directeur
M. Wu, Xian	Prof. Associé, Université Tongji (Chine)	Examinateur
M. Crispim-Junior, Carlos F.	Maître de Conférences, Université Lyon 2	Examinateur

TITRE : Monitoring de la posture du conducteur par des approches basées sur des données

Avec l'automatisation de la conduite, les conducteurs seront libérés des contraintes liées à la conduite et pourront effectuer de nouvelles activités comme lire un livre, utiliser un smartphone, travailler sur un ordinateur, dormir, etc. Pour un véhicule conditionnellement automatisé du niveau 3 selon la SAE (Society of Automotive Engineers), le conducteur doit être prêt à prendre le contrôle du véhicule en cas de besoin. Le monitoring postural du conducteur peut fournir non seulement sa posture dans le véhicule, mais aussi des informations pour évaluer son état cognitif et attentionnel. Ces informations sont nécessaires pour le développement de systèmes de protection et d'aide à la conduite pour une meilleure sécurité. Le monitoring du conducteur, notamment celui de la posture du corps par des capteurs non-invasifs tels que caméra de profondeur, capteur de pression, est un champ de recherche très actif ces dernières années. Divers systèmes de monitoring ont été proposés pour extraire des informations posturales du conducteur (par exemple, l'orientation de la tête, la position de la main, etc.) dans le but de reconnaître différentes activités du conducteur. Les méthodes existantes souffrent du problème du placement sous-optimal de caméra dans un habitacle restreint et d'occlusions corporelles dans le champ de vision. La plupart des études ont été consacrées à la surveillance des membres supérieurs, du tronc et de la tête du conducteur, et peu ont examiné la posture du corps entier. Pour faciliter la reconnaissance des activités des conducteurs, quelques bases de données sur la posture des conducteurs ont été proposés. Cependant, il y a un manque au niveau de méthodes efficaces pour générer des étiquettes d'annotation de vérité terrain qui sont indispensables pour les algorithmes d'apprentissage supervisé. A ce jour aucun dispositif de surveillance du conducteur n'a vraiment fait la démonstration de son efficacité.

La présente thèse vise donc (1) à créer une base de données labélisés correspondantes à un large panel d'activités liées non seulement à la conduite mais surtout à des activités autres que la conduite et (2) développer un système de monitoring postural du corps entier dans le véhicule à l'aide de caméras de profondeur et de capteurs de pression.

Dans ce travail, nous avons construit une large base de données labélisés sur les postures de conducteurs grâce à une procédure d'augmentation de données que nous avons mise au point. Les données d'origine, mesurées par 3 caméras de profondeur et 2 nappes de pression, ont été collectées sur une maquette en laboratoire auprès de 23 conducteurs effectuant 42 activités incluant aussi des tâches non-liées à la conduite. Les centres articulaires ont été reconstruits

grâce aux marqueurs posés au corps mesurés par un système de capture de mouvement. La procédure d'augmentation de données, basé sur des techniques d'infographie, permet la génération automatique d'images synthétisées avec des annotations 2D et 3D. Grâce à cette base de données, nous avons adapté plusieurs algorithmes d'apprentissage pour la reconnaissance de la posture des membres supérieurs, inférieurs et de la tête à partir des mesures par des capteurs de pression, des caméras de profondeur ou des deux. Une attention particulière a été accordée sur la sélection des paramètres pertinents quand les mesures de pression sont utilisées comme entrées, et sur la réduction des erreurs de la posture du haut du corps causé par des occlusions corporelles ou aux confusions de l'algorithme quand une caméra de profondeur est utilisée.

TITLE: Driver's postural monitoring using different data-driven approaches

With driving automation, drivers will be freed from the constraints of driving and will be able to perform new activities such as reading a book, playing with a smartphone, working on a computer, sleeping, etc. For a conditionally automated vehicle of SAE (Society of Automotive Engineers) Level 3, the driver must be prepared to take control of the vehicle when needed. Postural monitoring the driver's posture can provide not only body position in the vehicle, but also information to assess driver's cognitive and attentional state. Postural information is necessary in the development of protection and driving assistance systems for better safety. Driver monitoring, in particular that of body posture by non-invasive sensors such as depth camera, pressure sensor, has been a very active field of research in recent years. Various monitoring systems have been proposed for extracting driver postural information (for example, orientation of the head, position of the hands, etc.) in order to recognize different activities of the driver. Existing methods suffer from the problem of suboptimal placement of camera in the vehicle and body occlusions in the field of view. Most studies have been devoted to monitoring the driver's upper limbs, trunk, and head, and few have examined full-body posture. To facilitate the recognition of driver activities, a few databases on driver posture have been proposed. However, there is a lack in efficient methods for generating ground truth annotation labels which are essential for training posture estimation algorithms based on supervised learning. To date, none driver posture monitoring system has really demonstrated its effectiveness.

The present thesis therefore aims to (1) create an annotated posture database corresponding to a wide range of activities that take place in both traditional vehicles and autonomous vehicles, (2) to develop a postural monitoring system for the entire body in the vehicle using depth cameras and pressure sensors.

In this work, we have built a large database on driver postures with annotation and labels through a data augmentation procedure that we have developed. The original data, measured by 3 depth cameras and 2 pressure sheets, were collected on a laboratory mockup from 23 drivers performing 42 activities including non-driving tasks. The joint centers were reconstructed using markers placed on the body measured by a motion capture system. The data augmentation procedure, based on computer graphics techniques, allows the automatic generation of images synthesized with 2D and 3D annotations. Using this database, we have adapted several learning algorithms to recognize upper, lower and head posture from measurements by pressure sensors, depth cameras, or both. Special attention was paid to the selection of relevant features when

using pressure measurements as inputs and to the reduction of upper body posture prediction errors caused by body occlusions or confusion when using a depth camera.

MOT-CLES
Monitoring, Posture, Conducteur, Automobile, Véhicule autonome, Intelligence artificielle, Traitement d'images, Fusion de capteurs
Monitoring, Posture, Driver, Automotive, Autonomous vehicle; Artificial intelligence, Image processing, Sensor fusion

INSTITUT ET ADRESSE DE L'U.F.R OU DU LABORATOIRE :

LBMC – Laboratoire de Biomécanique et Mécanique des Chocs (UMR – T 9406) (Univ Eiffel-UCBL)
25 Avenue François Mitterrand, Case 24,
69675 BRON France

School of Automotive Studies – Tongji University
4800 Cao'an road
201804 SHANGHAI China

Résumé substantiel

Les futurs véhicules autonomes (VAs) sont censés réduire considérablement le nombre d'accidents liés aux facteurs humains, mais les accidents sont inévitables en raison d'autres facteurs. Avec l'automatisation de conduite, les conducteurs seront libérés des contraintes liées à la conduite et pourront effectuer de nouvelles activités : lire un livre, jouer avec un smartphone, travailler sur un ordinateur, dormir, etc ... Pour un véhicule conditionnellement automatisé du niveau 3 selon SAE (Society of Automotive Engineers), le conducteur doit être prêt à prendre le contrôle du véhicule en cas de besoin. On a besoin de l'information posturale du conducteur, comme l'orientation de la tête/des yeux, la position des mains et des pieds, pour détecter l'inattention du conducteur. Le monitoring de la posture du conducteur peut fournir non seulement sa position dans le véhicule, mais aussi des informations pour évaluer son état cognitif et attentionnel. Ces informations sont nécessaires dans le développement de systèmes de protection et d'aide à la conduite pour une meilleure sécurité.

Le monitoring du conducteur, notamment celui de la posture du corps par des capteurs non-invasifs tels que caméra de profondeur, capteur de pression, est un champ de recherche très actif ces dernières années, impliquant notamment des chercheurs en vision par ordinateur et en intelligence artificielle. Divers systèmes de monitoring ont été proposés pour extraire des informations posturales du conducteur (par exemple, l'orientation de la tête, la position de la main, etc.) dans le but de reconnaître différentes activités du conducteur. Les méthodes existantes souffrent du problème du placement sous-optimal de la caméra dans le véhicule et d'occlusions corporelles dans le champ de vision. La plupart des études ont été consacrées à la surveillance des membres supérieurs, du tronc et de la tête du conducteur, et peu ont examiné la posture du corps entier. Pour faciliter la reconnaissance des activités des conducteurs, quelques bases de données sur la posture des conducteurs ont été proposés. Cependant, il y a un manque au niveau de méthodes efficaces pour générer des étiquettes d'annotation de vérité terrain qui sont indispensables pour les algorithmes d'apprentissage supervisé. A ce jour aucun dispositif de surveillance du conducteur n'a vraiment fait la démonstration de son efficacité.

La présente thèse vise donc (1) à créer une base de données labélisés correspondantes à un large panel d'activités liées non seulement à la conduite mais surtout à des activités autres que la conduite et (2) développer un système de monitoring postural du corps entier dans le véhicule à l'aide de caméras de profondeur et de capteurs de pression.

Les principales contributions de cette thèse sont résumées ci-après :

1) Proposition d’une procédure de génération de données sur les postures du conducteur et une base de données labélisées pour le développement d’algorithmes de monitoring postural. Une expérience a été réalisée dans le cadre du projet national ANR AutoConduct pour collecter des données sur les postures du conducteur auprès de 23 conducteurs ayant une large variation en taille et en poids de la population d’adultes Français. Chaque participant a effectué 42 activités différentes en lien avec la conduite et non-conduite. Trois caméras de profondeur (deux pour le haut, une pour le bas du corps) et de deux nappes de pression sur le siège du conducteur ont été utilisés. Les mouvements du corps ont été enregistrés par un système de capture de mouvement optique et reconstruits à l'aide du logiciel RPx, un outil de reconstruction, d'analyse et de simulation de mouvement développé au LBMC à l'Université Gustave Eiffel (anciennement IFSTTAR). Les mesures de différents capteurs ont été synchronisées et fusionnées dans un même système de coordonnées. Grâce aux techniques d'infographie, les mouvements réels du conducteur ont pu être appliqués aux humains virtuels créés par MakeHuman, un outil open source pour le prototypage de modèles humains en 3D. Le logiciel MAYA, a été utilisé pour créer des séquences animées avec les humains virtuels dans un véhicule. La procédure d'augmentation de données a permis la génération des postures en 2D et 3D avec annotation à grande échelle. Enfin, les postures à l’intérieur d’un véhicule ainsi crées forme la base de données ‘AutoConduct’, qui comprend environ 130 000 trames de images et pression de contact directement issues de l’expérimentation avec des postures de vérité terrain en 3D, et aussi environ 12 millions trames d'images synthétisées avec la segmentation de parties du corps 2D et la localisation 3D des centres articulaires. En comparant aux bases de données existantes, la nôtre présente des annotations plus précises, une couverture du corps plus complète et une plage de variation plus grande de postures à l’intérieur d’un véhicule.

2) Identification des paramètres pertinents et proposition de méthodes pour la classification de posture à partir des mesures de pression de contact. En analysant la variation de pression pendant les mouvements du corps, cinq postures du tronc, deux positions du pied gauche et trois positions du pied droit ont été identifiées pour la classification. Pour extraire des caractéristiques pertinentes à partir de la distribution de pression, 12 et 8 zones de contact sur le dossier et l'assise sont définies respectivement. Environ deux cents paramètres sont extraits incluant le centre de pression, l’aire de contact, les ratios de la somme des pressions entre différentes zones ainsi que les valeurs relatives à une posture de référence. Ces paramètres sont classés en fonction de leur importance pour la classification des postures du tronc, des positions des pieds gauche et droit en utilisant la méthode d'estimation d’erreur Random Forest Out-Of-Bag (OOB). Les résultats montrent que seuls quelques paramètres de pression sont pertinents

pour discriminer les postures. Le meilleur model que nous avons obtenu permet une précision de classification de 0.91 (score F1), 0.93 et 0.74 en moyenne respectivement pour les postures du tronc, du pied gauche et du pied droit par validation croisée.

3) Proposition d'une méthode de détection plus précise de la posture du haut du corps basé sur une caméra de profondeur. Un réseau de neurones convolutifs (CNN) est adapté à partir du logiciel OpenPose (une méthode d'estimation de pose 2D en temps réel utilisant des champs d'affinité de la carte) pour localiser des cartes de confiance de sept parties des membres supérieurs (cou, épaule gauche/droite, coude gauche/droit, poignet gauche/droit). Les cartes de confiance sont ensuite utilisées par un modèle de régression de décalage que nous avons mis au point pour estimer la position 3D des sept articulations. Afin de réduire les erreurs de location des centres articulaires dues à des occlusions corporelles et à la confusion, la base de données AutoConduct est organisée à l'aide de la technique Filtered Pose Graph. Au cours d'un mouvement, la mesure de fiabilité de chaque articulation est utilisée pour sélectionner des postures similaires à partir de la base des données. Ces postures similaires sont ensuite utilisées pour corriger les postures ayant une mauvaise fiabilité. Enfin, les postures prédites sont lissées à l'aide d'un filtre de Kalman. Après correction, la proportion des échantillons ayant un écart de moins de 50 mm en moyenne des sept articulations est passé de 74% à 91%.

4) Proposition d'une méthode d'estimation de la posture de la tête utilisant la régression random forest (RFR). A partir des cinq points clés de la tête (oreille gauche/droite, œil gauche/droit, pointe du nez) détectés par le réseau neuronal convolutif que nous avons adapté du logiciel OpenPose, la régression Random Forest est proposée pour estimer l'orientation et la position de la tête. Les prédictions par la méthode RFR sont comparées avec celle basée sur la correspondance de corps rigide (Rigid Body Matching – RBM). La comparaison montre que la méthode RFR donne une meilleure estimation de la posture de la tête que la méthode RBM. Par la méthode RFR, les erreurs moyennes équilibrées sont inférieures à 11° et 2 cm respectivement pour l'orientation et la position de la tête dans 96.3% des cas.

5) Exploration d'une méthode de reconnaissance de la posture des membres inférieurs à partir d'une caméra de profondeur. D'abord, un algorithme de décalage moyen est utilisé pour extraire les points clés des jambes à partir du nuage de points dans l'espace pour les jambes. Ensuite, les points clés pertinents après une évaluation sont utilisées pour entraîner les classificateurs du random forest pour prédire la position des pieds. Par la validation croisée, la précision de prédiction (F1 score) des trois positions du pied droit est de 0.88 en moyenne, bien meilleure que celle par les capteurs de pression (0.74).

6) Exploration de la fusion de capteurs pour une meilleure prédiction de la posture. En termes de correction de la posture du haut du corps mesurée par une caméra de profondeur, l'utilisation de la classe de posture du tronc prédite basées sur les capteurs de pression comme connaissance préalable améliore légèrement la performation de prédiction. Avec la fusion de l'information provenant des capteurs de pression, le pourcentage des cas testés ayant une erreur de moins de 50 mm pour les sept articulations du haut du corps, est passé de 91% à 93%. En ce qui concerne la prédiction de la position des pieds, différentes méthodes de fusion entre les capteurs de pression et la caméra de profondeur ont été explorées. Les résultats montrent que la méthode de fusion au niveau de la décision surpasse la méthode reposant sur un seul capteur, soit la caméra de profondeur, soit les capteurs de pression.

Le mémoire est composé de 7 chapitres. Après une introduction générale (Ch1) et une revue bibliographique (Ch2) sur les systèmes de monitoring postural du conducteur et les bases de données, les contributions du présent travail sont organisées en cinq chapitres. La procédure de génération de données sur les postures du conducteur ainsi que la base de données ainsi générées (Ch3) sont d'abord décrites. Ensuite, cette base de données est utilisée pour développer et évaluer des méthodes de reconnaissance de postures du haut du corps (Ch4), de la tête (Ch5) et des membres inférieurs (Ch6). Enfin, le dernier chapitre (Ch7) résume les principaux résultats, les limitations ainsi que les perspectives pour la future recherche.

Acknowledgement

First of all, I would like to express my sincere gratitude to my foreign supervisor Dr. Xuguang Wang for suggestion of the topic and his patient guidance, encouragement, useful hints and discussions throughout the work. His kindness, generosity, enthusiasm for research and clarity of thought have been a great example for me.

I am profoundly grateful to my domestic supervisor Prof. Hongyan Wang for many years of guidance and support. She has been one of the most influential people in my academic career, responsible for kindling my interest in research. She helped me with my first steps in graduate life, and exposed me to the subjects of computer vision and machine learning.

I want to sincerely thank my co-supervisor Dr. Georges Beurier, a sensor guru and computer science wizard, for his support and contribution to the experiment and data processing of this thesis.

I am grateful to Philippe Beillas and Christine Buisson for their useful suggestions on the annual work progress meeting CST, as well as the anonymous reviewers for their constructive comments on my papers.

I would like to thank Ilias Theodorakos, Cyrille Grebonval, Richard Roussillon, Fabien Moreau for their technical support and contribution to the experiment.

I feel privileged to be a member of our research group at LBMC. I am greatly thankful to David Mitton and all the other colleagues who have helped me one way or another.

I would like to thank many friends from my three years at Université Gustave Eiffel, who did not have direct impact on this thesis, but made my time here really memorable.

Table of Contents

Chapter 1

1 Introduction

Over 1.35 million fatalities and 50 million serious injuries worldwide each year are claimed by road traffic accidents, most of which have been reportedly attributed to human driver errors in terms of recognition, decision and performance (Beanland et al. 2013; Singh 2015; Née et al. 2019; Dingus et al. 2006). Reducing the negative effects following the human factor uncertainties has become a top priority for many government agencies and automobile manufactures alike in order to improve road traffic safety.

Over the course of the past decade, the automobile industry has witnessed rapid development of driving automation. Nowadays vehicles equipped with Advanced Driver Assistance Systems (ADAS) are available. These systems can provide drivers with longitude or latitude driving support in their operational design domain. More advanced technologies that allow the driver to be out-of-the-loop for extended periods are now starting to be introduced. It looks like driving automation is on the verge of becoming a wide-spread reality. However, the current Level 1 and Level 2 vehicles (SAE International) require the full attention of the driver. Even in an oncoming Level 3 vehicle, the driver has to be ready for any fallback situations. In the foreseeable future, autonomous vehicles will be operated by shared control and they will coexist with the conventional vehicles until the commercial self-driving cars finally penetrate into the market.

It is still a wide-open question of how much safety benefit the automation can bring forth when taking into account the uncertainties of a human operator. Numerous naturalistic or simulated driving studies have shown that automation potentiates out-of-the-loop problems such as a decrease in situation awareness and increases in distraction because drivers tend to be engaged in non-driving related tasks (e.g., watching movies, using phones, reading books, etc.) during automated mode (Lu et al. 2016; Yoon and Ji 2019; Large et al. 2017; Lee, Yoon, and Ji 2021; Niu et al. 2019). With these considerations, the key safety benefit of automated vehicle technologies may easily get buried (Yang and Fisher 2021). To ensure a safe and successful

transition from human operators to automated diving systems or vice versa, it is imperative to understand how the humans behave behind the wheel and how they interact with the car itself.

Driver's postural kinematics is a strong indicator of driver's state. For example, driver head orientation can serve as an approximation of the gaze direction to determine the driver's intention, attention and alertness level (Tawari, Martin, and Trivedi 2014; Hu, Jha, and Busso 2020; Venturelli et al. 2017; Xing et al. 2017; Lee et al. 2011). The distance between driver body extremities and control interfaces (steering wheel and pedals) have been proved to be highly correlated with the driver's distraction or the readiness to take over (Deo and Trivedi 2019; Xing et al. 2017; Yamada et al. 2018; McGehee et al. 2016; Wu et al. 2017). Therefore, the monitoring of the driver's posture is of critical importance for the realization of better ADAS.

Another application of driver posture monitoring is related to the vehicle passive safety. In an autonomous vehicle, the new non-driving related postures such as highly reclined positions may pose a clear challenge to the traditional restraint systems that are usually calibrated only for standard driving posture (Wu et al. 2020; Jiang et al. 2020; Leledakis et al. 2021). In this case, the tracking of driver body locations can be used to modulate the collision response of the restraint systems so that the injury risk during unavoidable collisions can be mitigated (Filatov et al. 2019; Zhou et al. 2017).

In spite of the advancement of sensor technologies and artificial intelligence, a posture monitoring system for automotive applications, however, remains a challenging problem (Wang et al. 2019). Among the vast body of literature, most of the systems are based on computer vision techniques. Although promising progress has been made, the vision-based methods are still suffering from the sub-optimal placement of camera due to the limited space in the cabin and body occlusions in the field of view. To simplify the monitoring task of driver's posture, most of the studies only focused on specific body parts such as hands or head. The partial coverage of the driver's body may satisfy specific research purposes but can lead to biased understanding of driver's behavior. Furthermore, the postural information detected by most of the existing systems remain on the 2D image, which hinders the interpretation of the posture in 3D space. A few studies attempted to introduce successful generic posture recognition algorithms to monitor driver's upper-body. However, these generic methods did not perform well in the cabin because the training data and the testing data were not in the same domain. In addition, few of the existing systems were well validated due to the lack of data annotations in 2D or 3D. Apart from the vision-based methods, pressure sensor based methods have been investigated by several researchers to predict driver's posture. However, the relationship between pressure distribution and driver postures is still unclear.

Recently, the creation of in-vehicle driver posture datasets has become a trending topic (Feld et al. 2021; Borges et al. 2021), because the datasets allow the direct comparison of different methods with the state of the art and they open up challenging questions to the research community. Nevertheless, few of them are publicly available and even fewer can indeed benefit the community because of the insufficiency of the data and the incompleteness of data annotations.

The main goal of this thesis therefore is to develop a reliable driver posture monitoring system based on a domain-specific posture dataset.

The rest of the thesis is organized as follows:

- Chapter 2 provides a detailed literature review of existing studies related to driver posture monitoring. Advantages and disadvantages of current monitoring systems and driver posture datasets are analyzed and summarized. Future research and development directions to overcome the existing limitations are suggested.
- Chapter 3 presents a novel framework to create a state-of-the-art in-vehicle driver posture dataset which serves as the data foundation for the following chapters. Twenty-three male and female drivers were asked to perform 42 driving and non-driving related tasks on an experimental mockup. Three depth cameras and two pressure mats were deployed to monitor driver's full body posture. Driver motions were collected by a marker-based motion capture (Mocap) system. The raw Mocap data was reconstructed to obtain the ground truth driver postures and the postural measurement from different depth cameras is spatially aligned and temporally synchronized. Finally, a data augmentation approach based on computer graphics was proposed to further enrich the database while automatically gathering data annotations.
- Chapter 4 illustrates a prediction model for the pose estimation of driver's upper body. In this chapter, a 3D posture recognition model based on a depth camera was at first established by taking advantage of current prevalent methods. In parallel, feature engineering was performed to reveal the relationship between pressure distribution and driver trunk postures. Then a posture correction framework based on filtered pose graphs, sensor fusion and Kalman filter was designed to reduce posture recognition errors caused by body occlusions or the uncertainty of the recognition model.
- Chapter 5 proposes feature-based methods to predict the orientation and position of driver's head. Five head key points were extracted by using a convolutional neural network first, then a rigid-body matching algorithm and random regression forests were applied and their performance compared.

- Chapter 6 presents methods to predict driver's feet positions. Based on the pressure sensors, pressure features constructed in Chapter 4 were evaluated and selected to train classifiers for the prediction of three right foot positions and two left foot positions. Based on a depth camera, shank related keypoints as postural features were first identified by using Mean Shift clustering analysis of the point cloud in the legroom space. Then the features were evaluated and selected to train machine learning classifiers to predict the same predefined feet positions. The results using different sensors were compared, and sensor fusion methods were investigated for better detection of feet positions.
- Chapter 7 summarizes the main findings and limitations of this present research. Directions for the future research are discussed.

Chapter 2

2 State of the art

This Chapter gives an overview of the existing studies related to driver posture monitoring, up to the author's knowledge. The driver posture monitoring systems are presented and analyzed in Section 2.1. The issues concerning the current driver posture datasets are addressed in Section 2.2. Section 2.3 highlights several knowledge gaps in the hopes that future research efforts can help fill them. Finally, the objectives of this thesis are specifically formulated in Section 2.4.

2.1 Monitoring systems

Driver posture can be measured by invasive motion capture systems, such as wearable Inertial Measurement Units (IMU) (Sathyanarayana et al. 2008; Lee, Chong, and Lee 2017; Choi et al. 2018; Liu, Wang, and Qiu 2020; Ansari, Du, and Naghdy 2020). The main limitation of these techniques is that the reliance on devices attached to driver's body may degrade driving performance and compromising driving safety. They are preferred in an experimental research context, but are difficult to be implemented in real driving situations.

For this reason, the scope of this Section is limited to the systems that monitor driver postures in a non-intrusive manner, which is practically more feasible in a real car. Specifically, the review will be focused on the vision-based and pressure sensor based systems that have been frequently involved in recent studies.

2.1.1 Vision-based systems

Driver posture monitoring is one of the most active research areas in both machine learning and computer vision. In recent years, the rapid development of sensing technology and artificial intelligence has spawned many new ideas and concepts for the recognition of human body posture in a car. These systems can be divided into different groups according to the type of vision sensors involved, e.g., RGB camera, infrared (IR) camera, longwave infrared (LWIR) camera, stereo camera, depth camera or RGB-D camera.

At the early stage of the exploration, RGB camera was the most popular sensor used for in-vehicle posture monitoring. Veeraraghavan et al. (Veeraraghavan et al. 2005) presented a monitoring system utilizing the silhouette appearance obtained from skin-color segmentation. Unsafe driver activities, such as talking on a cellular telephone, eating, or adjusting the dashboard radio system were classified using supervised and unsupervised learning methods. Zhao et al. (Zhao et al. 2012) extended and improved Veeraraghavan's work to recognize four driving postures: grasping the steering wheel, operating the shift lever, eating and talking on a cell phone. Yan et al. (Yan, Coenen, and Zhang 2014) applied a motion descriptive shape based on a motion frequency image and the pyramid histogram of oriented gradients (PHOG) to detect the right hand in four different driving actions. Based on this work, the authors extended the class of the right-hand actions to eight and used a three-level hierarchal classification system to overcome the difficulties of some overlapping classes, with the accuracy being over 87.2% (Yan, Zhang et al. 2015). In order to locate the hands of the driver, Ohn-Bar et al. (Ohn-Bar, Martin, and Trivedi 2013) presented a fusion of classifiers to detect existence of hands in three regions: wheel, gear, and instrument panel (i.e. radio). Das et al. (Das, Ohn-Bar, and Trivedi 2015) assembled an annotated video-based naturalistic driving dataset and used Aggregate Channel Features (ACF) based on boosting decision trees over color and shape descriptors for the hand detection under challenging naturalistic driving settings. Guo et al. (Guo et al. 2007) utilized the face image templates distributed in the pose space to determine the head pose. This appearance-based method assumes that as long as a large number of training samples are used, the mapping relationship between the three-dimensional pose of the face and the two-dimensional image features of the face can be found. Similarly, Murphy-Chutorian et al. (Murphy-Chutorian, Doshi, and Trivedi 2007) proposed a static head detector based on image local gradient features and support vector regression. The system determines the three-dimensional position of the head by matching and tracking the face. Wu and Trivedi proposed a pose classifier that quantifies head yaw and pitch angles by using a coarse-to-fine strategy. First, a coarse pose estimate is obtained by nearest prototype matching with Euclidean distance in the subspace of Gabor wavelets. Second, the head pose estimate is refined by analyzing the finer geometrical structure of facial features. In a study conducted on naturalistic driving data by Martin et al. (Martin et al. 2012), the geometric configurations of prominent facial features (e.g., eye corners, nose corners, and the nose tip) in the image were analyzed to estimate the head pose. Tran et al. (Tran, Doshi, and Trivedi 2012) used image optical flow and hidden Markov models to characterize the temporal foot behavior during pedal deployment.

In 2017, Cao et al. (Cao et al. 2017) in Carnegie Mellon University released a real-time multi-person 2D pose estimation framework — OpenPose. OpenPose was trained on large-scale generic human posture datasets and it features a non-parametric representation referred to as Part Affinity Fields (PAFs) based on deep learning to associate body parts with individuals in the RGB image. Unlike the traditional methods, this bottom-up system can track the full body keypoints with high accuracy, regardless of the number of people in the image. Since the advent, OpenPose was deemed as the milestone of human posture recognition and it has given rise to a variety of applications in various fields. Inspired by the promising performance in an open area, Li et al. (Li et al. 2019) directly used OpenPose to extract the driver's upper body joints to predict abnormal behavior in autonomous vehicles. In a naturalistic study related to driver's readiness to take over (Deo and Trivedi 2019), the driver's upper body postural information was also from OpenPose. In another naturalistic driving study (Yuen and Trivedi 2019), the authors found that OpenPose experienced some performance issues (e.g., no output, confusion of left and right arms) due to the unique camera angle in the car. To resolve these issues, a derivative version of OpenPose was proposed to analyze the forearm postures of the driver and the front passenger in an autonomous vehicle.

The major challenge for RGB image based posture recognition algorithms is the varying illumination conditions in realistic driving scenarios where the excessive sunlight and headlight may wipe out the salient color information and most of the image features will disappear under shadows or complete darkness.

In order to improve the robustness of these methods, infrared (IR) cameras that do not rely on the natural light were introduced to monitor the driver's posture. Cheng and Trivedi (Cheng and Trivedi 2010) extracted histogram-of-oriented-gradients features from IR images and used support vector machines (SVM) and median filtering to identify whether driver's hand was in the center console area. Fu et al. (Fu et al. 2013) designed a system that categorizes the head pose into 12 different gaze zones based on facial texture features on the IR image. Lee et al. (Lee et al. 2011) used an elliptical face model to infer the yaw estimation when the head rotated away from the frontal pose. The authors trained classifiers in a supervised framework to determine 18 gaze zones. A few researchers also investigated the potential of using long-wave infrared (LWIR) cameras to predict driver head orientation (Kato, Fujii, and Tanimoto 2004), head position (Trivedi et al. 2004) and hand position (Cheng and Trivedi 2006), etc. The LWIR camera produces qualitative temperature measurements by generating images whose pixel intensity is proportional to skin's surface temperature, and thus does not exhibit problems with changing visible illumination either. However, the LWIR camera has a nonstationary skin

temperature-to-intensity mapping and seems not be a good choice in real driving situations, because the heat from car engine, sunshine, etc. could make thermal images very noisy.

Although the IR cameras overcome the lighting issues exhibited by RGB cameras, they share another limitation with the latter: the postural information is restricted to be on the 2D images. As driver's posture is three dimensional, the lack of the third dimension will inevitably lead to inaccuracies of posture recognition.

In order to accurately locate human's body parts in a car, stereo cameras have been employed. This system is based on the principle that depth information can be computed by triangulation from two or more lenses with a common area in their field of views. Using this system, Boverie et al. (Boverie, Quellec et al. 2000) developed a method to measure the distance between the occupant and the dashboard. In a study by Krotosky et al (Krotosky, Cheng, and Trivedi 2008), stereo cameras were used to track the occupant's head. Cheng and Trivedi (Cheng and Trivedi 2004) proposed a shape-from-silhouette (SFS) based method relying on volumetric voxel data from stereo cameras to describe the pose of the occupant's head and torso. Similarly, Tran and Trivedi (Tran and Trivedi 2009) proposed a driver's upper body posture recognition system. The system first tracks the spatial position of the driver's extremities (head and hands), and then uses inverse kinematics to estimate the entire upper body posture.

Compared with RGB and IR cameras, stereo cameras offer the potential to reconstruct detailed driver posture in 3D and has the advantage of being less sensitive to body occlusions due to multiple view inputs. The biggest challenge for stereo cameras lies in the high computational cost of the existing reconstruction algorithms. Perhaps due to this limitation, stereo cameras are rarely employed in recent studies.

As opposed to the traditional stereo cameras, depth cameras demonstrate a good compromise between accuracy and computational efficiency regarding 3D reconstruction. The depth camera comes with an infrared light source and uses Structured Light or Time of Flight (TOF) technology to generate depth images with distance information (Pagliari and Pinto 2015). Therefore, they are not sensitive to the illumination conditions. Nowadays, commercially affordable depth cameras are available, such as Kinect, ASUS Xtion and RealSense, among many others. The increasing maturity of depth sensing technology also brings new opportunities for 3D human posture recognition (Berretti et al. 2018; Wang et al. 2018). It is worth mentioning that the Kinect sensor is shipped with an official posture recognition algorithm (Kinect algorithm), which is trained on millions of depth images captured in an open area and features the extraction of 3D skeleton of full body from a depth image in real-time (Shotton et al. 2011).

Motivated by the advantage of 3D posture recognition, a few researchers attempted to directly use the Kinect algorithm to recognize the driver's upper body posture. For example, Toma et al. (Toma, Rothkrantz, and Antonya 2012) built a rule-based expert system to correct the novice driver's maneuver by analyzing the upper body posture information from Kinect along with the head, eye cues and simulated driving environment information. Similarly, Shia et al. (Shia et al. 2014) described a real-time semi-autonomous system that was able to correct a driver's operational input when possible in a safe manner by observing the driver's posture from Kinect in the laboratory.

In further studies, researchers introduced the Kinect sensor into a real car to monitor driver's upper body posture (Kondyli et al. 2015; Xing et al. 2017; Craye and Karray 2015). The results of these studies demonstrated that the performance of Kinect algorithm was far from satisfactory. This is mainly because the algorithm was originally designed for human-computer interaction in in-door scenarios. In order to ensure the reliability of posture recognition, the Kinect sensor has to be placed at 1.5~2m in front of the human body. This requirement is difficult to be fulfilled given the limited space in the cabin. In addition, driver's body is in close proximity of vehicle interior (e.g., steering wheel, driver seat and other objects), leaving critical challenges for foreground segmentation which is an important prerequisite for the implementation of Kinect algorithm. Moreover, the algorithm is also subject to body occlusions which take place more often than not in the cabin (Plantard et al. 2015). The extent of body occlusions presumably varies across different viewing angles, yet few of the previous studies have touched this topic to investigate the impact of camera positions on posture recognition.

To improve the driver posture recognition using a depth camera, Yamada et al. (Yamada et al. 2018) collected dome driver posture data to optimize the key parameters of the Kinect Algorithm. In a preliminary study of this thesis, a priori driver motion database was built and a data-driven approach was used to optimize the noisy upper body joints (head, center of shoulders, shoulders, elbows, wrists and hands) positions returned by the Kinect algorithm. The result was promising, suggesting the utility of motion databases for accurate posture tracking. In another study, Kondyli et al. (Kondyli et al. 2018) proposed a graph-based skeleton matching algorithm to extract the skeleton model of the driver's upper body.

Recently, the fusion of multi-modal image data for 3D human body posture monitoring has attracted much attention (Wang et al. 2018; Han et al. 2017). These systems attempt to merge the advantages of different image data (RGB, IR or depth) so that the robustness of human posture detection could be potentially increased. For example, in a recent study by Deo and Trivedi (Deo and Trivedi 2019), the authors used three RGB cameras, an IR camera and a depth

camera to monitor the driver's posture including head, upper limbs, hands and feet. Park et al (Park et al. 2020) proposed a method for real-time tracking of the head position of the occupants using both RGB and depth images from a depth camera Kinect v2. Martin et al. (Martin et al. 2019) proposed to use both RGB and depth images from a Kinect v2 to recognize the driver's upper body posture.

2.1.2 Pressure sensor based systems

The pressure sensor consists of a series of force sensor elements organized in a specific structure, usually a matrix. Using the pressure imaging technology, the pressure distributions in the contact area can be displayed in real time. As vision-based systems are sensitive to body occlusions, pressure sensors could be useful to provide complementary cues for robust driver posture monitoring.

Shin et al. (Shin et al. 2015) embedded eight force sensors on the seat pan and applied empirical criteria to analyze the pressure differences between left and right side, front and back side of the seat pan to classify four types of driver postures (sit straight, tilt left, tilt right and lean forward). The developed system showed an average accuracy of 77.5% for five subjects. In the system proposed by Ding et al. (Ding, Suzuki, and Ogasawara 2017), a Support Vector Machine (SVM) classifier was trained on the pressure values from both backrest and seat pan to classify eight driver activities (pressing accelerator, looking right/left, looking right/left rear, holding phone right/left and pressing brake) with a total accuracy of 80%. However, the classifier was only tested on the same participants, whose data were used for training. On the other hand, Okabe et al. (Okabe et al. 2018) also used the SVM technique. Apart from the original pressure values from the backrest and seat pan, authors added the changes of the Centre of Pressure (COP) with respect to their normal positions to distinguish three activities including cell phone use, forward gaze (normal state) and sleeping performed by 14 drivers. An accuracy of 76.8% by the LOO cross-validation was achieved. In these two studies (Ding, Suzuki, and Ogasawara 2017; Okabe et al. 2018), authors attempted to predict driver's head and hand movements. As pressure results from the contact between the trunk and thighs with the seat, activities mainly involving the movement of head or hands without trunk movement are difficult to be detected. This may explain why low classification accuracies were obtained by these two studies (e.g., 15% and 20% were determined for looking right/left in (Ding, Suzuki, and Ogasawara 2017) and 59.6% for the forward gaze in (Okabe et al. 2018). In the work by Vergnano and Leali (Vergnano and Leali 2019), sixteen force sensors were deployed onto driver seat (10 on the backrest) to detect if a driver was in an Out-Of-Position (OOP). By analyzing

the COP trajectories and vehicle acceleration, authors reported that three types of OOP (forward, left and right inclined trunk positions) could be detected with unspecified accuracy. In these studies, different pressure information such as the pressure distribution, COPs, etc. were used for recognizing driver's posture. However, the performance of most of the existing methods on a new driver was unknown. Moreover, few studies investigated the recognition of driver's feet position by using pressure sensors.

2.2 Datasets

Generic posture recognition methods such as OpenPose (Cao et al. 2017) and Kinect algorithm (Shotton et al. 2013) suffer from performance degradation when used in the vehicle cabin. This problem is also called the "curse of domain shift" because the training datasets and the driver posture data in a car did not live in the same domain. To take advantage of the learning power of existing models, a good solution is to perform domain adaptation by using in-vehicle driver posture datasets (Yamada et al. 2018; Yuen and Trivedi 2019; Torres et al. 2019). On the other hand, such datasets can benefit the development of new algorithms better suited to the automotive context (Borghi et al. 2017; Hu, Jha, and Busso 2020; Kondyli et al. 2018).

In the automotive contextualization, supervised learning techniques has proven to present the best results for human posture recognition, but they come with important requirements: 1) relevance of the dataset to the anticipated deployment of monitoring systems, 2) inclusion of annotations in 2D or 3D depending on the output of the system, 3) large-scale data size to avoid overfitting.

2.2.1 Existing driver posture datasets

Due to the increasing importance of driver monitoring, a wide range of in-vehicle datasets were collected and even made publicly available. These datasets fall into two groups: driver activity classification oriented and driver posture recognition oriented. Both are aimed to provide data support for developing driver monitoring systems, but they differ in the intended monitoring functions.

Among the first group is the StateFarm dataset launched by the US insurance group (StateFarm 2016), followed by the AUC dataset proposed by the researchers in the American university in Cairo (Eraqi et al. 2019), the Drive&Act dataset proposed by the researchers in Karlsruhe Institute of Technology (Martin et al. 2019), the more recent 3MDAD dataset proposed by the researchers in the Université de Sousse (Jegham et al. 2020) and many others. These datasets feature large volume of driver upper body images from one or more views when

drivers are performing various activities in naturalistic or simulated driving settings. The purpose of these dataset is to serve as benchmarks for investigating better methods to automatically detect if drivers are engaged in secondary activities (e.g., talking on the phone, eating, reaching object, etc.). This kind of methods attempts to directly infer the type of driver activity from the images without explicitly interpreting driver's posture. Accordingly, the image sequences in these datasets were only labelled by different driver action names. In fact, even when executed by the same driver, the same action may appear differently. In addition, some different in-vehicle actions may appear the same from a specific point of view. These inter- and intra-class similarities may lead to ambiguities for classification.

In contrast, the second group of datasets aim to help develop the algorithms that are able to specifically localize and track body parts in order to provide high level information for understanding driver's activities and behaviors, analyzing their vigilance, and investigating their arousal level. Das et al collected video-based datasets from their laboratory and YouTube channels, called VIVA Hand dataset (Das, Ohn-Bar, and Trivedi 2015) and VIVA Face dataset (Martin, Yuen, and Trivedi 2016), for the task of hand detection and head pose estimation respectively, under challenging naturalistic driving settings. The datasets included RGB images captured from different camera viewpoints and the hands' positions, face key points on the images were also annotated. The limitation of these datasets is that the recorded postural measurement is from 2D images which hamper the interpretation of driver posture enrolled in the 3D world.

For this reason, the sensors that can provide depth information have been integrated into more and more datasets so that more robust and accurate posture recognition can be achieved. Borghi et al (Borghi et al. 2018) introduced an annotated hand dataset called Turms, which consisted of 14k stereo infrared (IR) images of driver hands obtained during naturalistic driving through a stereo camera placed on the back of the steering wheel. Roth and Gavrila introduced a driver head pose dataset DD-Pose (Roth and Gavrila 2019), which features stereo IR images of 27 drivers during 12 naturalistic driving scenarios. The depth information was extracted from the IR image pairs from a front facing camera and the precise 6 degrees of freedom (DOF) head pose annotations are obtained by a motion capture sensor and a novel calibration device. Borghi et al proposed the Pandora dataset (Borghi et al. 2017) for the estimation of head and shoulder postures from depth image. Pandora contains 110 sequences collected from different subjects performing similar driving behaviors in a laboratory environment. Head and shoulder orientation were captured through inertial sensors. Schwarz et al introduced a driver head pose dataset — DriveAHead (Schwarz et al. 2017), which provided IR and depth frames from a

Kinect camera placed on the dashboard in a real driving scenario. A more extensive head pose dataset proposed by Selim et al (Selim et al. 2020) is called Autopose, which was collected from simulated driving scenario and provides ~ 1.1M IR images from the dashboard view, and ~ 315K from Kinect v2 (RGB, IR, Depth) from center mirror view. In the last two datasets, head position and orientation were captured by an optical Mocap system.

Recently, methods for generating upper body pose datasets have been published. For example, Feld et al (Feld et al. 2021) designed a DFKI cabin simulator to generate dataset. The simulator is equipped with an optical Mocap system to record the driver's body movements, and a wide-angle depth camera for postural measurement in both RGB and depth images. The dataset generated by this testbed can be potentially used for seat occupancy classification, 3D body pose tracking and recognition of driver's gesture, activity and intention. In a more recent study, Borges et al (Borges et al. 2021) proposed a system to generate the so called MoLa R8.7k InCar (InCar) dataset on a simulated mockup. The relative driver motion was recorded by an inertial suit and globally positioned by three head markers tracked by optical Mocap system. Similarly, a depth sensor was used to capture the driver upper body images. To date, there is no dataset that has covered the measurement of full body posture of the driver and there is no dataset that has integrated pressure distribution data in addition to the images.

2.2.2 Mocap systems and motion reconstruction

Human body's ground truth motion data serves as a reality check for the accuracy of posture recognition algorithms particularly for the 3D postural monitoring. In this research field, the lack of reliable source of driver motion data has hindered progress in research into achieving accurate driver posture monitoring. This explains why the Mocap data appeared in more and more driver pose datasets. The Mocap systems that were frequently used to create driver pose datasets are based on either optical cameras or Inertial Measurement Units (IMU). Here we review their principles and the respective limitations in the context of driver posture monitoring.

Optical-based Mocap systems are used by researchers across several R&D fields (Sigal, Balan, and Black 2010; Shotton et al. 2011; Plantard et al. 2015) and they can be separated in two types: marker-based systems and markerless systems. the marker-based systems can accurately measure the position of markers usually attached on the body surface, are not very sensitive to illumination change, and their set-up are longer than markerless systems. However, the markerless systems are much less accurate. The marker-based systems, such as the Vicon and OptiTrack, are usually regarded as the gold standard for motion capture. However, they suffer from the fact that they need to have the markers attached onto the body skin which

requires the subject's body to be exposed as much as possible. This will reduce the fidelity of driver postural measurement compared to the normal driving situation. Moreover, some markers will be inevitably occluded during driver body movement. When using these systems to track the markers on driver's head (Schwarz et al. 2017; Selim et al. 2020), this may not be a problem. But when it comes to the tracking of markers on the full body on the driver seat (Feld et al. 2021), the tracking will presumably turn out to be a disaster. In this circumstance, the plug-in gait model will fail to reconstruct the true posture because this model requires a confined and controlled space that does not resemble the in-vehicle scenario. Therefore, efficient motion reconstruction methods are needed to obtain the true postures from the incomplete marker sets in heavily occluded scenarios.

IMU-based Mocap systems are based on body-worn IMU units that can measure and report body segment orientation as well as joints' positions. However, it has been reportedly mentioned that they are prone to errors caused by drift or magnetic sensitivity (Liu, Wang, and Qiu 2020; Borges et al. 2021). Another issue for the estimation of full-body kinematics is concerned with the need of a biomechanical model and its initial calibration. During calibration, the subject needs to stand in a standard calibration posture to keep all segments strictly aligned and the coordinate systems of all joints parallel to one another. This calibration procedure is prone to introduce a systematic error that offsets the segments' orientations and joints' positions. In the study by Borges et al (Borges et al. 2021), the authors aimed at overcoming the drift issue by using an optical Mocap system in conjunction with the IMU-based Mocap system, however there were still observable errors related to sensor fixation and soft tissue movement. Considering the accuracy and reliability requirement of driver motion data, it is not practically feasible to use IMU-based Mocap systems for the in-vehicle environment.

2.2.3 Data annotation & augmentation

Data annotation is an indispensable task when preparing ready-to-use driver pose datasets. The 3D annotations from Mocap systems are usually enough for 3D pose estimation. Yet, the use of 2D image annotations, e.g., the categorization and labelling of body parts on the image, along with 3D annotations can lead to better accuracy and precision (Shotton et al. 2013).

Most of the previous works have relied on human annotators to locate the position or to identify the presence of hands (Das, Ohn-Bar, and Trivedi 2015; Borghi et al. 2018), face key points (Martin, Yuen, and Trivedi 2016) or upper body joints (Torres et al. 2019; Yuen and Trivedi 2019) on drivers' images. This is a tedious and time-consuming task that may prevent the dataset from reaching a large scale given a time limit. On the other hand, the hand-crafted

image labels could potentially be prone to errors, biases, inter- and intra-annotator differences. One alternative is to project 3D body joints or landmarks into the camera view provided that a joint calibration between the monitoring camera and the Mocap system has been performed (Borges et al. 2021). Although promising, this method requires highly accurate and reliable Mocap data. In addition, it cannot automatically handle the presence/absence of the body part in the image which is also an important cue for the models to learn to avoid false detection.

Data augmentation, by definition, is a strategy that enables one to increase the diversity and amount of training data available for avoiding overfitting, without actually collecting new data. Conventional data augmentation strategies rely on simple image manipulations such as image flip, rotation, scaling etc. For example, the DriveAhead dataset (Schwarz et al. 2017) was enriched by rescaling and cropping image patches. Similar approaches were used by Venturelli et al (Venturelli et al. 2017) to increase the size of the training input images also for head pose estimation. Yuen and Trivedi (Yuen and Trivedi 2019) used symmetry-mirror strategy to enrich the hand dataset while reducing the annotation time. In addition, an artificial cloud texture was overlayed onto the image to simulate the varying lighting conditions in a car. In order to increase the variability of the training dataset based on manually labeled samples, Torres et al (Torres et al. 2019) first converted the driver depth image in a 3D point cloud, which was then randomly translated in any directions and finally reprojected into a new depth image to simulate new camera positions. In the Pandora dataset (Borghi et al. 2017), random translations on vertical, horizontal and diagonal directions, jittering, zoom-in and zoom-out transformations have been exploited. Although it is an effective way of increasing image variability during training, these traditional strategies do not essentially change the postural context which is also one of the important indicators of data variability.

To date, computer graphics based posture image augmentation has attracted much attention. This novel technique involves the creation of a virtual scene that resembles the application scenario and the use of a rendering pipeline based on a virtual camera that mimics the real characteristics of the sensor. Shotton et al (Shotton et al. 2013) retargeted real Mocap data to a variety of virtual human character models to synthesize a great number of depth images and body part labels. The synthetic dataset was then used to train the posture recognition algorithm for Kinect sensor. A similar method was adopted by Martínez-González et al (Martínez-González et al. 2019). They used the generated synthetic human body depth images together with real sensor backgrounds to explore domain adaptation techniques for improving the existing posture recognition models. In a study performed by Chen et al (Chen et al. 2016), they also used this technique to boost human 3D pose estimation. Recently, a synthetic in-vehicle

dataset (SVIRO) has been released by Cruz et al (Cruz et al. 2020). The dataset contains depth images and infrared images rendered from simulated sceneries from the rear passenger compartment of ten different vehicles. Using SVIRO as benchmark, the authors addressed real-world engineering obstacles regarding the robustness and generalization of existing machine learning models for occupant classification and instance segmentation, and they demonstrated that machine learning models developed on SVIRO could be transferred to real applications.

Compared to the traditional strategies, this technique can automatically introduce variability in human shapes, body pose, background and view point configuration. Another advantage of using synthetic training images is that the annotation labels (e.g., image masks for segmentation, keypoints for posture estimation) can be automatically obtained almost for free, allowing one to scale up supervised learning to large scales.

2.3 Knowledge gaps and future research directions

In spite of the extensive effort devoted to develop driver posture monitoring systems, there are many research questions that still remain open. But this should not be construed as directed criticism of this body of work. Indeed, these studies represent the critical explorations in our understanding of driver posture monitoring systems. Future work should complement and build on this earlier work. Below, we discuss some of these issues from the global point of view, and identify several pertinent areas that merit future in-depth investigation for the topic mentioned in this thesis.

- Vision systems will continue to play an important role in driver posture monitoring. Particularly, there is growing interest regarding the depth camera. However, the research on the posture recognition algorithms better suited to in-car applications is far behind partly because of the over-reliance on the existing generic methods. In addition, few of the previous studies explored possible solutions to reduce the posture recognition errors caused by body occlusions.
- Most of the research effort has gone toward the recognition of driver's upper body including head and hands. To date, limited research focused on the monitoring of driver's lower limbs. In general, driver posture can be represented at several levels of resolution such as full body level, upper body, lower body, hands, head or feet depending on the research purposes. The monitoring systems relying on lower resolution of postural information may lead to false or biased evaluation of driver's state. Therefore, a monitoring system that is able to cover the full body of the driver is needed.

- In-vehicle driver posture datasets are of critical value for the development of better posture recognition algorithms. However, most of the currently available datasets are barely reusable due to the lack of efficient data annotation methods. Furthermore, none of them has covered driver's full body. To facilitate the research into achieving accurate driver posture monitoring, driver posture datasets with full-body measurement and high-quality annotations are hence needed.
- Pressure sensors may provide useful postural information, but the relationship between pressure measurement and driver postures is still unclear. In addition, exploring the benefit of fusing the pressure sensors with depth cameras could be an interesting topic.

2.4 Objectives of this thesis

Considering the importance of driver posture monitoring and the limitations of existing studies, the present work aims at exploring more efficient/robust posture recognition methods in order to bring driver posture monitoring systems closer to real-world applications. The objectives of this thesis can be specifically formulated as follows:

- Establishing an in-vehicle driver posture dataset, which covers the measurement of driver's full body along with high-quality ground truth posture annotations. The dataset will be made open access to facilitate the research activities in this field.

- Developing a reliable and accurate driver posture monitoring system, enabling the posture recognition of driver's head, upper-body and lower body at the same time.

Chapter 3

3 A framework for the creation of in-vehicle driver posture dataset

Driver posture datasets are crucial for developing robust and accurate driver posture monitoring systems in that they allow the direct comparison of various methods with the state of the art and they can even open challenging questions to the research community. To date, a full body posture dataset with high-quality ground truth annotations is still missing.

This Chapter proposes a novel framework to create a well-structured and extensive dataset, named AutoConduct. Section 3.1 presents an experimental setup to collect postural data from real drivers. Data processing steps are performed in Section 3.2, where the ground truth of full body postures is reconstructed and the postural measurement from different vision sensors are spatially aligned. Section 3.3 illustrates a data augmentation pipeline to synthesize driver posture images with high-quality annotations. The AutoConduct dataset is described and compared to the state-of-the-art driver posture datasets in Section 3.4. Finally, this chapter is summarized in Section 3.5.

3.1 Data collection

An experiment was first performed to collect real driver motions and real postural measurement. The experiment was conducted in the framework of the French National Project ANR AutoConduct which was aimed at designing a new Human-Machine cooperation strategy adapted to the driver's state. The experiment protocol was approved by the ethic committee of Université Gustave Eiffel.

3.1.1 Motion capture

Motivated by the accuracy of marker-based optical Mocap systems, the Vicon Nexus software (Version 2.7.5, Vicon, Oxford, United Kingdom) was used to track the spatial trajectories of the reflective markers on driver's body (Figure 3.1). The Vicon system consisted of 14 infrared cameras surrounding the scenery. The marker set was arranged in a way that there were at least three markers on each body segment. To label the markers, we created a marker template that could be used as a basis for setting up future participants with the same marker-set.

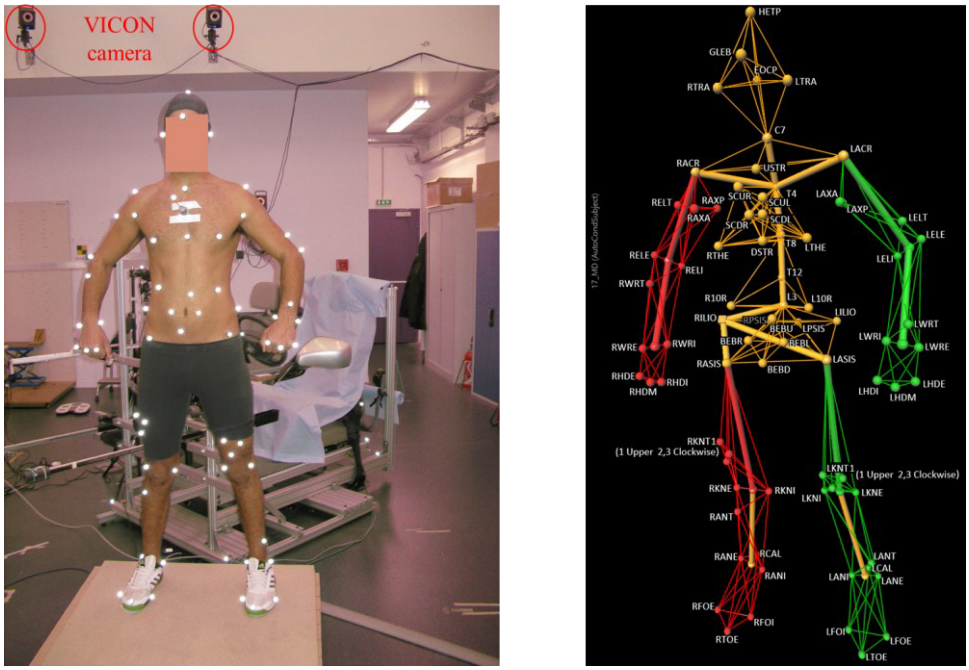


Figure 3.1. marker arrangement (left) and skeleton template (right)

3.1.2 In-vehicle sensors for postural monitoring

Three depth cameras (a Kinect v2 from Microsoft, a PMD CamBoard pico monster and a DS325 from SoftKinetic) and two pressure mats on seat pan and backrest were used to build the monitoring system. The use of these sensors was intentional and interesting because of their nature of non-intrusive detection. For a better monitoring of upper body postures, the Kinect v2 was located at the top of the right A pillar in a real car, while the PMD was mounted at the top of center rear mirror. The monitoring of driver lower body requires a depth camera that supports close-distance interaction, because the space of the leg room or footwell in a

car is quite limited. Therefore, a depth camera DS325 from SoftKinetic which has a lower limit of 15 cm as the depth range was selected. The DS325 was placed under the steering wheel and above the pedals to look at driver legs instead of feet. The specifications of the three depth cameras are given in **Table 3.1**. The two pressure mats used in this work were from Xsensor (PX100:48.48.02) (**Table 3.2**), and based on capacitive sensing mechanism. To prevent pressure mats from moving during experiment, a double-sided adhesive tape was used to fix them onto the seat.

Table 3.1. Technical specs of the depth cameras used in the present work.

Camera	L/W/H (cm)	Position	Lens	Supported image	Resolution (Pixel)	FOV (deg)	Depth resolution (m)
Kinect v2	24.9/6.7/6.6	Top of right A pillar	RGB	RGB	1920×1080	84×53	--
			Depth	IR	512×424	70×60	--
				Depth			< 0.03 @ 0.5 – 4.5
PMD	6.2/6.6/2.9	Center rear mirror	Depth	IR	352×287	100×85	--
				Depth			< 0.02 @ 0.5 – 6
DS325	10.5/3.0/2.3	Under steering wheel	Depth	IR	320×240	74×58	--
				Depth			< 0.014 @ 0.15 – 1.0

Table 3.2. Technical details of Xsensor pressure mat (PX100:48.48.02).

Property	Value
Sensor technology	Capacitive pressure imaging
Pressure range	0.14*-2.7 N/cm2
Spatial resolution	12.7 mm
Accuracy	± 10% full scale
Sampling frame rate	39 frames/s (maximum)
Total area	81.3 × 81.3 cm
Sensing area	60.9 × 60.9 cm
Thickness (uncompressed)	0.08 cm
Sensing points	2304 (48 rows by 48 columns)
Time compensation	Yes

*Pressures smaller than the lower threshold were not recorded in this study.

3.1.3 Experimental mockup

The experimental mockup is illustrated in **Figure 3.2**. This was a simplified testbed including the steering wheel, three pedals, gearshift, dashboard, as well as a tablet at right of the steering wheel to simulate future user interface for automation control. The driver seat and the steering wheel could be adjusted according to individual's preference.

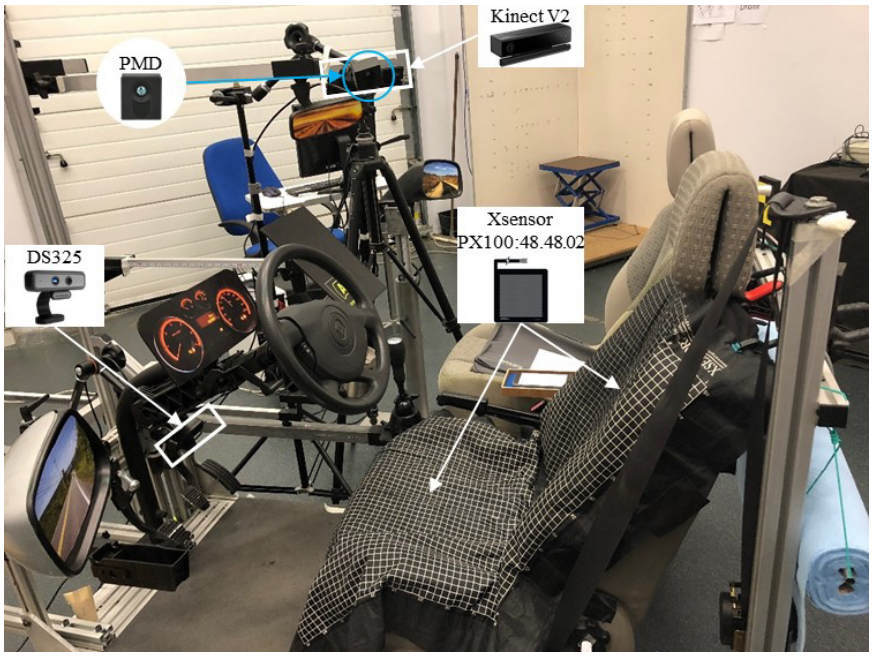


Figure 3.2. Mockup configuration.

3.1.4 Sensor calibration and synchronization

Sensor calibration is an important step for data localization, data transformation and data alignment of vision systems. For the Mocap system, we followed the Vicon calibration procedure to track the reflective markers within a precision of 3 mm.

For the depth lens within each depth camera, the standard camera calibration (Zhang 2000) was performed to obtain their respective intrinsic and extrinsic parameters. With these parameters, a pixel (u, v) in the depth image can be transformed into a point (X, Y, Z) in the world coordinate system by using Equation (3.1).

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1/dx & 0 & u_0 \\ 0 & 1/dy & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\text{Intrinsic matrix}} \underbrace{\begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}}_{\text{Extrinsic matrix}} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3.1}$$

Where dx and dy are the scale factors in image u and v axes, f is the focal length, u_0 and v_0 are the coordinates of the principal point on the image, Z_c is the depth value of the pixel at (u, v). (R, t) are the rotation and translation which relate the world coordinate system to the camera coordinate system. It should be noted that the Kinect v2 is equipped with an extra RGB lens in addition to the depth one. In order to realize the color-depth registration, a joint calibration was performed to find the relative rotation and translation between the two lenses.

Considering the computation and memory requirements, the data recording was performed on two computers, one for the Mocap system, PMD and DS325 and the second for the rest of the monitoring systems. The Mocap system was configured to record data at a frequency of 50 Hz, while the pressure mats and depth cameras shared the same frequency of 25 Hz. In order to achieve proper temporal synchronization, an electronic trigger was used. All the data recording tasks simultaneously started as soon as they received the starting signal from the trigger.

3.1.5 Participants

In this work, twenty-three drivers (12 males and 11 females) were selected by body height and mass for data collection. They ranged in age from 22 to 65 years (40 ± 11.5 for Mean ± Standard Deviation), in height from 153 cm to 195 cm (171 ± 13), in Body Mass Index (BMI) from 18.2 kg/m^2 to 43.4 kg/m^2 (27.8 ± 6.7). **Figure 3.3** gives the distribution of participants' height and BMI. Compared to the American adult population from NHANES 2015-16, the participants in this work covered a large range of population. Written informed consent from participants was obtained prior to the experiment. Prior to data recording, they were instructed to find the preferred seating configuration, seating position and steering wheel position.

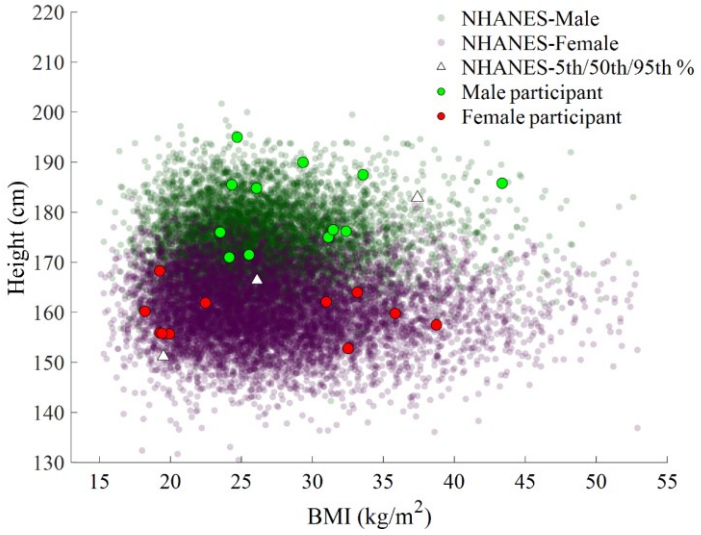


Figure 3.3. Distribution of participants' height and BMI with respect to the American adult population from NHANES (2015-16)

3.1.6 Posture variations

Posture variation is one key attribute of the human body pose datasets. A training dataset with a larger range of posture variations will enable better generalization of the posture recognition algorithms during deployment. Based on the report from the NHTSA 100-car naturalistic driving study (Dingus et al. 2006) and the recent literature review on the driver activities that may take place in an autonomous car (Naujoks et al. 2017), in-vehicle driver actions can be extensively categorized into three groups:

(1) Primary driving tasks that are important for operating the vehicle, such as changing lanes, braking, switching gear, checking rear mirrors etc.

(2) Secondary driving tasks that concurrently compete for the same resources (perceptual, cognitive or physical) required by safe driving and degrade driver performance, such as using hand-held cell phone, drinking, adjusting navigation system, etc.

(3) Non driving related tasks that take driver's hands off the steering wheel and feet off the pedals during automated driving, such as reading books, resting, sleeping, etc.

Since the purpose was to include large range of posture variations in the dataset, the tasks that require a similar posture were grouped and only one of them was selected as representative. In addition, nine general body actions such as arm abduction, head rotation, trunk rotation, etc. were designed. This led to 42 tasks in total for each participant, as listed in **Table S1**. This list was by no means exhaustive, but covered as many posture variations as possible.

During experiment, participants were guided to perform these tasks one by one without replication in a randomized order. In order to familiarize the participants to the different tasks, a hand-out with illustrations was provided.

3.2 Data processing

Extra effort is needed to transform raw trajectories of markers into an understandable format for motion interpretation. More specifically, joint centers need to be located. In addition, there is a need to spatially align the postural measurement from different sources with the ground truth data for training and evaluating posture recognition algorithms.

3.2.1 Motion reconstruction

Using marker-based optical Mocap systems, such as Vicon, to record seated body motions is challenging due to unavoidable occlusions. To obtain a good quality of ground truth driver postures from the raw Mocap data, RPx, a motion reconstruction, analysis and simulation tool developed at Univ-Eiffel (Monnier et al. 2009). The motion reconstruction process in RPx consists of two steps. First, a subject specific articulated skeleton was created for each participant from a reference standing posture. The joint positions related to hips, spine, shoulders and center of two shoulders were estimated by statistical models (Reed, Manary, and Schneider 1999; Peng et al. 2015). Regarding limbs and head, joint positions were simply determined as the center of the marker pair attached close to the target joint or body part. Once the personalized skeleton template was created, joint angles were calculated by minimizing the distance between model-based marker positions and measured ones. After motion reconstruction, driver body posture was represented as a skeleton model with 28 articulated joints from head to foot. See **Figure 3.4** for examples.

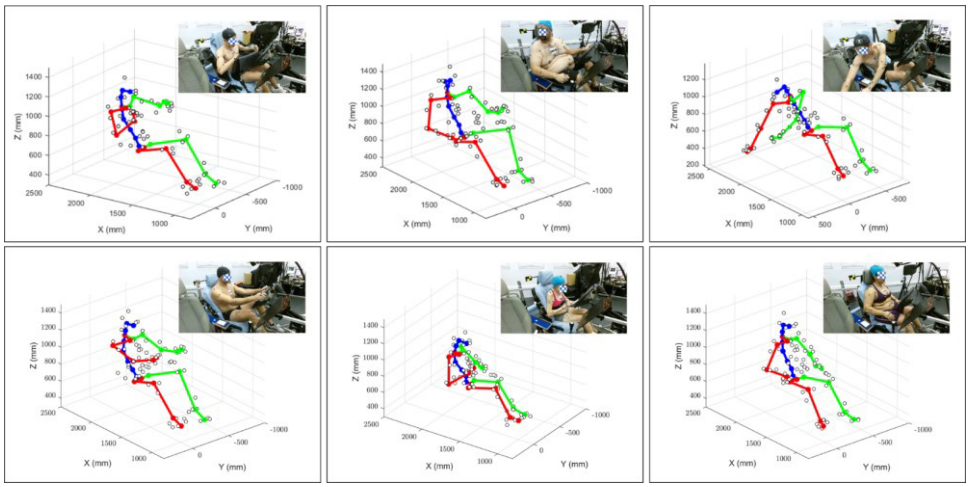


Figure 3.4. Reconstructed driver postures.

3.2.2 Spatial alignment

Using the camera parameters obtained from the calibration step, the depth image from each depth camera can be converted to a point cloud in the camera coordinate system (CS_c). The alignment of these point clouds with the ground truth driver motion data in CS_v required

the definition of a common target coordinate system (CS_t) and the calculation of transformations (rotation and translation) between the source coordinate systems and CS_t.

To this end, a customized calibration board was introduced (**Figure 3.5**). The calibration board was perpendicularly mounted in the middle of a base board. An identical chessboard calibration pattern same as the one used for camera calibration was symmetrically attached onto both faces (face A and face B) of the calibration board. Face A was used to define the world coordinate system CS_w for the depth camera. On face B, three markers were attached at three inner corners of the chessboard to establish the correspondence with the Vicon system. The physical dimensions of the chess pattern, the calibration board and the marker size were pre-recorded.

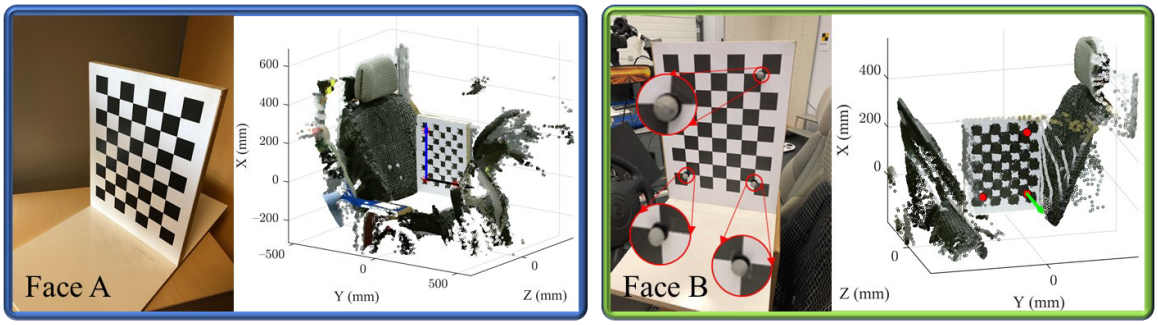


Figure 3.5. Definition of CS_t using calibration board.

The calibration board was first placed at the middle of the driver seat to define the world coordinate system for Kinect v2 (CS_w^{kin}) by using the Matlab camera calibration toolbox which gives the rotation and translation between CS_c^{kin} and CS_w^{kin}. CS_w^{kin} is regarded as the target coordinate system CS_t in this work (**Figure 3.5**), where the X axis goes upwards, the Y axis directs to front and the Z axis goes leftwards relative to the driver. With the help of the three markers on face B, the transformation between CS_t and CS_v can be obtained through an optimization problem (Equation 3.2) which minimizes the Euclidian distance between the three markers and their corresponding inner corners on face A.

$$(\mathrm{R}, \mathrm{t}) = \operatorname{argmin} \sum_{\mathrm{i}=1}^{\mathrm{n}} \|(\mathrm{R}a_i + \mathrm{t}) - \mathrm{b}_i\|^2 \qquad (3.2)$$

Where R and t denote rotation and translation, respectively. a_i $(\mathrm{i} = 1,2,\ldots,\mathrm{n})$ are the coordinates of a point in coordinate system A while b_i $(\mathrm{i} = 1,2,\ldots,\mathrm{n})$ are the corresponding

coordinates in system B. Similarly, we used the calibration board and Matlab camera calibration toolbox to find the rotation and translation between CS_c^{pmd} and CS_w^{pmd}. Then the three markers again were used to connect CS_w^{pmd} and CS_t.

Regarding the calculation of the relative positions between CS_c^{ds325} and CS_t, this method cannot be used because the calibration board was not visible to DS325 due to its placement. In this case, correspondence points were manually selected from the registered point cloud in CS_t and the point cloud in CS_c^{ds325} to find the rotation and translation between CS_c^{ds325} and CS_t so that approximate superimposition could be achieved. **Figure 3.6** illustrates the data alignment procedure and the ultimate result.

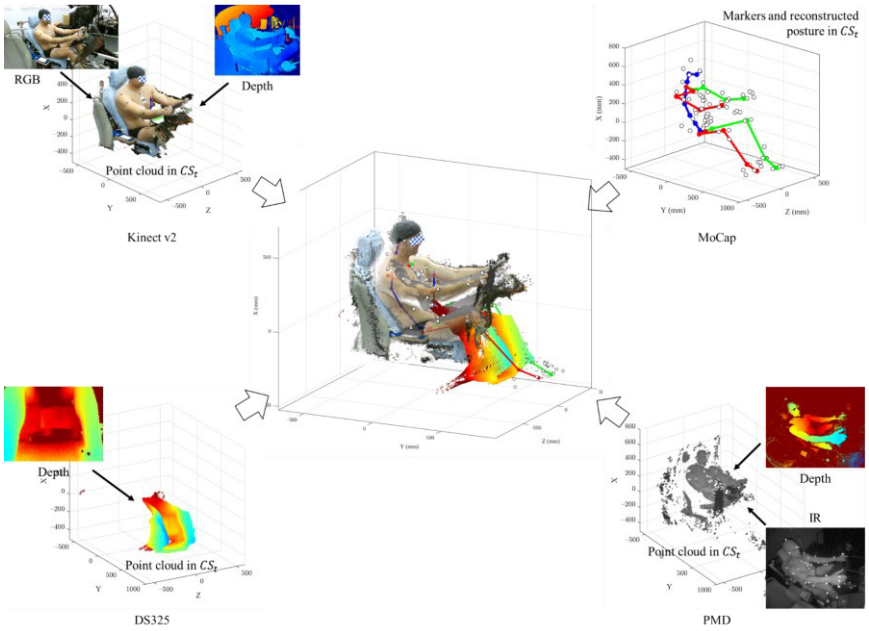


Figure 3.6. Spatially aligned experiment data

3.3 Data augmentation

The real driver upper body images collected from the experiment cannot be directly used as training dataset for several reasons. First, the participants wore a gym-suit so that reflective markers could be attached directly to the body surface in order to reduce tracking errors caused by relative movement between marker and bone. Second, the mockup structure was simplified to have as many markers in line of sight to ensure accurate tracking. These actions were designed to facilitate the motion capture at the expense of image fidelity loss with

respect to realism. As a result, the posture recognition algorithms developed on the experimental images will be biased and may not perform well in a real car. In addition, there lacks an efficient method to annotate the body parts on the real images. Due to high variability in body size and high number of postures that a drive may adopt, it is challenging to have a thorough representation of in-vehicle driver postures just by experiment.

Inspired by the previous studies (Shotton et al. 2011; Martínez-González et al. 2019), we retargeted the real driver motions to digital human characters and established a rendering pipeline to generate synthetic images by using standard computer graphics techniques (Zhao et al. 2020c). Our goals are threefold: image realism, posture variety and efficient collection of high-fidelity annotations.

The digital human characters were created in MakeHuman (an open-source tool designed to prototype realistic 3D human models, http://www.makehumancommunity.org). The basic human character was composed of a rigged skeleton (similar to the skeleton model in RPx) and a skin mesh that forms the body surface shape (**Figure 3.7** left). The skeleton was linked to the mesh using the linear blend skinning technique so that changes of joint positions led to an adaptation of skin mesh. To impersonate the real participants in our experiment, the attributes for each virtual model (e.g., age, gender, weight, height, body proportion etc.) were adjusted accordingly. We also randomly configured clothes and hairstyle materials for each base character to yield more realistic driver appearances in images (**Figure 3.7** right).

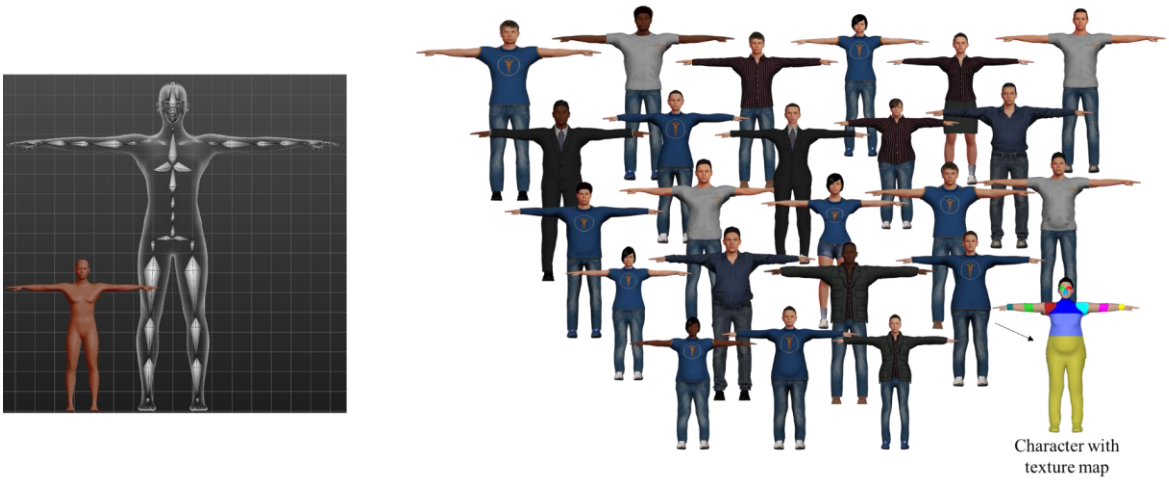


Figure 3.7. Digital human models. Left: base character. Right: characters created for real participants.

In order to collect body part labels, a color-coded texture map was designed and attached onto the body skin and the exterior accessories, as shown in **Figure 3.7** (right). The same texture map can be applied to all the characters. The localized body part labels are consistent with the mainstream posture recognition algorithms, such as Kinect algorithm (Shotton et al. 2013) and OpenPose (Cao et al. 2017) for upper body tracking. To simulate the background of in-vehicle scenery, a rigged vehicle model with complete interior was used (**Figure 3.8**).

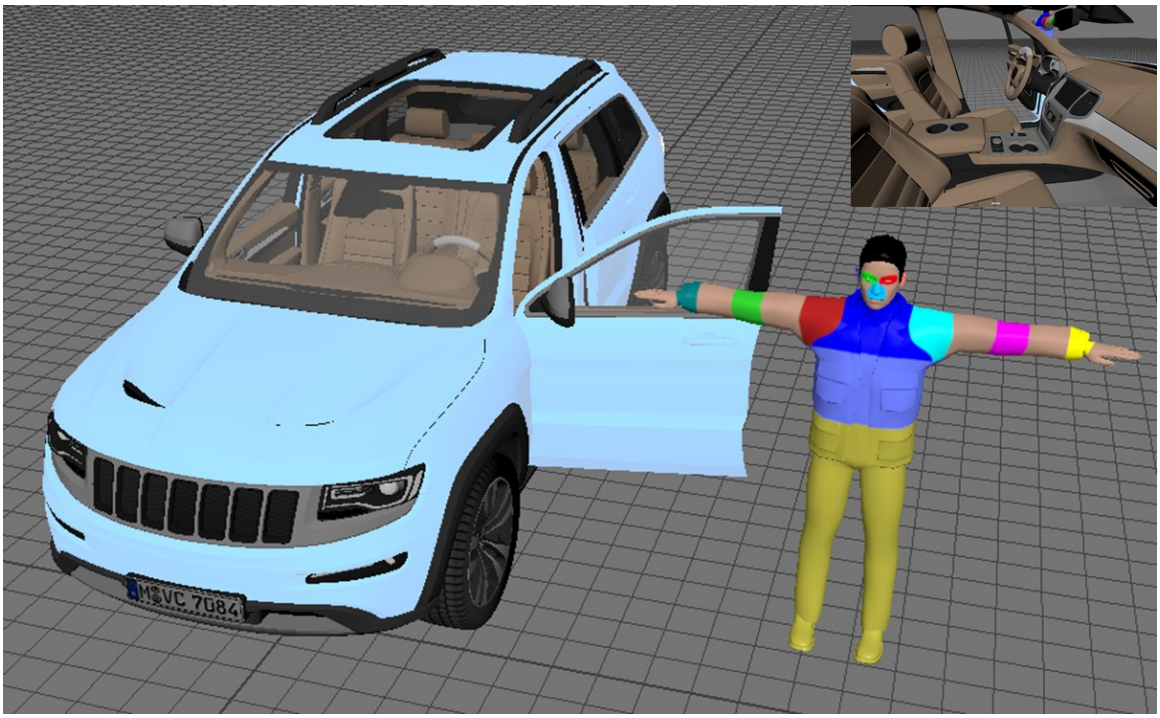


Figure 3.8. Simulated driving scenery.

Finally, we imported the real driver motions reconstructed by RPx, the digital human characters from MakeHuman and the car model into MAYA, a 3D computer animation, modeling, simulation, and rendering software package. The RPx motion was retargeted to the digital human characters by using Autodesk® HumanIK® (HIK) animation middleware (a full-body inverse kinematics (IK) solver and retargeter) (**Figure 3.9**).

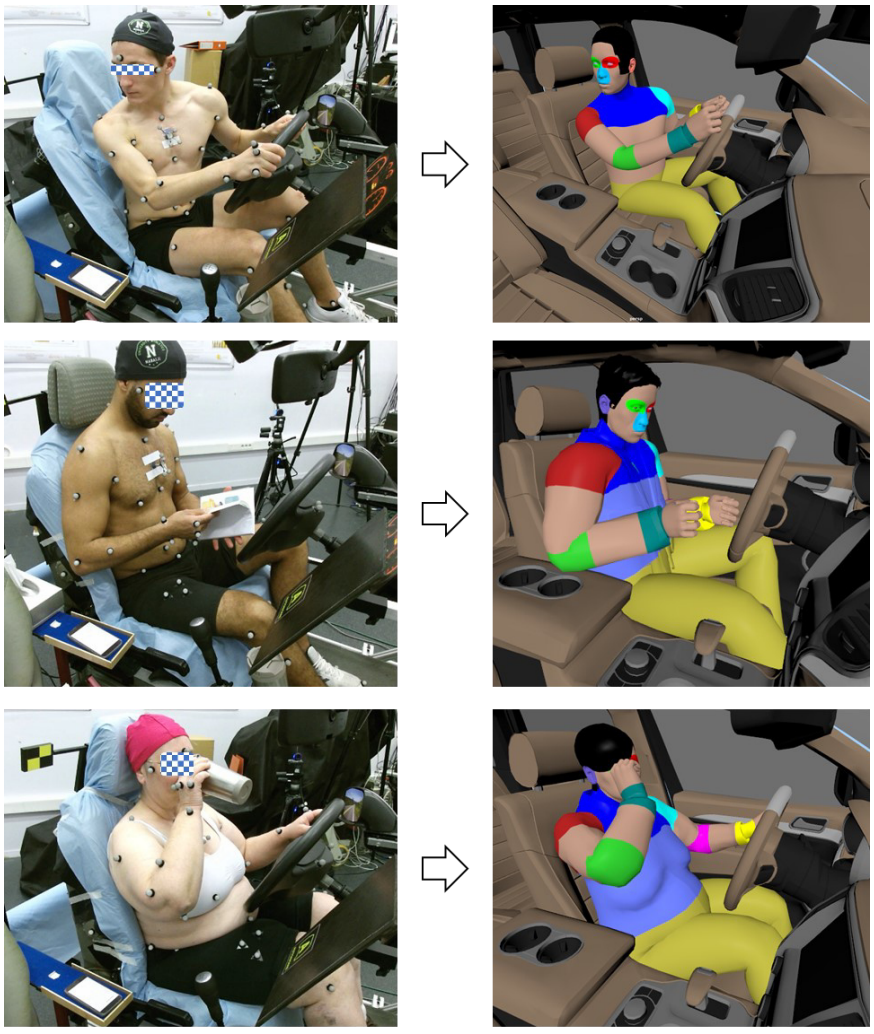


Figure 3.9. Motion retargeting to the corresponding virtual model.

Given the 23 digital human characters and the reconstructed driver motions from real participants, we not only retargeted the real motions of driver X to the corresponding character but also to the other 22 characters. This cross-retargeting strategy allows the character to perform the same task with a 23-time repetition, each time with a different style. The posture variations are therefore enriched 22 times more with respect to the real driver postures collected from the experiment. During animation, two virtual cameras, imitating the real cameras in our experiment, were used to render the scene into RGB, gray scale, depth and body segmentation images. Note that, the pixel intensity in gray scale images can be adjusted to obtain infrared images by referring the pixel depth information. In addition, the joint centers and some keypoints (e.g., nose tip, eye centers, ear centers) were trivially

recorded in the meantime. Examples are given in **Figure 3.10**. Thanks to the reference library provided by the MAYA documentation, this data augmentation procedure was automized by Python scripts.

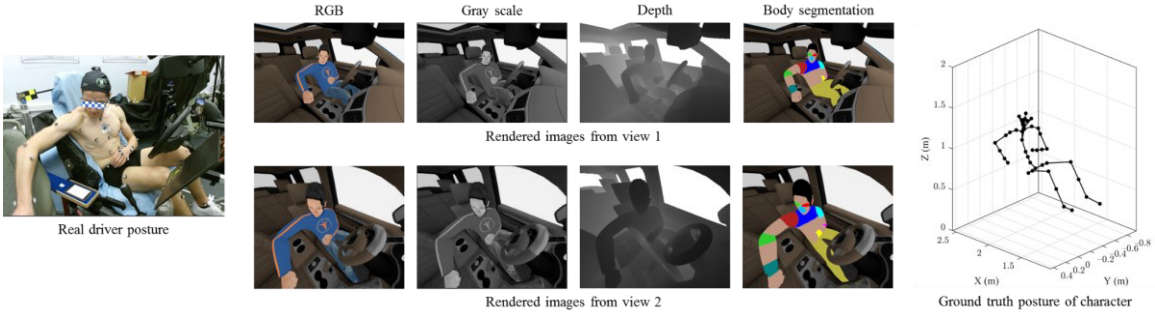


Figure 3.10. Rendered frame.

3.4 Results and Discussion

Using the proposed framework, the dataset created in the present work, named 'AutoConduct', was mainly composed of two sets: real data and synthetic data. The real data was collected from 23 drivers performing 42 activities in a controlled environment to include ~130K frames of postural measurement from the monitoring systems and ~260K frames of reconstructed driver motion data. Each frame of postural measurement consists of synchronized upper body images from Kinect v2 (RGB, depth) and PMD (IR, depth), lower body images from DS325 (IR, depth), as well as pressure distributions from backrest and seat pan. The visual postural measurement was spatially aligned with the ground truth motion data. The synthetic dataset consists of about 12 million (260K*23*2) data frames generated in a simulated in-vehicle scenario. Each frame includes RGB, infrared, depth images and 2D/3D posture annotations. **Table 3.3** shows an overview of our AutoConduct compared to other prominent datasets for driver posture recognition.

Table 3.3. Comparison of selected in-vehicle datasets for driver posture recognition

Dataset	Year	Exp settings	#Driver (f)	#Sensors	Modality	Data vol.	Annot.	Data aug.	Supported monitoring functions
VIVA hand (Das, Ohn-Bar, and Trivedi 2015)	2015*	Naturalistic driving	N/A	1 RGB camera	RGB	4k	Manual, 2D	Conventional	Hand detection
VIVA face (Martin, Yuen, and Trivedi 2016)	2016*	Naturalistic driving	N/A	1 RGB camera	Video sequences	39	Manual, 2D	–	Head pose estimation
Pandora (Borghi et al. 2017)	2017*	Laboratory	22(12)	1 depth camera	Video sequences	110	IMU, 3D	Conventional	Head pose and shoulder orientation estimation
DriveAhead (Schwarz et al. 2017)	2017*	Naturalistic driving	20(4)	1 depth camera	IR + depth	1M	Optical, 3D	Conventional	Head pose estimation
Turms (Borghi et al. 2018)	2018*	Naturalistic driving	7(2)	1 stereo camera	IR	14k	Manual, 2D	–	Hand detection
DD-Pose (Roth and Gavrila 2019)	2019*	Naturalistic driving	27(6)	1 stereo camera	IR	660k	Optical, 3D	–	Head pose estimation
Yuen & Trivedi (Yuen and Trivedi 2019)	2019	Naturalistic driving	N/A	1 RGB camera	RGB	8.5k	Manual, 2D	Conventional	Hand detection
Torres et al (Torres et al. 2019)	2019	Laboratory	5(-)	1 depth camera	depth	12k	Manual, 2D	Conventional	Upper body pose estimation
AutoPose (Selim et al. 2020)	2020*	Laboratory	21(11)	1 IR camera and 1 depth camera	RGB + IR + depth	2M	Optical, 3D	–	Head pose estimation
DFKI (Feld et al. 2021)	2020	Laboratory	N/A	1 depth camera	RGB, IR, depth	N/A	Optical, 3D	–	Upper body pose estimation
InCar (Borges et al. 2021)	2021*	Laboratory	N/A	1 depth camera	Depth	N/A	IMU, 2D, 3D	–	Upper body pose estimation
AutoConduct (our)	2021	Laboratory & simulation	23(11)	3 depth cameras and 2 pressure mats	RGB, IR, depth and pressure distribution	12.4M	Optical, 2D, 3D	Computer graphics	Head, upper body and lower body pose estimation

* Publicly available. N/A Information not clarified by the authors. – Not addressed by the authors.

In terms of the real dataset, the main advantage of AutoConduct over the others is the availability of reliable full-body motion data reconstructed by RPx, which enables the accurate estimation of joint centers in 3D. Regarding the extent of body part coverage, Pandora (Borghi et al. 2017), DFKI (Feld et al. 2021) and InCar (Borges et al. 2021) are the nearest neighbors of AutoConduct. The authors of Pandora and InCar unanimously used IMU-based Mocap system for motion tracking of the upper-body. As previously mentioned, IMU systems are subject to systematic errors that may cause data shift and there still lacks reliable method to compensate these errors (Borges et al. 2021). The authors of DFKI used the default plug-in gait model to obtain true postures from the optical motion data, but they failed to address the issues related to marker occlusions. In contrast, we used a personalized articulated skeleton model and other prior knowledge such as joint range of motion (ROM) and reference posture in RPx to reconstruct movements. Our method is well adapted for the heavily occluded situations of in-vehicle motion capture, as shown in **Figure 3.4**. Correctly reconstructed motion data is important as it lies at the core of the data augmentation method used in this work. In addition, AutoConduct is the first dataset that has provided visual measurement including driver's entire body, allowing the investigations of holistic posture monitoring approaches. Furthermore, the inclusion of body pressure distributions may be of help to provide complementary cues for robust posture monitoring.

In terms of the synthetic dataset, the tedious manual annotation work is avoided in our data augmentation pipeline thanks to the computer graphics techniques. Meanwhile, the synthetic dataset exhibits richer posture variations and more realistic image appearances with respect to the experimental data. In fact, our data augmentation pipeline can generate an infinite number of data samples by combing different viewing angles, driver poses, digital characters, garment assets, vehicle models, etc. However, the current pipeline exhibits several limitations that need to be addressed. First, the configurations of the driver seat and the steering wheel in the car model were not personalized. Second, the motion constraints imposed by the interior of the car model was not considered in the motion retargeting process. Third, body occlusions caused by objects like handbags, smartphones, tablets or bottles etc., were not considered in the current work. In addition, only the augmentation of vision data was considered when developing the pipeline. Regarding the pressure distribution data on the driver seat, it could also be virtually simulated by using an finite element model (Ren et al. 2017), which will be one of the directions of the future work.

3.5 Summary

This Chapter presents a novel framework for the creation of in-vehicle driver posture dataset. The dataset was composed of synchronized and aligned body posture data from an experiment and a large number of synthetic images from a data augmentation pipeline. The experiment data was collected from multiple depth cameras and pressure sensors covering the full body, coupled with ground truth body motions reconstructed from recorded makers attached to the body. The data augmentation pipeline enables one to generate a large number of synthetic annotated body images. In contrast to the state-of-the-art driver posture datasets, the AutoConduct dataset has advantages regarding the coverage of body parts, data modalities, posture variations and the quality and completeness of annotations. In terms of the future work, further effort is needed to improve the data augmentation pipeline and most importantly, various posture recognition algorithms will be tested and the potential of adapting existing posture recognition models will be evaluated. The author believes that the development of more performant driver posture recognition algorithms can benefit from this dataset. Once validated, the synthetic data from AutoConduct dataset and the software will be made publicly available to encourage and stimulate further research.

Chapter 4

4 Posture monitoring of driver's upper-body

Accurate upper-body pose estimation is a key task for automatic monitoring of driver's attention, intention and seated position. Due to body occlusions, sub-optimal placement of cameras and also the lack of high-quality driver posture dataset, existing upper-body posture estimation methods are not satisfactory.

Based on the AutoConduct dataset, this Chapter aims to develop an accurate and robust method to monitor driver's upper-body posture. An overview of the system is depicted in **Figure 4.1**. The system first uses a Convolutional Neural Network (CNN) adapted from OpenPose (Cao et al. 2017) to process the stream of RGB/IR frame from the depth camera in order to locate upper body keypoints. With help of the depth frames, body part confidence maps (PCM) computed from the keypoints are projected into the 3D space. An Offset Joint Regression (OJR) model adapted from a previous work by Girshick et al. (Girshick et al. 2011) is then used to find the joints' locations beneath the surface of PCMs. Meanwhile, a posture classification method based on pressure measurement is deployed. To reduce the posture estimation errors based on the depth camera, a posture correction framework based on the priori knowledge from pre-built motion graphs and Kalman filters is applied. The benefit of sensor fusion from the depth camera and pressure sensors is investigated. The ultimate output is represented by a skeleton with joints of center of shoulders, shoulders, elbows and wrists in 3D space.

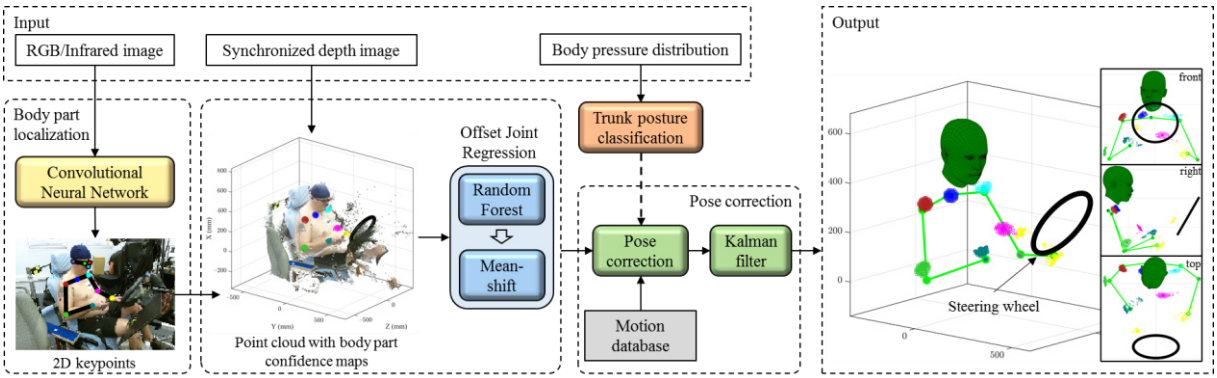


Figure 4.1. System architecture for upper-body posture monitoring.

This Chapter is organized as follows. Section 4.1 describes the posture estimation method based on a depth camera. Section 4.3 presents a posture classification method using pressure sensors. The posture correction framework is detailed in Section 4.3. Finally, main results are summarized in Section 4.4.

4.1 Posture estimation based on a depth camera

Posture estimation of human body is one of the most active research fields in both computer vision and machine learning. Although 3D skeleton-based human representations have been intensively studied (Wang et al. 2018; Han et al. 2017; Berretti et al. 2018), methods that are better suited to in-vehicle applications are rarely investigated. Based on the analysis of the existing generic posture estimation models, this Section proposes a novel method for the 3D posture estimation of driver's upper body using a depth camera. The method consists of an adapted Convolutional Neural Network (CNN) and an adapted Offset Joint Regression model (OJR).

4.1.1 Adapted Convolutional Neural Network (CNN)

To locate body part on the image, we adapted the network architecture presented in OpenPose (Cao et al. 2017) which was pretrained using the COCO 2016 keypoints challenge dataset (Lin et al. 2014). It was selected for the following reasons: 1) real-time performance, 2) well-established network design by coupling the Part Affinity Fields with the estimation of body part locations for robustness and better accuracy, 3) good domain adaptability proven by previous works (Yuen and Trivedi 2019; Torres et al. 2019) in the same research context. The network begins with an input image, which passes through the VGG-19 network followed by stages of convolutional kernel layers. The feature maps from VGG-19 are also passed directly to the beginning of each stage by using a feature concatenation layer. The loss is computed with the ground truth body part maps at the end of each stage. The network is designed to simultaneously learn body part locations and the association between matching parts using PAFs.

As driver's lower body was not fully visible in the camera view, the network was designed to output 12 keypoints: left eye (LEY), right eye (REY), left ear (LEA), right ear (REA), nose tip (NOT), center of shoulders (CS), left shoulder (LS), right shoulder (RS), left elbow (LE), right elbow (RE), left wrist (LW) and right wrist (RW). To further improve the real-time performance which is important for driver monitoring, the network was modified to utilize only 2 stages with less than half the parameters compared to the original design, as suggested by

Yuen and Trivedi (Yuen and Trivedi 2019). The adapted architecture with hyperparameters is shown in **Figure 4.2**.

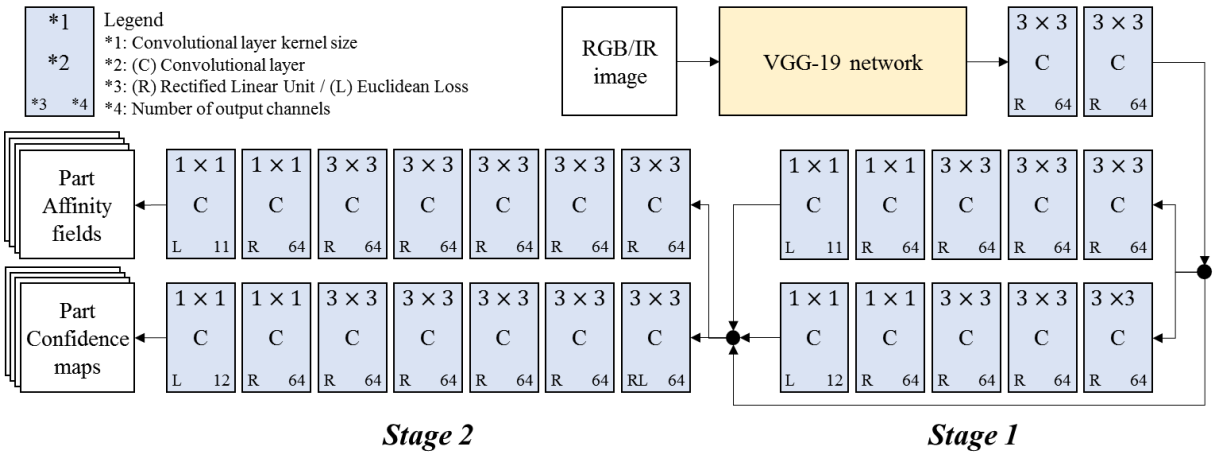
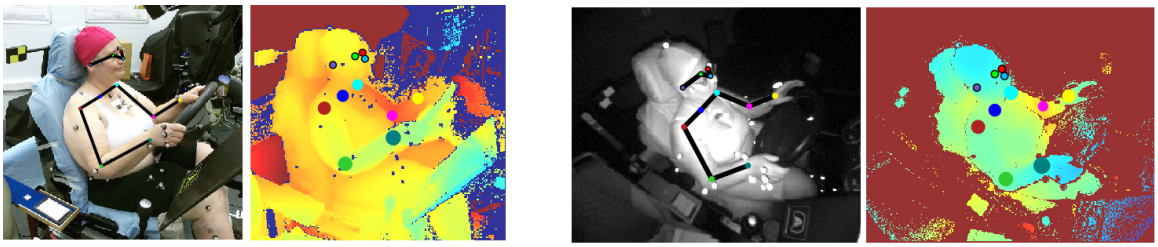


Figure 4.2. Adapted CNN architecture for the localization of driver body parts.

The adapted network was trained on the synthesized Autoconduct data (RGB, IR images along with the body segmentation labels) using the transfer learning technique (Pan and Yang 2009) and it was tested on the real driver images from the AutoConduct dataset. **Figure 4.3** shows the recognition results on the RGB image from Kinect v2 and the IR image from PMD. The Part Confidence Maps (PCMs) which are essentially Gaussian blobs surrounding the keypoint centers are superimposed on the corresponding depth images. In Chapter 5, the five keypoints including LEY, REY, LEA, REA and NOT will be used to estimate the head orientation and position. This Chapter addresses the extraction of seven upper body joint centers representing the center of the shoulders, left and right shoulder, elbow and wrist (CS, LS, LE, LW, RS, RE, RW) in 3D.



(a) Kinect v2 (b) PMD

Figure 4.3. Test of the adapted CNN on real driver RGB images (left) and IR images (right).

4.1.2 Adapted Offset Joint Regression (OJR)

The seven keypoints localized on the RGB/IR image by the adapted CNN could be projected into 3D space by using the synchronized depth image. However, the 3D keypoints still lie on

the body surface and their positions relative to the true joint centers are not consistent across different postures. The purpose of the Offset Joint Regression (OJR) model was to find the joint center locations beneath the body surface. The OJR model was originally proposed by Girshick et al. (Girshick et al. 2011) to estimate the full body 3D pose of the humans from a single depth image. The basic idea is to find the body parts that wrap the joints of interest first from the depth image, then the body part pixels will cast learned votes for the possible locations of corresponding joint centers, as illustrated in **Figure 4.4**. The algorithm behind the OJR model is briefed below, refer to (Girshick et al. 2011) for technical details.

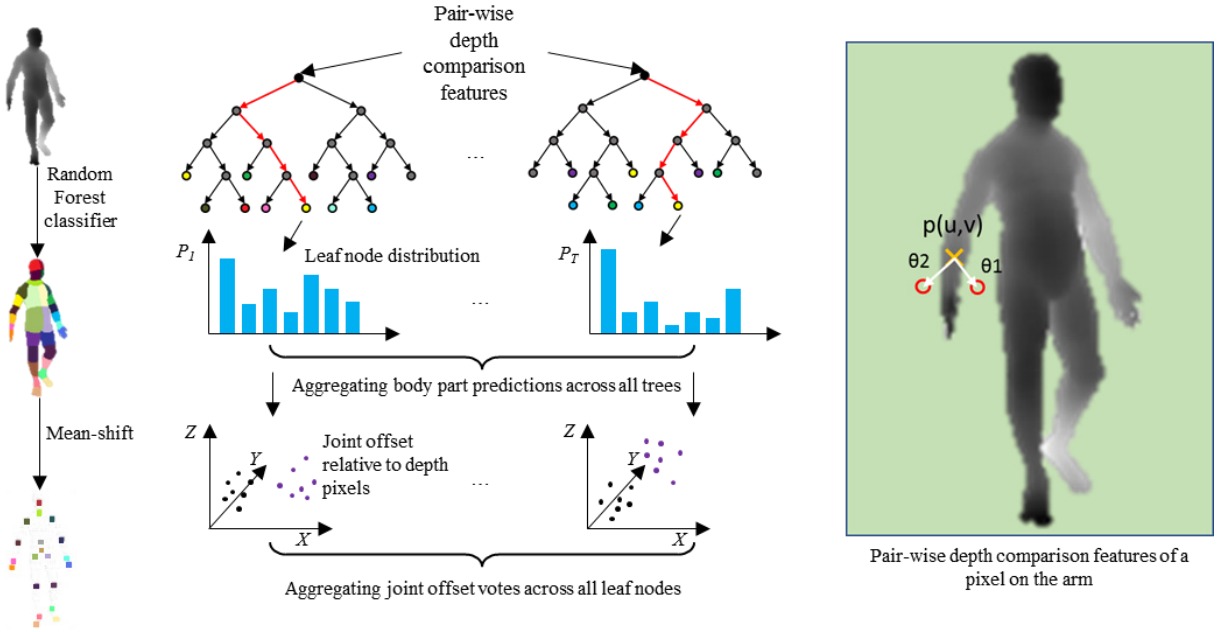


Figure 4.4. Offset Joint Regression (OJR).

Given a human body depth image, the OJR first uses a Random Forest classifier (Breiman 2001) and simple pair-wise depth comparison features (Equation (4.1)) to transform the difficult problem of pose estimation into an easier problem of per-pixel semantic segmentation.

$$f(p|\theta) = z\left(p + \frac{\vartheta_1}{z(p)}\right) - z\left(p + \frac{\vartheta_2}{z(p)}\right) \tag{4.1}$$

Where feature parameters $\theta = (\vartheta_1, \vartheta_2)$ describe 2D pixel offsets $(\Delta u, \Delta v)$ in two different directions and $z(p)$ is the depth value at pixel $p(u, v)$. The depth-weighting compensates for observing fewer pixels when imaging a person further from the camera, and ensures the following aggregation step is depth invariant. Individually, these features provide only a weak signal about which part of the body the pixel belongs to, but in combination in a Random Forest classifier they are sufficient to accurately identify the body parts.

At the leaf node reached in each tree, image patches with similar appearances are stored. Also stored at the leaf node is a learned distribution over the relative 3D offset (vote) from the projected pixel coordinate to the closest body joint of interest. Such votes can directly predict interior rather than surface positions. Then a local clustering analysis based on mean-shift (Comaniciu and Meer 2002) is employed to aggregate the votes across all decision trees in the forest for the final prediction of the joint position.

Since no complex computations are involved, this model can easily reach super real-time performance. In the original work (Girshick et al. 2011), the OJR model was trained on a large quantity of synthesized Mocap data to avoid overfitting. The OJR adapted in this work was trained on the synthesized depth images along with the corresponding ground truth body segmentations and 3D skeletons in the AutoConduct dataset. The structure of the adapted OJR was modified so that only the body parts that wrap the seven upper-body joints of interest were considered.

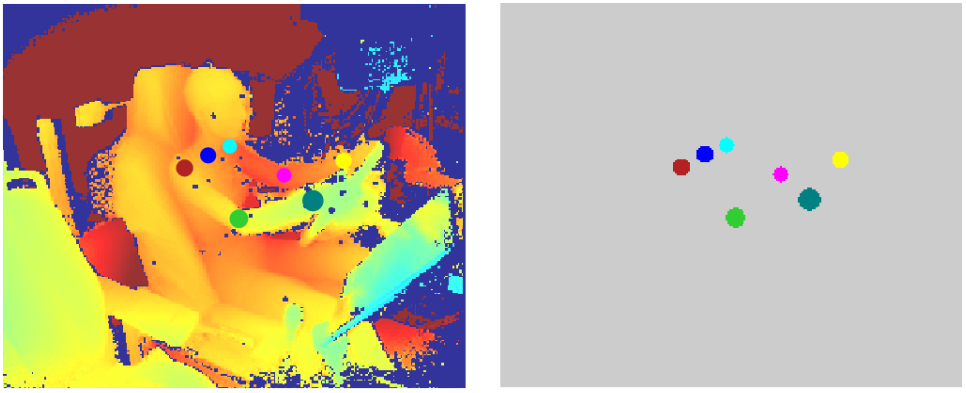
It should be noted that although the body parts were already determined by the adapted CNN in form of PCMs, the adapted OJR model needed to re-classify these body parts so that the relevant votes could be stored in the correct leaf nodes of the trees in the Random Forest classifier. To this end, this work introduced another set of relative features $f(p|r,j)$ by integrating the predetermined label information from the adapted CNN in addition to the pair-wise depth comparison features $f(p|\theta)$ for each foreground pixel p belonging to body part $j(j = 1,2,\dots 7)$. The $f(p|r,j)$ is defined as follows:

$$f(p|r,j) = \left\{ {pc_1}^{3d} - p^{3d}, {pc_2}^{3d} - p^{3d}, \dots, {pc_j}^{3d} - p^{3d}, \dots, {pc_7}^{3d} - p^{3d} \right\} \quad (4.2)$$

Where p^{3d} is the projected position of pixel p in the camera coordinate system, ${pc_j}^{3d}$ is the projected position of the body part center in the camera coordinate system. During training, this center position was projected from the geometric center of the PCM of body part j on the depth image. If a body part was not visible due to occlusion, the features related to this body part were assigned with NaN values. During testing, it was projected from the keypoint of body part j on the depth image which was available from the adapted CNN. Similarly, if a body part was not identified by CNN, the features related to this body part were assigned with NaN values.

$f(p|r,j)$ revealed the position of each body part relative to each other in the posture space, thus were strong indicators for the adapted OJR to discern the correct attribution of each body part given the PCMs from the adapted CNN. As shown in **Figure 4.5**, the classification results by the adapted OJR are highly consistent with the adapted CNN. **Figure 4.6** gives the result of upper-body posture recognition based on the images from PMD. Note that the use of these extra

features assumed that the results from the adapted CNN were always accurate. This assumption did not hold in reality because of the uncertainty of CNN itself under challenging situations, which may lead to posture recognition errors as discussed in the following.



(a) PCMs recognized by the adapted CNN (b) PCMs re-classified by the adapted OJR

Figure 4.5. Re-classification of PCMs by the adapted OJR.

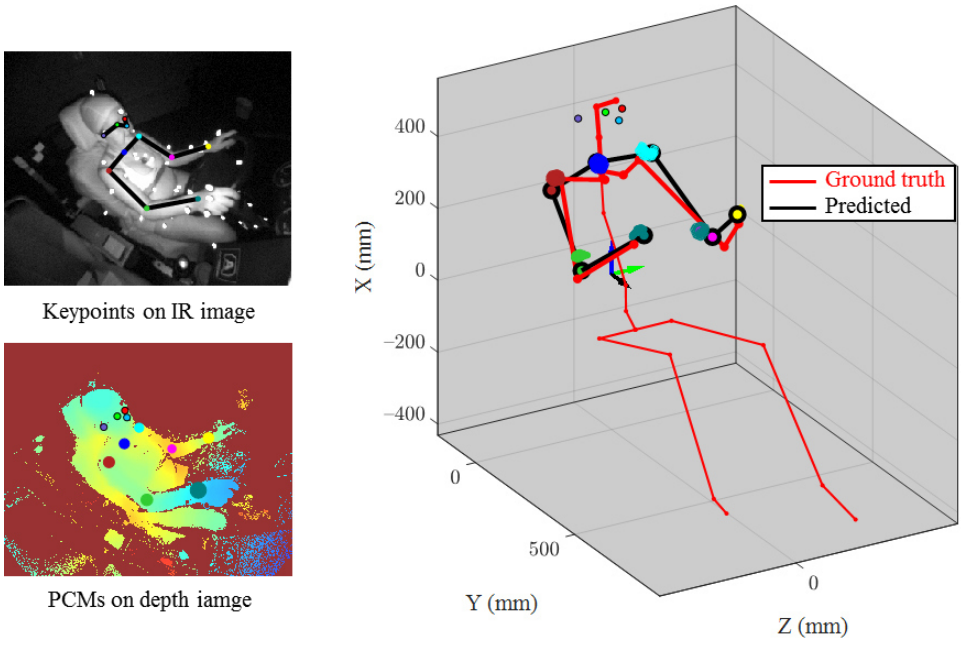


Figure 4.6. Upper-body posture recognition. Only the center of the shoulders (CS), the left shoulder (LS), the left elbow (LE), the left wrist (LW), the right shoulder (RS), the right elbow (RE) and the right wrist (RW) are considered in this Chapter. The posture is transformed into the target coordinate system as defined in Section 3.2.2.

4.1.3 Results and discussion

The proposed posture recognition method was tested on all the real image data from PMD and Kinect v2 separately. To highlight the advantage of using an in-vehicle focused dataset over a generic one, the proposed method was benchmarked against the one with a combination of

the original OpenPose network pretrained on the COCO dataset and the adapted OJR module. The detection accuracy of a joint was quantified as the percentage of data samples where the predicted joint position was within D mm of the ground truth position. In a previous work (Shotton et al. 2011), the authors set $D = 100\,mm$ to quantify the accuracy of the generic Kinect algorithm. In this work, we set $D = 50\,mm$ because we were aiming at a domain-specific posture recognition algorithm related to driving safety.

The results are given in **Table 4.1**. Using the proposed method, the mean prediction accuracy of the seven joint centers on the data of PMD and Kinect v2 are 61% and 74%, respectively, higher than their counterparts obtained by the benchmark method (55% and 69%), suggesting the utility of AutoConduct dataset for domain adaptation. The results also show that the proposed method performs better on the data of Kienct v2 than on the data of PMD. In addition to the difference in the camera placements, the image resolutions, field of views, image modalities used for posture recognition were also different. Therefore, it can be difficult to conclude that the camera placed close to the right A pillar (Kinect v2 in this work) was better than the one placed on top of the center rear mirror (PMD in this work). As the data augmentation pipeline proposed in this work allows one to simulate the camera properties and positions, it could be interesting to take advantage of this power to investigate the optimum camera positions in a car, as suggested by Plantard et al. (Plantard et al. 2015).

Table 4.1. Percentage of data samples with joint center prediction errors less than 50mm (N = 129282).

	CS	LS	LW	LE	RS	RE	RW	Mean
PMD**	81%	48%	51%	61%	60%	64%	61%	61%
PMD*	75%	42%	48%	55%	50%	61%	53%	55%
Kinect v2**	80%	69%	61%	81%	79%	74%	73%	74%
Kinect v2*	75%	63%	57%	73%	72%	71%	69%	69%

* Original OpenPose network + adapted Offset Joint Regression (benchmark method)
** Adapted CNN from OpenPose + adapted Offset Joint Regression (proposed method)

It is worth mentioning that the OJR method was originally a standalone solution for predicting the 3D posture of the human body directly from the depth image. However, it shares a crucial prerequisite with the Kinect algorithm that the background of human body is already segmented and excluded from the input depth image. This is a challenging task for the in-car situation because the driver's body is in close proximity of vehicle interior (e.g., driver seat, steering wheel, door, etc.). The problem is further escalated if the background is dynamic (e.g., driver interacts with objects, presence of passenger's body in the view, etc.). To take advantage of the learning power of the OJR, this work sidestepped this challenge by using the adapted

CNN to first identify the seven body parts of interest as foreground from an RGB image or IR image, as shown in **Figure 4.3**. We will keep looking for better solutions that rely solely on depth images in the future.

In terms of the posture recognition errors, some of the typical issues were identified:

1) High frequency vibrations of joint center locations, which happened when the adapted CNN could not track the body parts accurately.
2) Some body parts were not recognized by the adapted CNN when they were getting close to the border of the image.
3) Abnormal behavior of 3D skeleton when the PCMs on the depth image were not correctly projected to the corresponding body part due to occlusions, including self-occlusion and the occlusion imposed by objects such as a handbag. Occlusions also introduced recognition confusions of body parts.
4) Body segment lengths were not consistent across frames because both the body part localization and the joint center inference were performed in an independent manner without considering the kinematic constraints of human body.

These are the common issues observed in many applications of pose estimation. For the monitoring of the driver in a car, posture recognition accuracy is of essence. Therefore, efforts are needed to reduce the recognition errors as much as possible.

In the context of in-vehicle posture monitoring, some special prior knowledge could be used to reduce recognition errors. For instance, the lengths of body limbs can be considered constant. Physical activities of the driver are restricted by the seat and vehicle interior. For instance, the standard driving posture with the hands on the wheel and right foot on the gas pedal is highly constraint and can be predicted correctly. The incorporation of such prior knowledge into the posture recognition model may allow one to exploit the temporal consistency and the geometric constraints. In addition, the pressure sensors on the driver seat may be able to provide extra information useful for improving the posture recognition accuracy based on depth camera, which leads to the investigations in Section 4.2.

4.2 Posture classification based on pressure sensors

Due to the sensing mechanism of pressure sensors, the posture recognition in this context is usually transformed into a classification problem. Which posture classes can be possibly inferred by body pressure distribution (BPD)?

4.2.1 Definition of posture classes

A cursory review of related works demonstrated that the driver posture classes to be predicted varied from study to study. In the study by Shin et al. (Shin et al. 2015), four trunk postures (normal sitting posture, trunk tilted to left, trunk tilted to right and trunk inclined forward) were identified. Similarly, Vergnano and Leali (Vergnano and Leali 2019) also focused on the trunk posture to determine whether the driver is out of position. Ding, Okabe and others attempted to predict driver's head postures and hand positions (Ding, Suzuki, and Ogasawara 2017; Okabe et al. 2018). The results showed that the prediction accuracy of the head posture and hand position was inadequate. Based on the observation of the pressure data in the AutoConduct dataset, it was found that BPD patterns were sensitive to the movements of driver's trunk and feet while not much sensitive to the head or hand motions alone. Therefore, the driver postures here are mainly described by trunk and feet positions. The estimation of driver's feet positions will be particularly addressed in chapter 6. This Section thereby focuses on the recognition of driver's trunk posture.

To identify the typical trunk postures for classification, a trunk coordinate system was built based on the reconstructed posture (**Figure 4.7**). The posture class definition scheme is illustrated in **Figure 4.8**. The trunk posture was characterized by three trunk angles (rotation, inclination, and lateral tilt) with respect to the standard trunk position (TP0) at the beginning of each task. By analyzing the trunk angles, four additional trunk posture classes (TP1 to TP4) were defined with noticeable deviations from TP0 in terms of rotation, inclination, or lateral tilt.

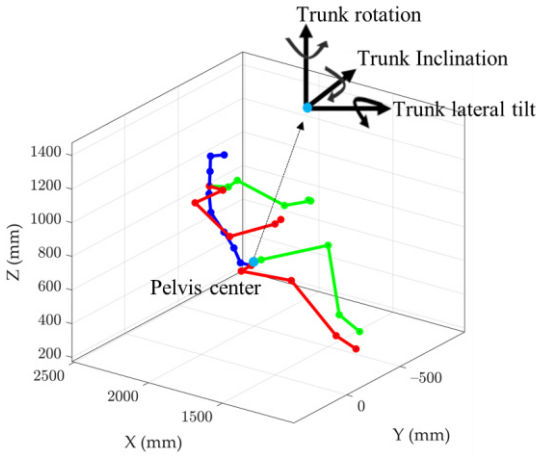
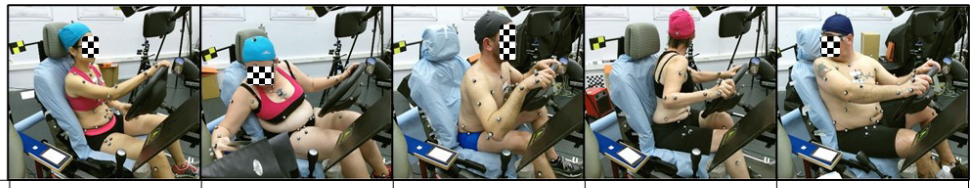


Figure 4.7. Driver posture after reconstruction using Mocap data. A coordinate system located at the driver's hip center was used to describe different trunk positions.



Typical trunk postures					
Posture class	TP0	TP1	TP2	TP3	TP4
Trunk rotation	[–5°, 5°]	[20°, 50°]	[–20°, 10°]	[–45°, –25°]	[25°, 45°]
Trunk inclination	[–3°, 3°]	[0°, 20°]	[20°, 45°]	[5°, 20°]	[5°, 20°]
Trunk lateral tilt	[–2°, 2°]	[5°, 25°]	[–5°, 20°]	[–15°, 10°]	[–10°, 15°]

Figure 4.8. Definition of posture classes for driver's trunk.

To reduce data redundancy and overfitting risk of the classifiers, an intra- and inter-motion filter was applied to remove similar postures of the same participant. Two trunk postures were regarded as different if one of the three trunk angle differences was higher than 3°. Finally, 3999 trunk postures were extracted from the experimental dataset.

4.2.2 Feature extraction and evaluation

According to the measured contact areas, 42 by 44 cells for both seat pan and backrest mats were retained. In one preliminary study (Zhao et al. 2020a), the BPD pairs were converted into images to train a deep learning model for posture classification. Deep learning model is very good at automatically extracting high-level features from the raw data, yet the results showed that it was prone to overfitting on driver's BPD data. The main reason is that the BPD pattern is not only shaped by driver's posture, but also influenced by body size and driver's seating preferences (seat position, seat pan angle, backrest angle, etc.). With these confusions, the deep learning method clearly failed to figure out the most relevant features for posture recognition. To achieve a robust solution, a feature engineering process based on domain expertise therefore is needed.

Inspired by the work by (Mergl et al. 2005) for studying seating comfort, the whole backrest (B) and seat pan (S) pressure mats were segmented into 12 (B1-B12) and 8 (S1-S8) sub areas (**Figure 4.9**). Based on this segmentation scheme, two-hundred relevant pressure features were defined (**Table S2**). They fell into three categories: Contact Area Proportion (CAP, the proportion of cells activated by body contact on each pressure mat), Centre of Pressure (COP) of an area and Pressure Ratio (PR, the ratio between the sums of pressure from different sensing

areas). These features covered critical information that can be extracted from BPD and they were specifically constructed to account for driver's body part motions by observing the experiment data.

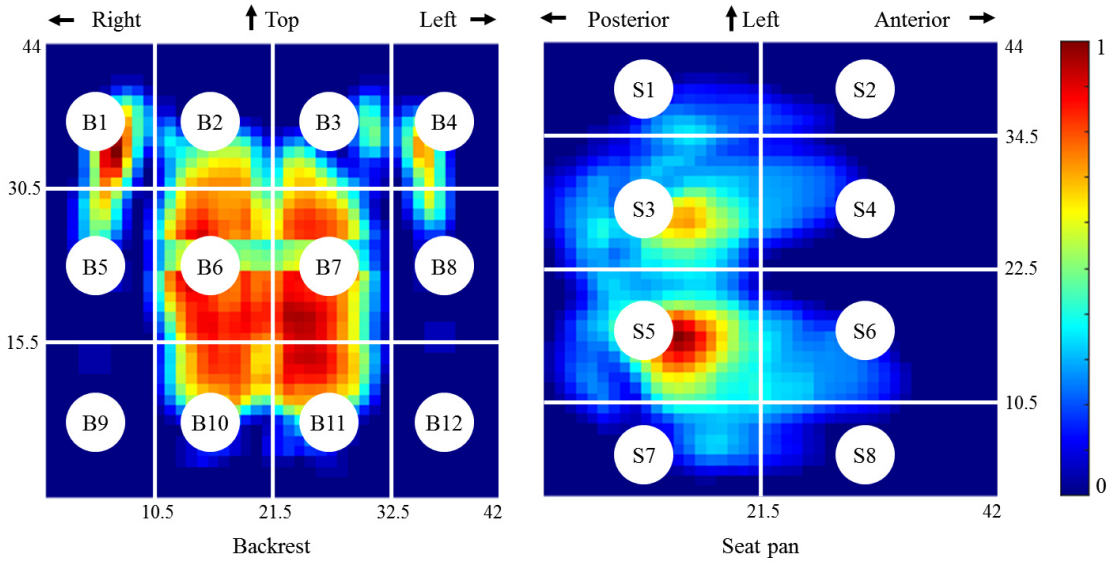


Figure 4.9. Pressure mats segmentation. The BPD on the backrest (left) and seat pan (right) were normalized by the respective peak pressure.

It should be noted that some of the features (**Table S2**) were correlated with each other. For example, a more forward trunk inclination will reduce contact area proportion on the backrest and lower the center of pressure at the same time. To better understand the relationship between BPD and driver postures, the features needed to be evaluated and selected. Feature selection is a major research area of machine learning. By selecting a subset of feature vectors, machine learning models can be trained more efficiently, and better performance can be obtained.

In another preliminary study (Zhao et al. 2021a), the classifiers trained on relative features with respect to the standard driving postures have shown better prediction results than those trained on absolute features. The relative values were therefore used to evaluate the importance of the feature candidates.

The feature evaluation was performed under the framework of RF Out-of-bag (OOB) error estimation (Breiman 1996, 2001), as proposed in a previous work (Zhao et al. 2020b). Here we briefly review the feature evaluation method.

During training, the RF classifier randomly selects a subset of features and draws N out of N observations in the data set with replacements for growing each decision tree. The left-out observations, approximately one third, are called *Out-Of-Bag (OOB)* observations. By randomly permuting OOB observations across one parameter at a time, The larger the OOB

error is, the more important the feature is for postural classification. The importance of a feature p is quantified by Equation (4.3).

$$Importance_p = \bar{d}_p / \sigma_p \tag{4.3}$$

Where $\bar{d}_p$ and σ_p are the mean and standard deviation of increased OOB error (d_p) for all the decision trees when permuting the feature p.

Once the features were ranked by their importance, the classifiers were trained with a different combination of features, starting from the first two most important until all features were included by adding a less important one each time.

4.2.3 Classifier and evaluation metric

The RF classifier was again used to classify the trunk postures. In order to objectively evaluate the generalization capability of the classifier, LOO cross-validation was performed, where the data of all participants except one were used for training and the excluded one for testing. Iterating this procedure for each participant resulted in the 23-fold cross-validation.

To quantify and evaluate the classification accuracy, the *F*1 *score* (a harmonic mean of *precision* and *recall*) was calculated from the confusion matrix. The *precision (PREC)* is referred to as the proportion of data samples that the classifier predicts true actually are true (Equation (1)), while the *recall (REC)* is referred to as the ability to predict the true results given the data samples of a specific class (Equation (2)). As opposed to the *accuracy* which is simply calculated as the number of all correct predictions divided by the total number of data samples, the *F*1 *Score* (Equation (4.6)) takes both false positives and false negatives into account and is, therefore, usually more useful, especially when the classes are unevenly distributed.

$$PREC = \frac{TP}{TP + FP} \tag{4.4}$$

$$REC = \frac{TP}{TP + FN} \tag{4.5}$$

$$F1\ Score = \frac{2 \cdot PREC \cdot REC}{PREC + REC} \tag{4.6}$$

where *TP* denotes the true positives, *FP* denotes the false positives, and *FN* denotes the false negatives.

4.2.4 Results and discussion

Using the average F1 score across all the five classes as a proxy for accuracy, the best combination of pressure features for the RF classifier was determined. **Figure 4.10** shows that the classifier achieved the highest average F1 Score (0.91) when only 27 important features

were used, better than other similar studies (Ding, Suzuki, and Ogasawara 2017; Okabe et al. 2018; Shin et al. 2015) which ranged between 76.8% and 80.5%. The confusion matrix is given in **Figure 4.11**. The accuracy was notably compromised when more than 40 features were involved. The selected important features are given in **Table S3**.

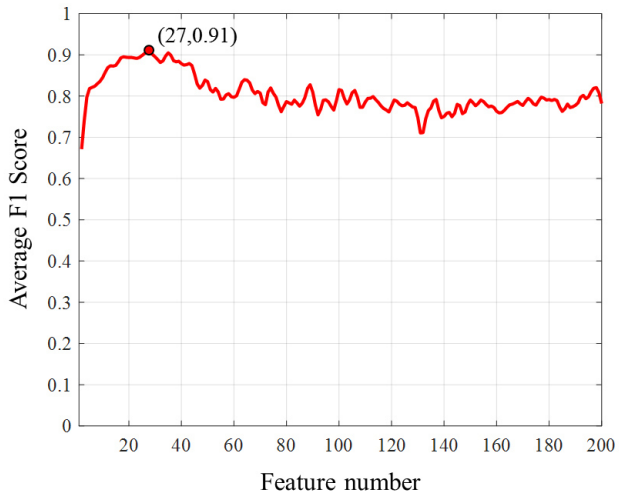


Figure 4.10. Average F1 Score vs feature number for trunk posture classification.

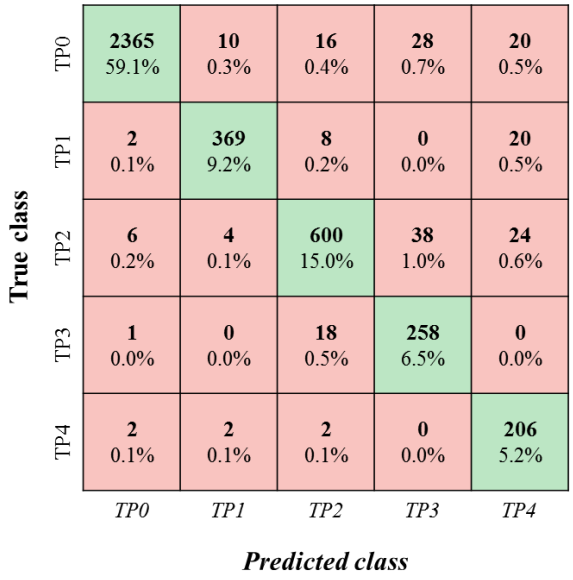


Figure 4.11. Confusion matrix of the classification results from the 23-fold LOO cross-validation tests using RF (N=3999). The green downwards diagonal showed the number and proportion of correct detection cases for each class.

In addition to the RF classifier, other supervised classifiers including support vector machine (SVM) (Chang and Lin 2011), multilayer perceptron (MLP) (Hastie, Tibshirani, and Friedman 2009), k-nearest neighbors (k-NN) (Altman 1992), and naïve Bayes (NB) (Hastie, Tibshirani, and Friedman 2009) were also tested in the previous work (Zhao et al. 2021b).

Results showed that none of these methods outperformed the RF classifier on this dataset, suggesting that RF classifier is a suitable choice to deal with the pressure features.

When it comes to body pressure monitoring system, one important issue is related to its performance on the data from drivers with different body sizes. In previous studies (Bourahmoune and Amagasa 2019; Ma et al. 2017), the researchers have tested their monitoring systems on the subjects grouped by different BMI ranges. To investigate the effects of participant's BMI on monitoring performance using our proposed method, a regression model between classification accuracy and BMI was built. However, no clear relationship was found, suggesting that the proposed method is robust with respect to body size variations. This could be, in large part, attributed to the relative features we used, which could reduce the effects of body size, as well as seat configuration, on the BPD.

The main limitation of this method is that the RF classifier was built on relative pressure features which required a reference sitting posture for initialization. Interestingly, if the RF classifier was trained on the absolute pressure features, it was found that the standard trunk posture for driving could be recognized with a high *F1 score* of 0.98, though the average value was much lower (0.83). This could be a way to identify the reference trunk position without imposing a manual calibration procedure.

The RF classifier also provides a theoretical framework for determining class scores (membership probabilities), which allowed one to continuously predict driver postures in a dynamic manner, see **Figure 4.12** for a few examples. By tracking the class scores overtime, the RF classifier generalized well on new data, including not only the typical trunk posture classes but also the intermediate postures that have been excluded from training dataset.

In summary, pressure sensors based classifiers could provide reliable qualitative inference about driver's trunk posture. In Section 4.3, this trunk postural information will be incorporated into the posture correction framework to reduce the unreliable postures recognized by a depth camera.

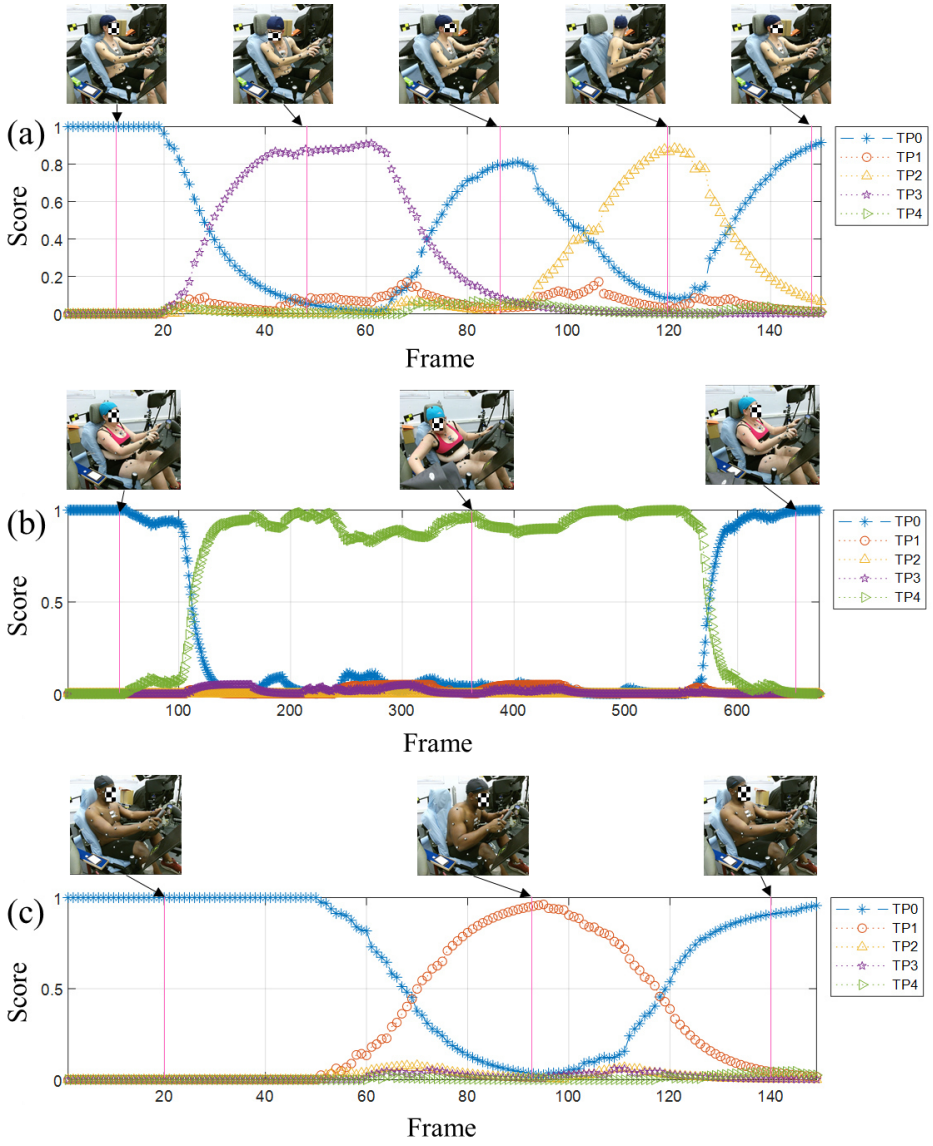


Figure 4.12. Continuous prediction of trunk postures. The probability scores of trunk classes are visualized as a function of time.

4.3 Posture correction based on motion databases

Posture correction is an endeavor to obtain natural-looking and accurate results from low-dimensional or noisy measurement. One solution is to use pre-recorded human motions to constraint the solution space. For instance, Chai and Hodgins (Chai and Hodgins 2005) proposed a method to reconstruct the full posture based on a small set of reflective markers on the body. In this work, the pre-recorded accurate posture samples that were a close match to the input signals were selected to construct a locally linear posture space, which was then used as prior knowledge for solution seeking. In a study by Shum et al. (Shum et al. 2013), reliability measurement of noisy Kinect skeleton data was incorporated into the motion database query in

order to obtain relevant postures that were also kinematically valid. In a further study by Plantard et al. (Plantard, H. Shum, and Multon 2017), a structure called Filtered Pose Graph (FPG) was proposed to model the pre-recorded motion database. Such a structure can benefit the efficiency of the posture selection process as well as the quality of posture reconstruction.

Inspired by these studies (Chai and Hodgins 2005; Shum et al. 2013; Plantard, H. Shum, and Multon 2017), this Section proposes a correction framework based on pre-built driver motion databases to correct the incomplete and corrupted driver upper-body skeleton data predicted by the depth camera. An overview of the framework is presented in **Figure 4.13**. Offline, Motion retargeting is performed to further enrich the real driver posture variations. The driver motion databases are established using the reconstructed postures from the extended dataset and they are organized by the FPG structure. They are coded by driver's seating position (P) and body size (L), and conditioned on five different trunk posture classes consistent with Section 4.2.1. The posture samples are decomposed by body part (shoulder, left arm and right arm). Online, the relevant motion database sets are first selected based on the seating position estimated by pressure sensors and body segment lengths estimated from the recognized posture from the depth camera. The reliability of the joint positions estimated from the depth camera and the trunk positions predicted by the pressure sensors are evaluated to select relevant posture samples from the relevant motion databases with which reliable postures are then synthesized. Finally, Kalman filters are applied to further improve the temporal consistency of the synthesized posture.

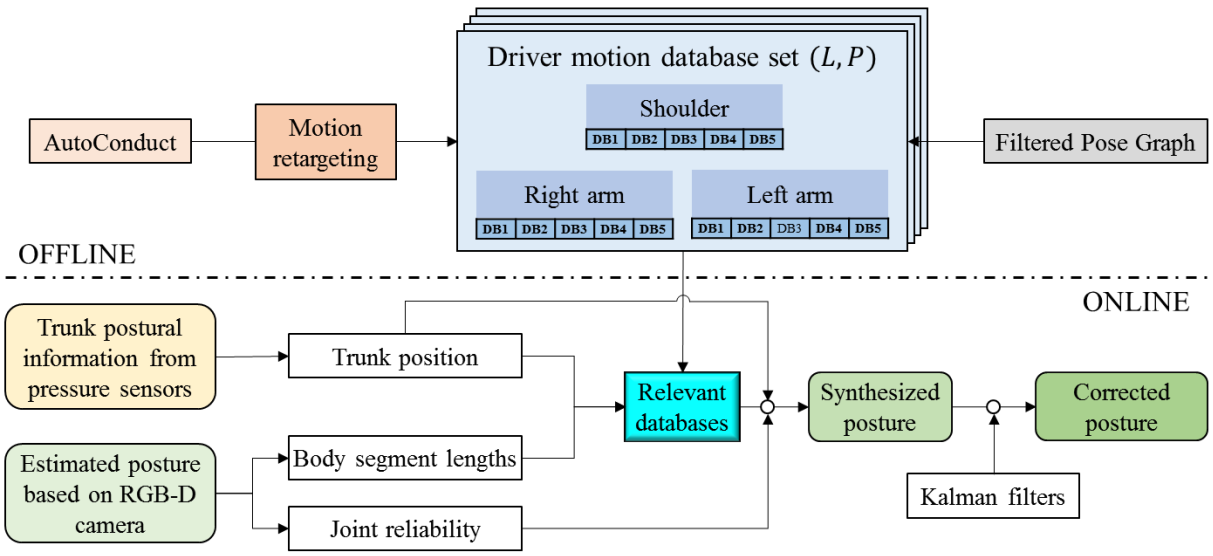


Figure 4.13. Motion database based posture correction framework.

4.3.1 Motion databases

The purpose of a motion database is to provide posture candidates similar to the one driver performs during a task for reducing prediction error. Depending on age, gender, anthropometric characteristics, and seating preference, there is a large range of variation in driver postures. In addition, a same action/task can be executed in different ways by different persons or even by a same person. Therefore, a database with large range of posture variations therefore is needed.

Based on the reconstructed driver postures from the AutoConduct dataset, where 23 drivers and 42 in-vehicle tasks were involved, motion retargeting was performed in this work to further enrich the posture variations thanks to the RPx motion simulation module. Specifically, the motion of one participant in the experiment was used to drive the RPx skeletons of the other 22 participants. This led to a motion dataset 22 times larger than the original one. Theoretically, this motion retargeting method can generate accurate posture data for any drivers provided that their body dimensions are known. Each posture was represented with a set of joint center positions. As this Chapter focuses on driver's upper-body, only the positions of the seven joints were used to build the motion databases.

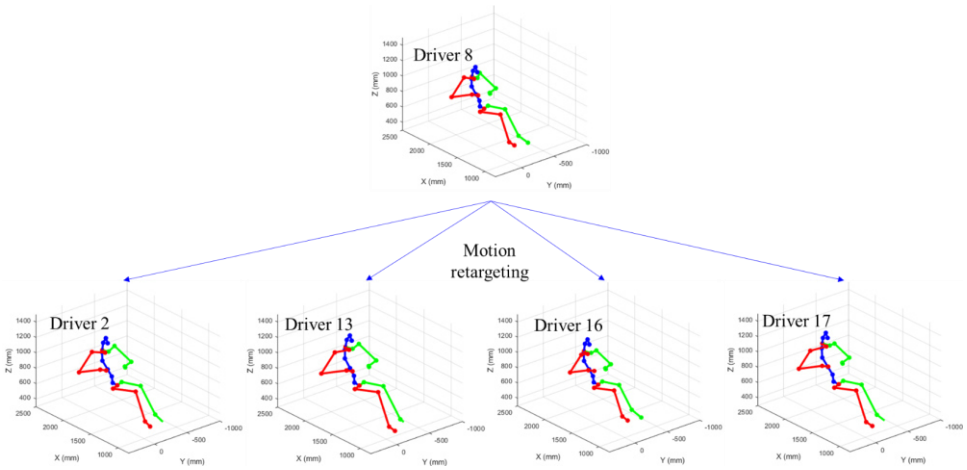


Figure 4.14. Motion retargeting in RPx.

The FPG was computed by using a procedure as illustrated in **Figure 4.15**. For a more detailed presentation of the FPG, refer to Plantard et al. (Plantard, H. Shum, and Multon 2017). This Section briefly review the key steps used for the computation.

First, an intra- and inter-motion filter was applied to remove redundant similar postures of each driver. The similarity between two postures p_a and p_b was defined as the maximum Euclidian distance of all the joint positions:

$$d(p_a, p_b) = max_{j=1\ldots N} \|p_a(j) - p_b(j)\| \tag{4.7}$$

Where N is the total number of joints. If $d(p_a, p_b) < \varepsilon_1$, one of the postures was discarded. The remaining postures were called Filtered Nodes. The main purpose of this step was to control the density of the motion database so that the database query could retrieve samples with necessary variations.

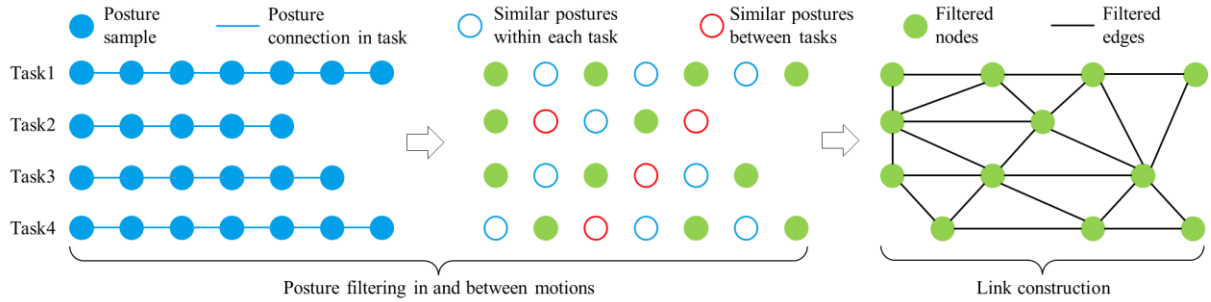


Figure 4.15. Filtered Pose Graph (FPG) construction.

In the second step, the Filtered Nodes were reconnected by evaluating the joint position difference $d(p_a, p_b)$ to compute the Filtered Edges, leading to the final Filtered Pose Graph. There was connection between two nodes if $d(p_a, p_b) < \varepsilon_2$. The Filtered Edges could synthesize new artificial motions while ensuring continuity. Most importantly, they can speed up the posture selection process by tracking the previously selected nodes, as shown in **Figure 4.16**. Consequently, this method can be applied to large-scale databases with real-time performance.

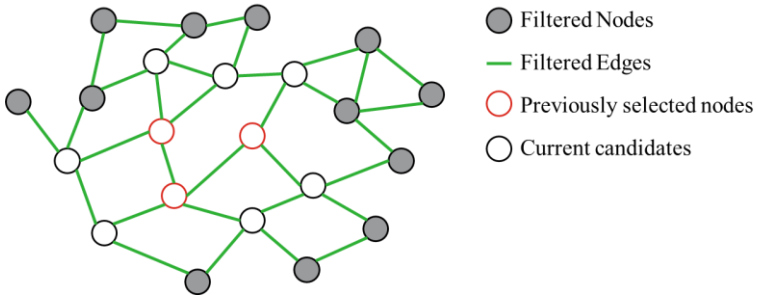


Figure 4.16. Posture selection process based on FPG.

In the studies by Shum et al. (Shum et al. 2013) and Plantard et al. (Plantard, H. Shum, and Multon 2017), the similar postures extracted from a motion database were used to create a natural posture space to synthesize the full body posture as a whole. The limitation of this approach was that if only a part of the body was incorrectly tracked, the posture of all the correctly tracked body parts may be compromised by the overall correction strategy. For the problem of driver posture monitoring, the recognition errors usually take place on certain body parts depending on driver's action and the view angle of the camera. In a preliminary study

(Zhao et al. 2018), driver upper-body motion databases were individually built for each specific driver action so that more relevant posture samples could be provided. However, this method required that the driver action class could be reliably estimated before, which is challenging due to the high dimensional nature of human body movement and the noisy input data from Kinect.

With these considerations, this work instead proposes to decompose the upper-body posture samples by regional body parts, namely, the shoulder, the left arm and the right arm. Each regional part consists of three joints, and they are connected by the left shoulder or the right shoulder, as illustrated in **Figure 4.17**. These postures were also transformed into the target coordinate system defined in Section 3.2.2.

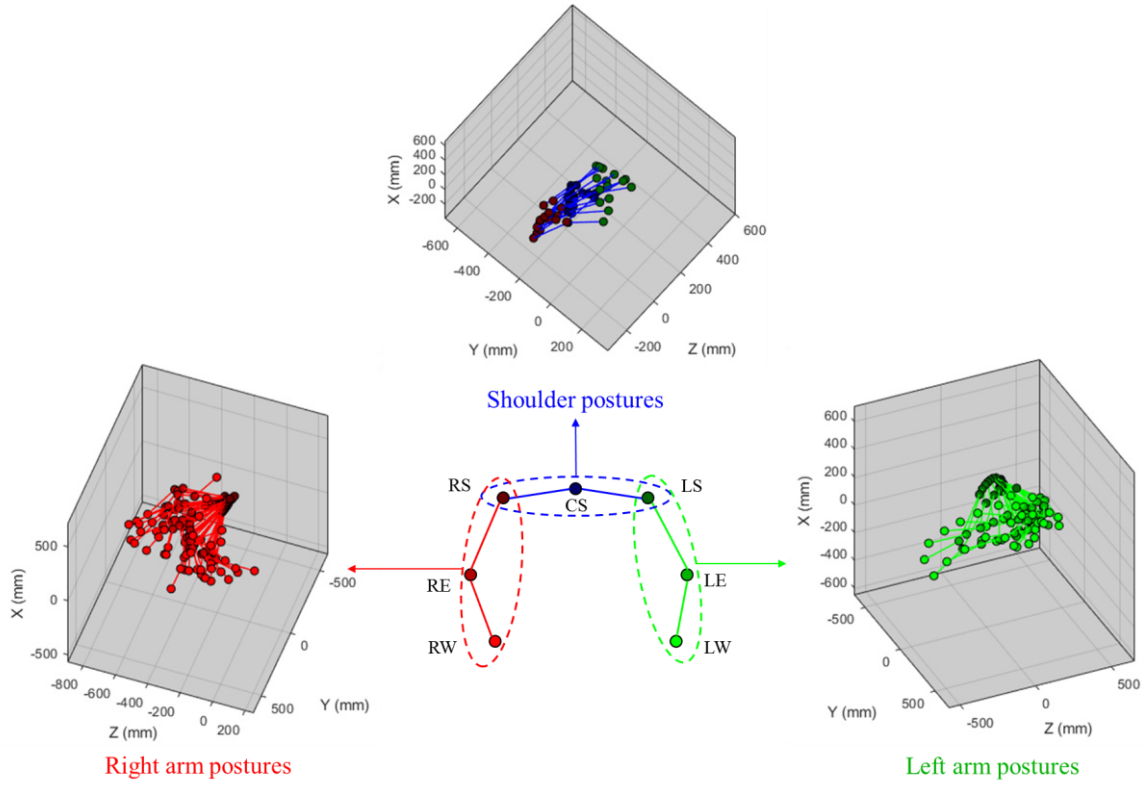


Figure 4.17. Motion databases for regional body parts.

To provide more relevant candidates for posture correction, we further conditioned the posture samples of all the three regional body parts on the five trunk posture classes as defined in Section 4.2.1, which could be predicted by using pressure measurements. It was observed that some occlusions were introduced by trunk orientation. For instance, when driver rotated the trunk to left, the left shoulder and the left arm were occluded in the view of the camera on the right A-pillar. It is therefore tempting to see if prediction performance can benefit from the sensor fusion.

Following the method explained above, we built five conditional motion databases with the FPG structure for each regional body part to form the motion database group. The three motion database groups then constitute the motion database set for each driver. Finally, twenty-three sets of motion databases were obtained. For the convenience of the optimization of the FPG structure, we unified the parameters ε_1 and ε_2 across different drivers. They were empirically set so that the motion databases have a reasonable density while the filtered Nodes from all the tasks can be connected without discontinuity. **Table 4.2** gives the values of these two parameters.

Table 4.2. Parameter configurations for building the motion databases.

Body part	$\boldsymbol{\varepsilon_1}$	$\boldsymbol{\varepsilon_2}$
Shoulder	20 mm	70 mm
Left arm	100 mm	200 mm
Right arm	100 mm	200 mm

To retrieve similar posture candidates, it is important to first identity the relevant database sets that have similar body size with the test driver. To compare the recognized posture with those in the motion databases, they have to be located in the same position. These steps could be manually performed as done by Shum and Plantard (Shum et al. 2013; Plantard, H. Shum, and Multon 2017), yet this work attempted to automatically estimate these latent variables by analyzing the available information during run time. To this end, the motion databases of each participant was coded by body size (L) and seating position (P), where L is the length between center of shoulders and wrist (assume the driver body has left-right symmetry) and P is the position of the center of shoulders (CS) when the participant is in the standard driving posture. **Figure 4.19** illustrates the structure of a motion database set.

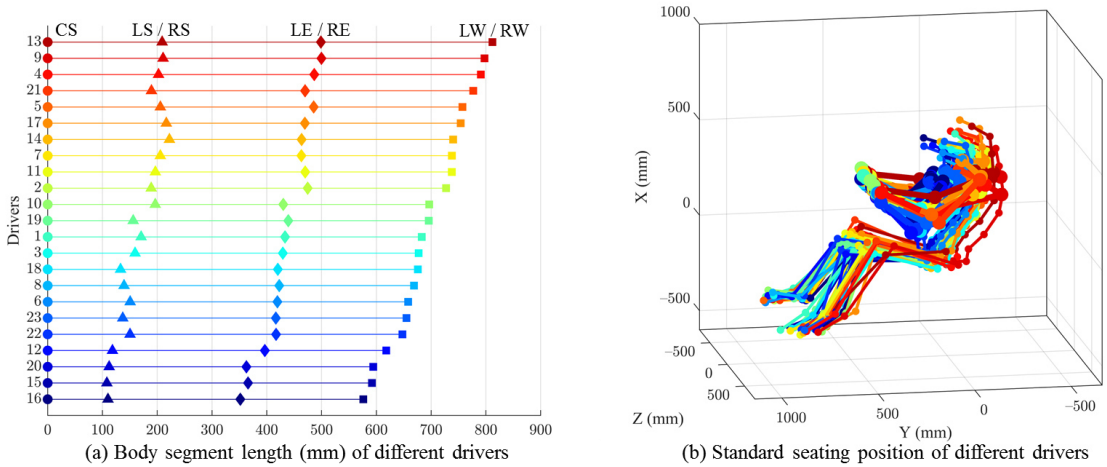


Figure 4.18. Body size and standard seating position of different drivers.

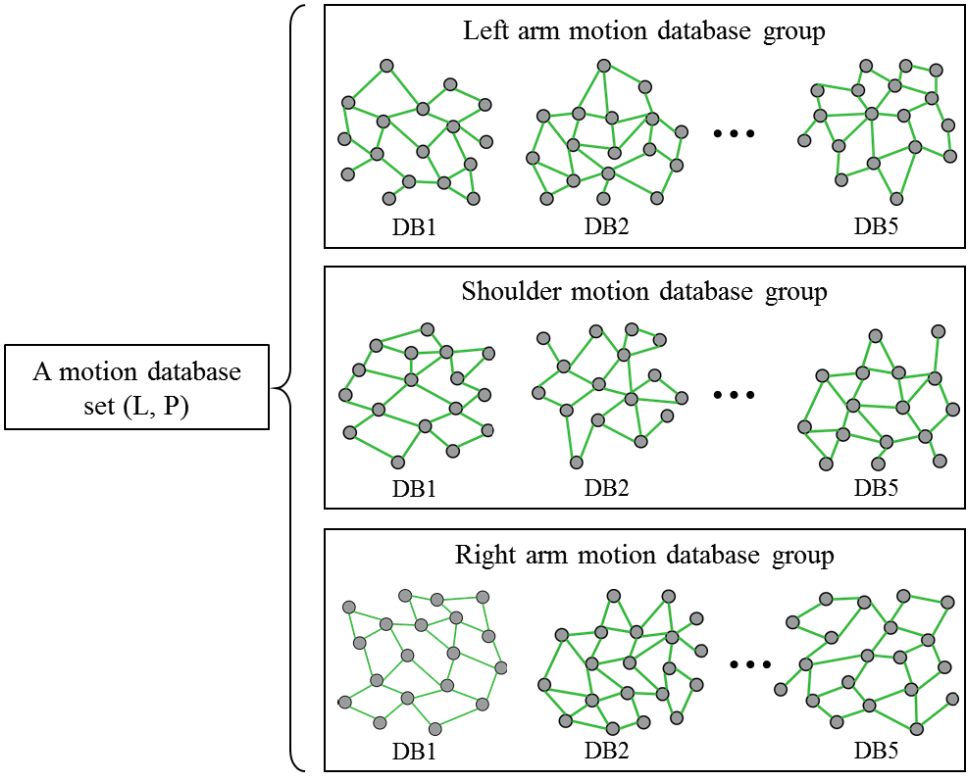


Figure 4.19. Structure of the motion database set for each driver. The motion databases DB1, DB2, DB3, DB4 and DB5 are conditioned on the trunk posture classes TP0, TP1, TP2 TP3, and TP4, respectively.

4.3.2 Estimation of body size and standard seating position for driving

To reliably estimate the body size from the recognized driver posture, the body segment lengths from pre-recorded true postures are used as reference. By analyzing the existing data included in the AutoConduct dataset, driver's shoulder length (half of the shoulder breadth) is in the range [108, 222] (mm), upper arm in [233, 290] (mm), forearm in [220, 313] (mm). They were similar to the driver anthropometric statistics collected by Park et al. (Park et al. 2016) though we had only 23 drivers involved in this work. During run time, the prior belief about the lengths of the body segments was aggregated and refined across a video sequence until the estimated body segment lengths converged in the corresponding ranges. After testing, we found that this process was efficient and incurs almost no additional cost. Then the refined belief (L_{est}) was recorded and used to find the relevant sets of motion databases.

In terms of the estimation of standard seating position, two conditions were tested simultaneously. The first one was the same as used for estimating the body size L as explained above. The second was the classification results from pressure sensors. If L_{est} was determined while the predicted class by pressure sensors was in the standard trunk position (TP0), then the position of CS was recorded as P_{est}.

These automatic estimation approaches were tested by using the Leave One Out (LOO) cross-validation. The mean estimation precision across 23 drivers on Kinect V2 (PMD) for L and P was 24.0 (26.4) mm and 23.7 (29.8) mm respectively, thus was considerable.

4.3.3 Reliability measurement

To retrieve relevant posture samples from the motion databases, the input postural information needs to be evaluated. This is because incorrectly tracked body parts may guide the database query to find irrelevant candidates. This section is based upon the method adapted from Shum et al (Shum et al. 2013) where three terms including tracking state, behavior and kinematics were used to evaluate the reliability of the recognized joint j at current frame f.

The tracking state term ($Wt_j(f) \in [0.0, 1.0]$) is returned by the adapted CNN model, indicating the reliability of the recognized keypoints estimated by the model itself. This value is low when the model is less confident about the prediction. If the joint was not recognized, $Wt_j(f)$ equals to 0.

The behavioral term ($Wb_j(f) \in [0.0, 1.0]$) refers to the abnormal behavior of the tracked joints in 3D space, such as the high frequency vibrations due to the uncertainty of recognition model and the projection errors because of occlusion. It was defined as the normalized joint displacement with respect to last frame:

$$Wb_j(f) = \begin{cases} \frac{D_j(f)}{D_j^{max}} & if\ Wt_j(f) \neq 0 \\ 0 & if\ Wt_j(f) = 0 \end{cases} \tag{4.8}$$

Where D_j^{max} is the maximum displacement of joint j between temporal frames observed on the AutoConduct real image dataset.

The kinematics term ($Wk_j(f) \in [0.0, 1.0]$) aims to evaluate the consistency of body segment lengths. To quantify the length's variation $d_i(f)$ of body segment i, the estimated body size dimensions are used as reference. Assume joint j is connected to m other joints, the kinematics reliability is defined as:

$$Wk_j(f) = 1 - \frac{\sum_{i=1}^{m} d_i(f)}{m} \tag{4.9}$$

$$d_i(f) = \min\left(\frac{|l_{rec}(i,f) - l_{est}(i)|}{l_{est}(i)}, 1\right) \tag{4.10}$$

Where $l_{rec}(i,f)$ is the segment length calculated from the recognized posture, $l_{est}(i)$ is the estimated segment length.

Finally, the reliability rate of joint j is defined as:

$$W_j(f) = min\ (Wt_j(f), Wk_j(f), Wb_j(f)\ ) \tag{4.11}$$

Notice that the calculation of the behavioral term ($Wb_j(f)$) needs to reference the recognized posture in history. If joint j in the previously recognized posture is actually not reliable, the reliability of $Wb_j(f)$ will hence be compromised. To compensate this risk, a sliding window with one frame length is used to track the well-recognized recent joint position (x, y, z) for reference. The window starts with a recognized joint position, where the $min\ (Wt_j(f), Wk_j(f))$ of the joint is bigger than an empirical threshold (0.8). If the maximum Euclidian difference between the current predicted joint position and the one in the window is smaller than a threshold (100 mm), the values in the window will be replaced by the current recognized joint position to be used as reference for evaluating the $Wb_j(f)$ for the next frame. Otherwise, the window will remain unchanged. It will be re-initiated when the $min\ (Wt_j(f), Wk_j(f))$ is again bigger than 0.8 and the update procedure will carry on, so on and so forth. This window is used according to the fact that body motion must be continuous.

4.3.4 Posture selection and synthesis

To test the generalization capability of this correction framework, LOO cross-validation is performed, where the test driver's motion databases are intentionally excluded. This is because the new driver's body size dimensions are difficult to be accurately obtained in reality. The relevance of the databases from the other drivers is identified by evaluating the difference between L of the database and the L_{est} estimated from the recognized posture. If the difference is smaller than a threshold (ε_l) the database will be regarded as being relevant. In this work, ε_l was set to 50 mm, equal to the threshold used for evaluating the accuracy of posture recognition, which led to 1.2 relevant database sets on average for the test drivers. Afterwards, the selected relevant motion databases are globally translated to the estimated position P_{est} in the target coordinate system and they are merged by body part (each has three joints) and five trunk posture classes (TP0, TP1, TP2, TP3 and TP4), see **Figure 4.20**.

Take the shoulder posture database in DB1 as example, the similarity (S) between the recognized posture ($pos^r(f)$) and a node ($pos^d(n)$) stored in the corresponding motion database is measured by a Weighted k-NN algorithm (WkNN, Equation 4.12). The use of the reliability rate $W_j(f)$ allows the similarity to rely more on the correctly tracked joints. The votes in **Figure 4.20** are averaged joint positions (CS, LS and RS) from the K similar shouler postures. The same method applies to the three joints on the left arm (LS, LE, and LW) and right arm (RS, RE and RW).

$$S(pos^r(f), pos^d(n)) = \sqrt[2]{\sum_{j=1}^{3} W_j(f) \left\| pos_j^r(f) - pos_j^d(n) \right\|^2} \tag{4.12}$$

Note that when selecting similar body part postures, a whole database search is only needed at the first time step or when the sliding window is reinitiated. In the following time steps, the database search will be limited to the nodes that are connected to the previously selected ones by using the Filtered Edges.

Once the votes for each body part is obtained, they are assembled to form the full upper-body posture by using the left shoulder (LS) or the right shoulder (RS) as junctions. The final positions of LS and RS are averaged across the votes from the related body part databases.

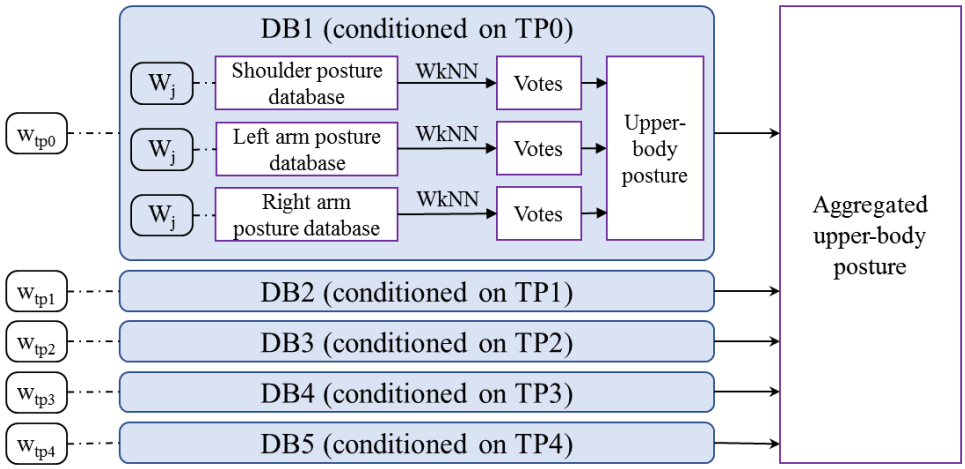


Figure 4.20. Posture selection process within a motion database set.

The posture synthesis procedure is iterated through all of the conditional datasets to obtain the upper-body postures. Finally, they are aggregated by using the trunk class probabilities returned by pressure sensors. If more than one relevant database sets are selected for the test driver, the posture selection process will be performed independently, and the aggregated upper-body posture from different database sets will be averaged.

4.3.5 Motion smoothing

The synthesized postures could still exhibit some small noise like temporal spikes during motions. To deal with this problem, a Kalman filter (Kalman 1960) is applied for each joint. Kalman filter is a recursive method that can be used to solve the problem of linear filtering of time series data (Bishop and Welch 2001).

Assume that the driver body movement was linear and the noises followed the normal distributions, the state of the joint was expressed as a combination of the current position (x, y, z) and velocity (v_x, v_y, v_z).

$$\hat{x}_k = [x, y, z, v_x, v_y, v_z] \tag{4.20}$$

The transition matrix was defined by Equation (4.21).

$$A = \begin{matrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix} \tag{4.21}$$

Where the Δt is the interval time between the previous and the current frames. The measurement noise covariance R for each joint was measured prior to the filter operation. The process noise covariance Q was empirically estimated.

Figure 4.21 shows the prediction of the right wrist trajectory when the driver drinks coffee. As it can be seen, most of the prediction errors from the recognized posture can be eliminated by posture synthesis based on the motion database. The jitter movements of the synthesized posture can be further smoothed out by using a Kalman filter.

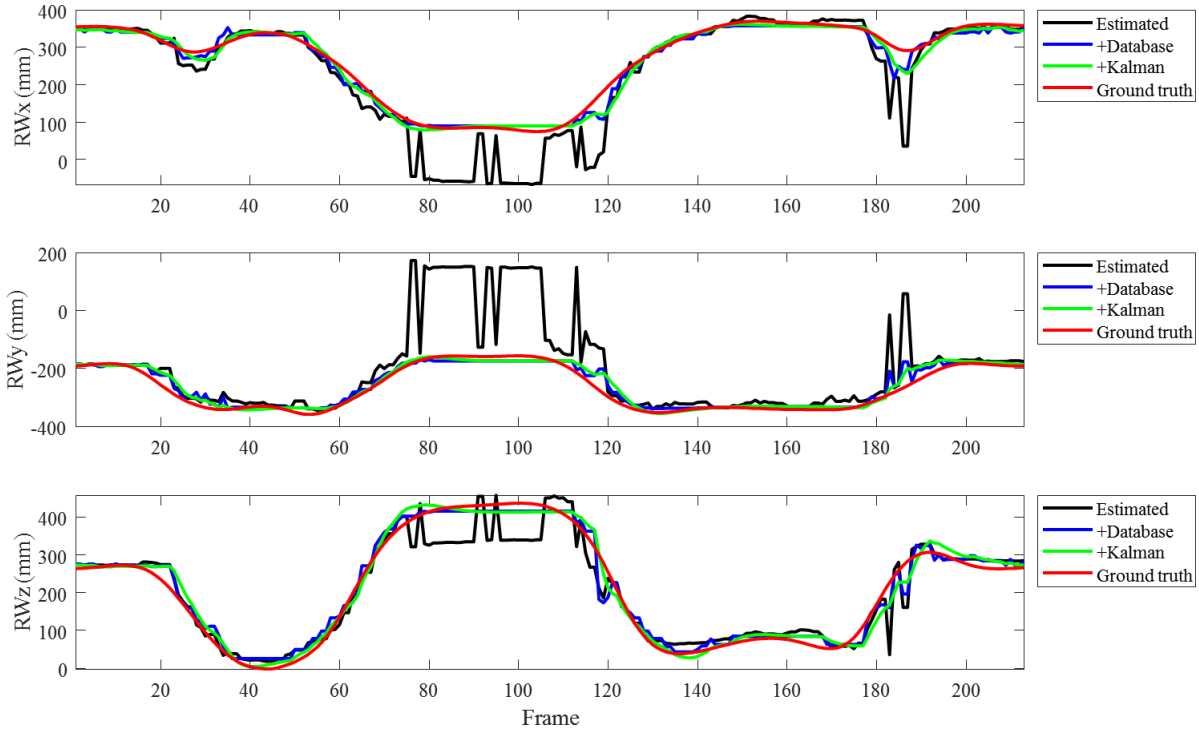


Figure 4.21. Prediction of right wrist position (x, y, z) during movement. (*Black*) Recognized joint position from depth camera. (*Blue*) Synthesized joint position by using motion database. (*Green*) Smoothed joint position by Kalman filter after posture synthesis. (*Red*) Ground truth joint position from RPx.

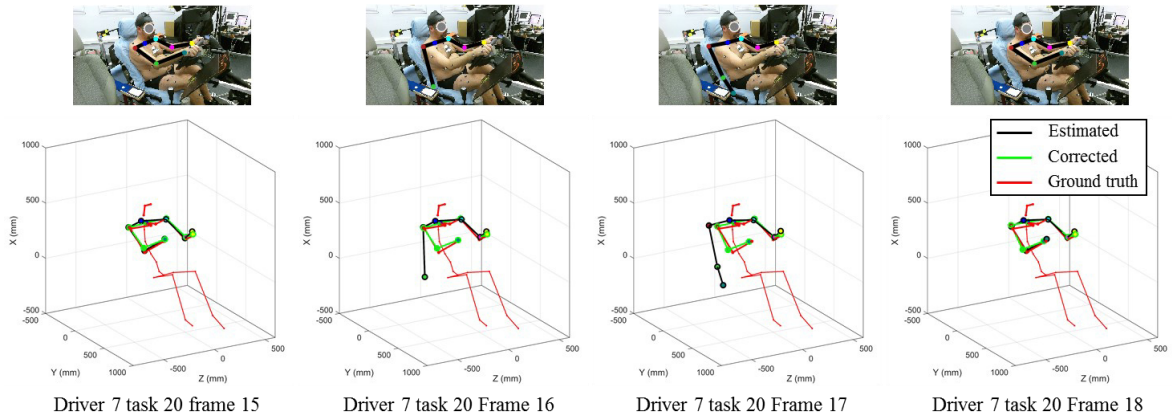
4.3.6 Results and discussion

The proposed posture correction framework along with the depth camera based posture estimation model were tested on the real data from all the tasks performed by 23 drivers. Since the posture recognition model performed better on Kinect v2 than PMD, the recognized posture on Kinect v2 was used as baseline.

To highlight the usefulness of the posture correction framework, particular attention was paid to the postures with recognition errors mentioned in Section 4.1.3. The jitter movements of a joint can be effectively smoothed out by the proposed framework (**Figure 4.22**). When occlusions or confusions occur (**Figure 4.23**), more accurate joint positions can be restored. Since the similar posture samples used for postural correction are from motion databases, the full body posture can be reconstructed though some of the joints are not tracked by the posture recognition model (**Figure 4.24**). For the same reason, the segment lengths in the posture now are consistent across frames after correction.

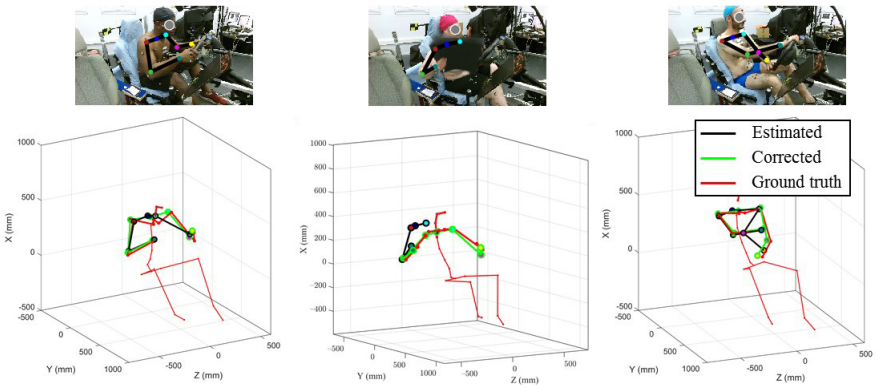
To better quantify the performance of the correction framework, the occurrence (among the whole data frames) of the prediction errors on different ranges were computed for each joint. The prediction errors were defined by the Euclidian distance between the optimized joint position and ground truth. To test if the use of the trunk postural information from pressure sensors really benefits the correction framework, another optimization method based on a general motion database for each body part (shoulder, left arm or right arm) without being conditioned on the trunk posture classes was also tested. The results are shown in **Figure 4.25**. The corrected postures had a much better error distribution, more cases of small errors (<= 50 mm) and less high errors (> 50 mm). Using the correction method without sensor fusion, the mean sample proportions of the seven joints that are within 50 mm from the ground truth positions is 91%, which can be further improved to 93% if the postural information from pressure sensors is incorporated as prior knowledge. By inspecting the accuracy improvement of each joint, it can be found that center of shoulders (CS), left shoulder (LS) and right shoulder (RS) gain more benefits from sensor fusion than the other joints that are further away from the trunk. This is because the elbows and wrists are less constrained by the trunk position.

To the best of the author's knowledge, this is the first work that attempts to quantify the prediction errors of driver's upper-body in 3D. Consequently, the results of this work cannot be directly compared to the previous studies.



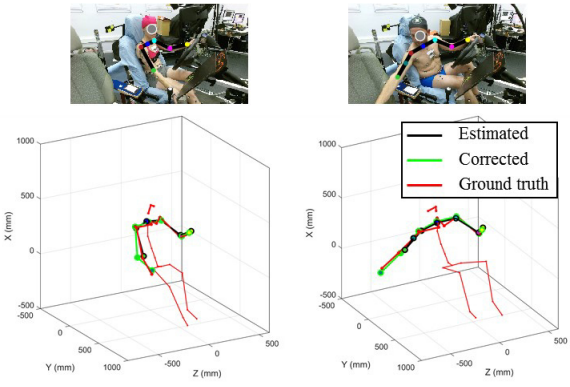
Driver 7 task 20 frame 15 | Driver 7 task 20 Frame 16 | Driver 7 task 20 Frame 17 | Driver 7 task 20 Frame 18

Figure 4.22. Correction of postures with jitter movements. Each column represents a frame. The backrest is occasionally misrecognized as right arm.



Driver 4 task 19 frame 73 | Driver 12 task 35 frame 223 | Driver 13 task 11 Frame 45

Figure 4.23. Correction of the postures with occlusions or confusions. Left column: The left shoulder is occluded when driver's trunk inclines forward. Middle column: The upper-body are partially occluded by a handbag during fast movement, leading to incomplete posture recognition. Right column: The left elbow is occluded by the right forearm when driver turns left.



Driver 6 task 24 frame 42 | Driver 10 task 35 Frame 57

Figure 4.24. Correction of postures with right arm approaches the image border. Left column: the right wrist is not recognized. Right column: the right elbow and right wrist are misrecognized.

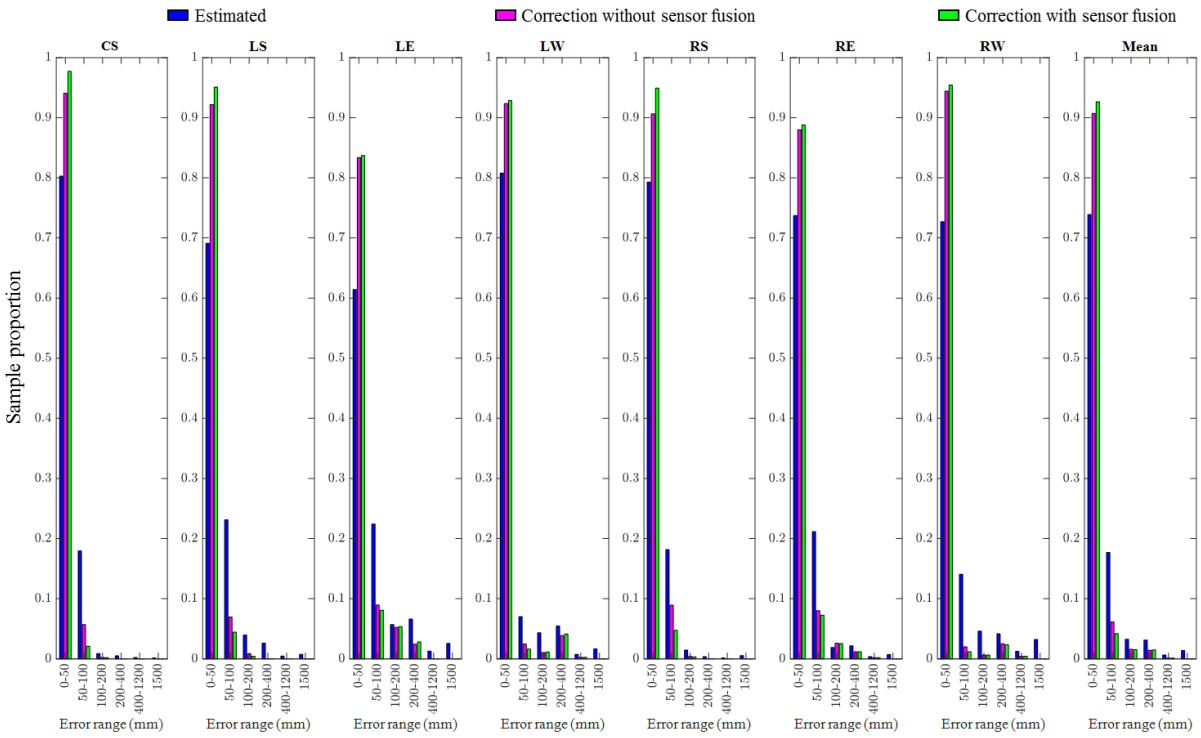
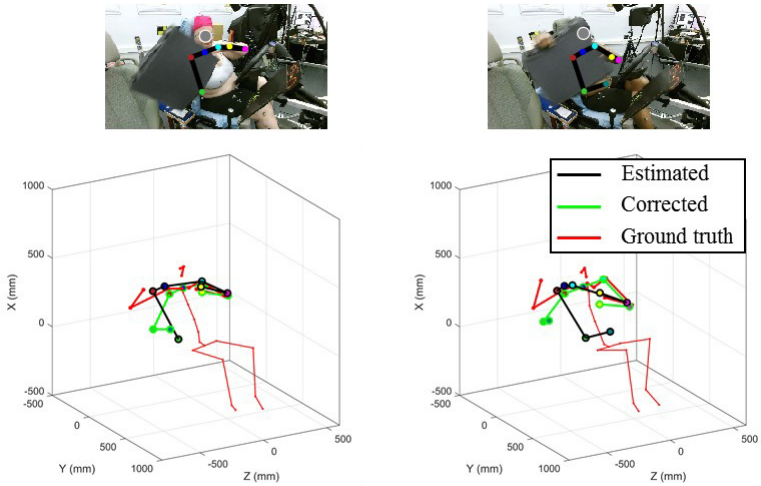


Figure 4.25. Error range distributions of posture prediction (N = 129282).

The limitation of the proposed correction framework is that if the challenging postures persists a long time or most of the upper-body are occluded by objects, the true postures cannot be recovered because too few reliable information is currently available, as illustrated in **Figure 4.26**. Under these circumstances, a possible solution is to apply a dynamic model based on the analysis of pre-recorded driver motions. The incorrectly tracked joint positions will be predicted based on the reliable trajectories in history.



Driver 12 task 39 frame 77 Driver 19 task 35 Frame 65

Figure 4.26. Failure cases of posture correction.

4.4 Summary

This Chapter proposed a system for the accurate estimation of driver's upper-body posture. First, a posture recognition model based on body part localization and offset joint regression was proposed to extract 3D joint positions from a depth camera, and the posture recognition errors were analyzed. Second, this Chapter showed that reliable trunk postural information can be inferred from pressure measurement by using a machine learning classifier based on relevant pressure features. Third, the correction framework based on sensor fusion, motion databases and Kalman filters demonstrated promising performance for reducing the posture recognition errors using a depth camera. On average, the seven upper-body joints were accurately (within 50 mm from the ground truth) predicted in 95% of the cases.

Chapter 5

5 Posture monitoring of driver's head

Head pose estimation (HPE) plays an essential role in the understanding of driver's attention, awareness and intention. In spite of a large body of literature on this topic (Selim et al. 2020; Hu, Jha, and Busso 2020), the development of an accurate HPE system is still an ongoing effort. Researchers are still struggling with the low accuracy caused by extreme head postures and occlusions. This Chapter aims to provide methods to predict the orientation and position of driver's head with help of the AutoConduct dataset.

Existing HPE systems can be divided into feature-based, appearance-based, and 3D model based classes, depending on the methods used (Selim et al. 2020). Feature-based methods rely on some specific features that are visible on the face like eyes and nose for pose estimation. Appearance-based methods takes the whole information provided into consideration and attempt to regress a pose. The input information can be a raw 2D image or depth map. 3D-model based methods derive a head model from the 2D image or depth map in order to regress a head pose.

In this Chapter, the feature-based method is adopted as the baseline. This is mainly because the facial keypoints can be obtained from the adapted CNN model proposed in Chapter 4. The rest of this Chapter is organized as follows. Section 5.1 describes the head posture reference system and explains how the ground truth head postures are obtained. Based on the available facial keypoints, head pose estimation methods are presented in Section 5.2. The evaluation metric and results are given in Section 5.3 and Section 5.4, respectively. Finally, this Chapter is summarized in Section 5.5.

5.1 Ground truth head posture

The definition of a Head Coordinate System (HCS) is crucial for determining the absolute head orientation and position measurements. The proposed HCS is illustrated in **Figure 5.1**. Its origin is located at the nose tip (NOT), the X-axis directs upwards, the Y-axis forwards and the Z-axis leftwards with respect to the driver. Using the NOT as the origin is because the NOT

was visible for most of cases we studied. The HCS corresponding to the standard driving posture is viewed as the reference coordinate system (HCS_o) for calculating the head yaw, head roll, head pitch, and head position.

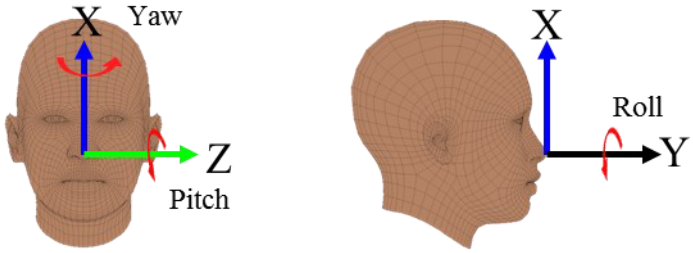


Figure 5.1. Head Coordinate System.

During head movement, the HCS defined at the standard head posture followed the rigid body transformations with the three head markers to become a new coordinate system HCS_n. The calculations of the ground truth head orientation and position thus was transformed into the problem of obtaining the relative rotations (Rx, Ry and Rz) and translations (tx, ty and tz) between HCS_n and HCS_o, as shown in **Figure 5.2**.

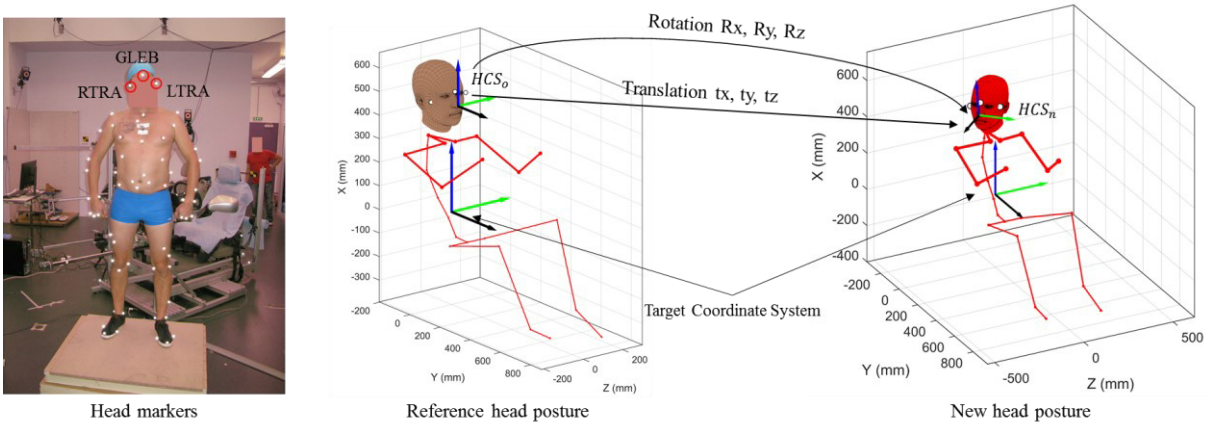


Figure 5.2. Coordinate systems and transformations.

5.2 Methods

Using the adapted CNN model, five keypoints including the right ear (REA), right eye (REY), left eye (LEY), left ear (LEA) and nose tip (NOT) were extracted from the RGB/IR image and they were then projected and transformed into the target coordinate system by using the corresponding depth image, as shown in **Figure 5.3**. The feature-based HPE method, in essence, used these keypoints instead of the reflective head markers to calculate the coordinate transformations. Unlike the head markers, the localization of the keypoints were noisy

(confidence < 0.5), and some of the keypoints could be missing (confidence = 0) due to occlusion, as shown in **Figure 5.4**.

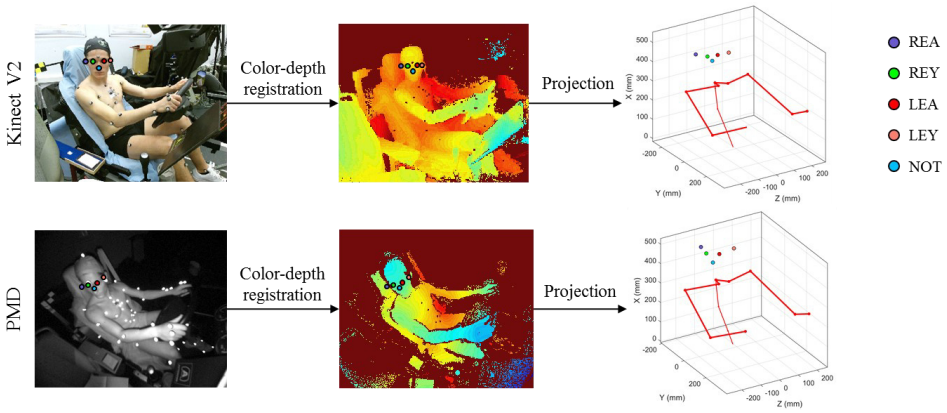
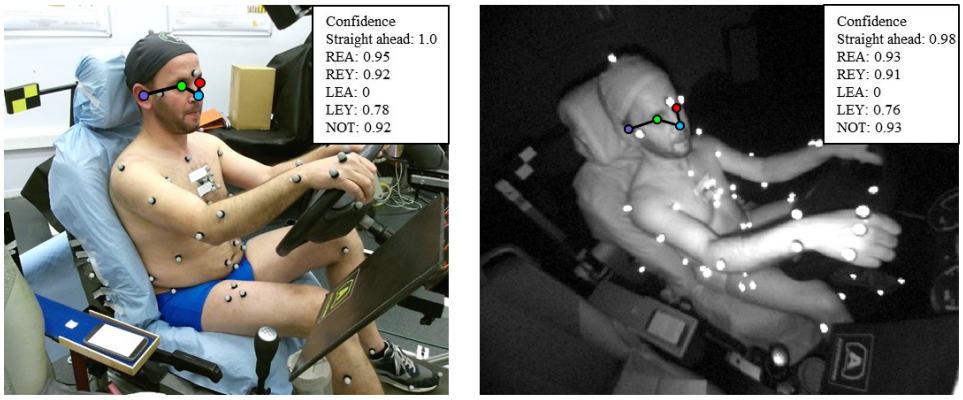


Figure 5.3. Facial keypoints from the adapted CNN model in Chapter 4.

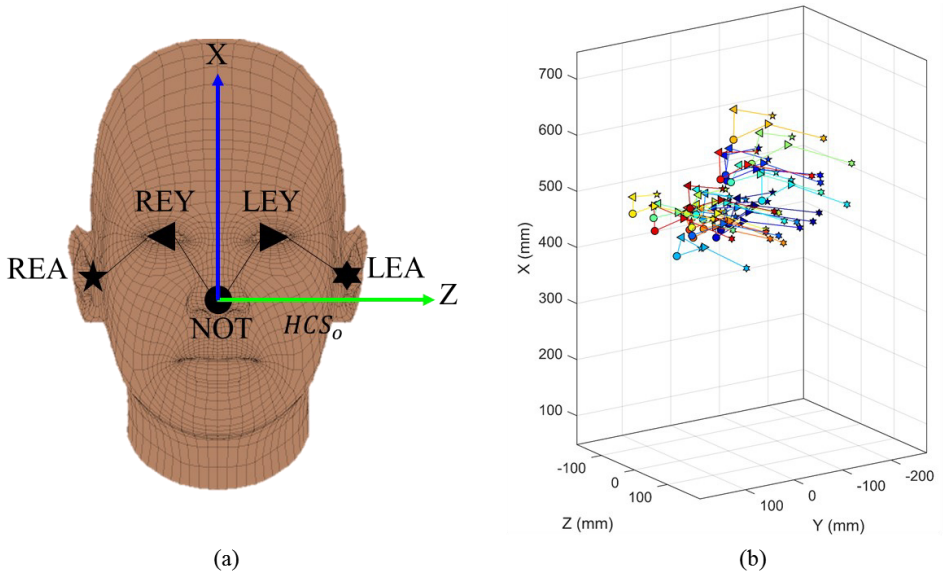
Figure 5.4. An overview of the recognized keypoints from different views. The confidence scores are returned by the adapted CNN model.

To ensure the head coordinate system during pose estimation is consistent with that used for obtaining the ground truth, a personalized head template consisting of keypoints was created for each participant when the head was in the standard driving position as used for defining the reference head coordinate systemHCS_o. Given a new set of available keypoints, a Rigid Body Matching (RBM) was performed to find the optimal rotation and translation between the new keypoints and the head template. In order to accommodate the occasional absence of some keypoints during head movement, the head template was supposed to include all of the five keypoints. However, when the head was in the standard position, only the REA, REY, and NOT can be reliably (confidence > 0.9) tracked by both cameras, as shown in **Figure 5.5**. To obtain a complete head template, the reliable keypoints on the right part of the face (REA, REY) were mirrored to the left side about the median plane passing through the NOT. The created templates are shown in **Figure 5.6**.



Kinect V2 PMD

Figure 5.5. Keypoints recognition when the head is in standard position.



(a) (b)

Figure 5.6. (a) Keypoints based head template. (b) Head templates of all the participants in the target coordinate system, where each color denotes a participant.

Note that the RBM method is sensitive to the number of the keypoints and at least three point pairs are required to find the rotation and translation between two frames. When less than three keypoints are recognized, the current frame will be replaced by the most recent frame where at least three keypoints are available.

Apart from the straightforward RBM method, a supervised learning method based on Random Forest Regression (RFR) was proposed. Random forests are collections of decision trees, each trained on a randomly sampled subset of the available data, which reduces over-fitting in comparison to trees trained on the whole dataset, as shown by Breiman (Breiman 2001). The randomness is introduced either by the subset of training examples provided to each tree, or by the subset of features used for the split optimization at each node, or in both.

The features were defined as the displacements (dx, dy and dz) of the new keypoints with respect to those in the head reference template. To predict the head orientation, quaternions ($q \in \mathbb{R}^4$) were opted as the output of the regression model because Euler angles suffer from the gimble lock problem. As the regression model only supports single output, seven models were built to individually predict the four elements of the Quaternions and three elements of the head positions. Finally, the Euler angles were obtained from the quaternions.

During training, the features were obtained by moving the head template along with the head markers. If some keypoints were in fact lost, the training features related to the lost keypoints were assigned with NaN values. The purpose of this step was to help the regression models to be better adapted to the real applications during testing.

Similar to Chapter 4, Kalman filters were used to smooth the predicted head motions from RBM or RFR. The state of head position (origin of the HCS) was expressed as a vector consisting of the current position (tx, ty, tz) and velocity (v_x, v_y, v_z).

$$\hat{x}_p = [tx, \mathrm{ty}, \mathrm{tz}, v_x, v_y, v_z] \tag{5.1}$$

The state of the head orientation was a vector consisting of the current Euler angles (Rx, Ry, Rz) and angular velocity ($\omega_x, \omega_y, \omega_z$).

$$\hat{x}_r = [Rx, Ry, Rz, \omega_x, \omega_y, \omega_z] \tag{5.2}$$

The transition matrices are the same as the one in Equation (4.21). Similarly, the measurement noise covariance was measured prior to the filter operation. The process noise covariance was empirically estimated. As seen in **Figure 5.7**, the Kalman filters can effectively smooth out the high frequency vibrations of the predicted results.

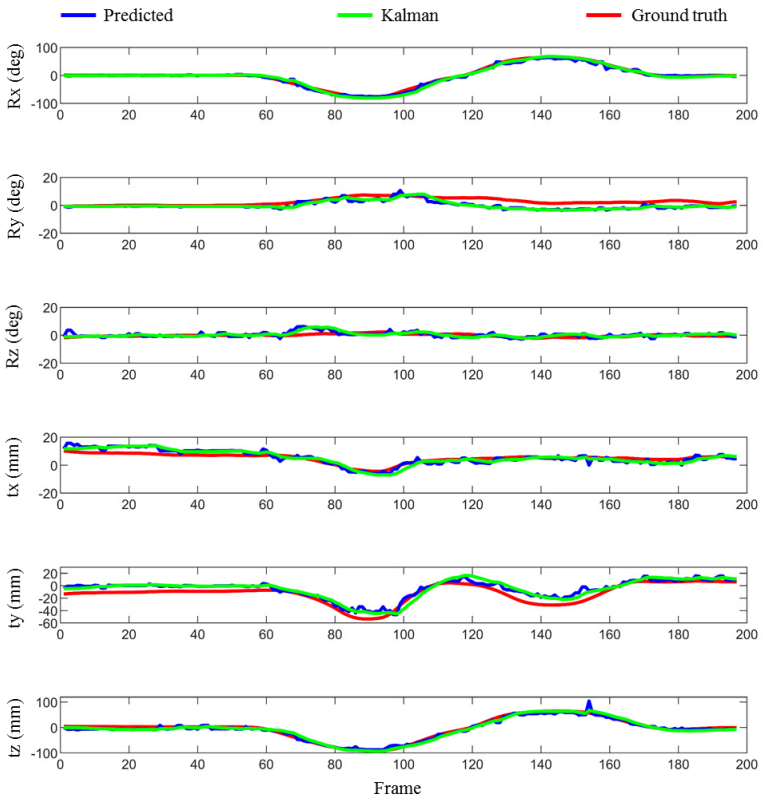


Figure 5.7. Head motion smoothed by Kalman filters. The head orientation and position are predicted by RFR method.

5.3 Evaluation metrics

To provide a good benchmarking foundation, two metrics from the previous studies (Selim et al. 2020; Schwarz et al. 2017) were adopted to evaluate the performance of the two HPE methods.

The first metric was the Mean Absolute Error (MAE) defined as Equation (5.3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \tag{5.3}$$

Where y is the predicted value of Rx, Ry, Rz, tx, ty or tz, $\hat{y}$ is the corresponding ground truth, n is the total number of data frames.

The second metric was the Balanced Mean Error (BME), which evenly splits the data frames based on the true angular or positional deviations with respect to the reference posture into multiple smaller bins and averages the MAE of each of the bins:

$$BME = \frac{1}{k}\sum_{i} \emptyset_{i,i+d} \tag{5.4}$$

Where d is size of the bin, k is the number of the bins, $\emptyset_{i,i+d}$ is the MAE of the angular or positional prediction when the ground truth is in the i^{th} bin. As the dataset is filled with many

postures easy for accurate pose estimation and only a few challenging postures, the BME can thus give un unbiased evaluation for the HPE methods.

To allow for a fine-grained evaluation of the HPE methods, the dataset was further split into three subsets: *easy*, *moderate*, and *hard* depending on the number of recognized keypoints (N_{kp}). *easy*: $N_{kp} = 5$; *moderate*: $N_{kp} = 3$ *or* 4; *hard*: $N_{kp} \leq 2$. Since the keypoints recognition varied between different views, the dataset split was performed for Kinect v2 and PMD separately, as shown in **Figure 5.8**. The proportions of the dataset splits for Kinect v2 were 9.6%, 86.7%, 3.7% for *easy*, *moderate*, *hard,* while they were 11.7%, 81.1%, 7.2% for PMD.

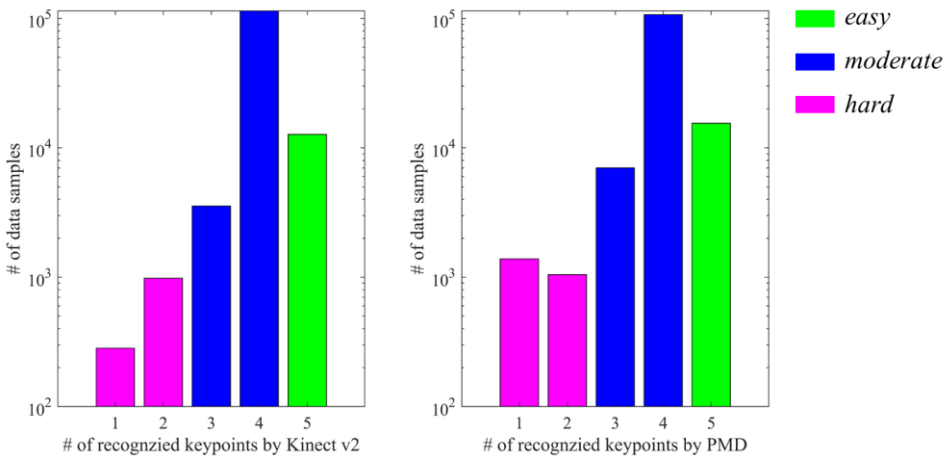


Figure 5.8. Dataset split by the number of recognized keypoints.

Another common issue related to HPE is the occlusions by glasses. To further reveal the performance of the HPE methods, the dataset was also split by the presence of glasses to have two subsets: wearing glasses (8 drivers) and not wearing glasses (15 drivers).

5.4 Results and discussion

Table 5.1 gives the performance of the two HPE methods (RBM and RFR) evaluated on the whole dataset. A detailed distribution of the MAE of each of the six head parameters can be found in **Figure S1**. Using the Kinect v2, the RFR method outperforms the RBM method for the prediction of both head orientation and position. Concerning the PMD, the RFR method performs better than the RBM method for predicting the head orientation but slightly worse for the head position. Like the posture recognition of the upper-body, the Kinect v2 that is close to the right A pillar performs better than the PMD placed on top of the center mirror for the pose estimation of driver's head. Comparing the drivers wearing glasses with those without glasses (**Table 5.2**), the average prediction error in head position by using Kinect v2 increases from 15.3 mm to 23.1 mm, and the error by using PMD increases from 57.7 mm to 76.3 mm. In

contrast, the average prediction error in head rotation with glasses is slightly smaller than that without glasses regardless of whether the Kinect v2 or the PMD is concerned.

Table 5.1. Balanced Mean Error (BME) of different HPE methods evaluated on the whole dataset using different cameras.

Method	Camera	Orientation (deg)					Position (mm)			
		Rx	Ry	Rz	Avg		tx	ty	tz	Avg
RBM	Kinect v2	47.8	14.3	25.4	29.2		17.0	23.2	34.1	24.8
	PMD	52.3	17.9	29.9	33.4		55.4	58.7	85.6	66.6
RFR	**Kinect v2**	**19.1**	**8.0**	**6.3**	**11.1**		**14.4**	**19.3**	**28.6**	**20.8**
	PMD	27.5	16.5	22.3	22.1		60.4	76.4	71.2	69.3

Table 5.2. Balanced Mean Error (BME) of the RFR method for drivers wearing glasses and not wearing glasses.

Camera	Glasses	Rotation (deg)					Position (mm)			
		Rx	Ry	Rz	Avg		tx	ty	tz	Avg
Kinect V2	0	19.1	9.3	5.8	11.4		10.7	18.1	26.0	15.3
	1	18.2	8.1	6.5	10.9		17.5	19.3	32.4	23.1
PMD	0	25.7	13.5	21.9	20.3		55.1	75.9	42.1	57.7
	1	27.1	18.7	13.5	19.7		43.7	62.9	79.3	76.3
Note: 0 – without glasses (15 drivers), 1 – with glasses (8 drivers)										

Table 5.3 shows the evaluations of the RFR method on the three subsets defined by the number of keypoints recognized by Kinect v2. Average BME of head orientation ranges from 8.4° on the *easy* subset to 17.2° on the *hard* subset. With respect to the head position, the average BME ranges from 11.1 mm on *easy* subset to 40.2 mm on *hard* subset. Results clearly show that the head pose estimation performance is highly dependent on the number of recognized keypoints increasing by a factor close to 4.

Table 5.3. Balanced Mean Error (BME) of RFR method evaluated on the subsets using Kinect v2.

Subset (proportion)	Orientation (deg)					Position (mm)			
	Rx	Ry	Rz	Avg		tx	ty	tz	Avg
easy (9.6%)	13.0	7.7	4.5	8.4		5.8	13.9	13.5	11.1
moderate (86.7%)	18.8	7.9	6.0	10.9		11.1	16.9	27.6	18.5
hard (3.7%)	30.1	6.3	15.3	17.2		25.6	43.1	51.9	40.2

Unlike the previous studies (Hu, Jha, and Busso 2020; Selim et al. 2020; Roth and Gavrila 2019; Schwarz et al. 2017), this work attempted to use a non-frontal camera to estimate driver's head posture including orientation and position. The frontal camera is indeed better for the monitoring of driver's head but cannot cover driver's upper-body due to the unique view angle. In spite of the suboptimal camera position, the proposed method in this chapter was able to provide promising results for head posture estimation in most of the cases. In Chapter 4, the

same camera is used to monitor the driver's upper-body. This allows one to monitor the driver's head and upper-body at the same time by using a single camera.

Future work is needed to explore other methods for the estimation of driver's head posture based on the AutoConduct dataset. For example, the OpenFace 2.0 algorithm (Baltrusaitis et al. 2018) that is able to provide up to 68 facial keypoints from the image, with which the RFR method is believed to perform even better. Other methods such as the PointNet++ framework proposed by Hu et al. (Hu, Jha, and Busso 2020) and the CNN model proposed by (Venturelli et al. 2017) can be tested as well. In addition to the regression methods, the estimation of driver's head posture can be transformed into a classification problem by discretizing head postures into different classes, such as rotating left, rotating right, bending down, etc.

5.5 Summary

In this Chapter, a feature-based method was employed to extract the head orientation and position based on the AutoConduct dataset. The head coordinate system was defined and the ground truth head postures were obtained by using a Mocap system. Baseline methods based on rigid body matching or Random Forest regression were proposed. The predicted head motions were smoothed by exploiting Kalman filters. Finally, benchmarking results were provided and analyzed. Further effort is still needed to find better methods for the estimation of driver's head posture with big rotation or translation with respect to the standard posture.

Chapter 6

6 Posture monitoring of driver's lower-body

The monitoring of driver's lower-body can give rise to a variety of promising applications, ranging from the detection of pedal errors such as pedal misapplication (Collins, Evans, and Hughes 2015; Jonas et al. 2018), and the evaluation of the driver's readiness to takeover in an autonomous vehicle (Deo and Trivedi 2019) to the estimation of injury risks related to the non-nominal sitting posture such as crossing legs (Leledakis et al. 2021). In the field of driver posture monitoring, tremendous work has gone toward the posture recognition of driver's upper body such as head and hands (Wang et al. 2019; Hu, Jha, and Busso 2020; Yuen and Trivedi 2019; Xing et al. 2019; Billah et al. 2018; Torres et al. 2019). To date, only a limited number of researches focused on the monitoring of driver's lower-body.

The earliest research on the monitoring of driver feet positions was performed by Tran et al (Tran, Doshi, and Trivedi 2012; Tran, Doshi, and Trivedi 2011). They used the optical flow computed from grayscale images for foot tracking and Hidden Markov Model for the prediction of seven pedal application states: Neutral (hover off pedal), moving towards brake, brake engaged, release brake, accelerator engaged, moving towards accelerator and accelerator released. The evaluation of their system showed a mean correct rate of 93.77% for classification, which degrades to 82.44% by LOO cross-validation across 12 subjects. Thanks to their system, they were able to predict ~74% of the pedal presses 133 ms prior to the actual foot-pedal contact. The main limitation of this system is attributed to the nature of optical flow approach. If the shoes lack texture or the shoes appearance is indistinguishable from the background, the system will not work properly. Rangesh and Trivedi (Rangesh and Trivedi 2019) proposed a network to associate relative spatial locations of right foot in the RGB image with one of the five foot activities: away from pedals, hovering over accelerator, hovering over break, on accelerator, and on break. The test on three drivers' data led to an overall classification accuracy of 97.49%. In a naturalistic driving study, Deo and Trivedi (Deo and Trivedi 2019) employed infrared sensors on the accelerator and brake pedal to measure the distance of driver's foot from each pedal. Frank and Kuijper (Frank and Kuijper 2019) developed a feet gesture prediction system

using capacitive proximity sensor arrays surrounding the footwell space. They used Random Forest (RF) classifiers for the recognition of four feet gesture classes namely toe tap, heel tap, toe rotation and heel rotation. Ten-fold cross-validation on six subjects' data showed a positive classification rate of 93.03%. However, their system was intended to use driver's feet for gesture control when the feet are in idle state. Ansari et al. (Ansari, Du, and Naghdy 2020) developed a system to monitor right foot trajectories between accelerator and brake pedal by using XSENS motion capture system. In this work, a new modified bidirectional long short-term memory (Bi-LSTM) deep neural network was designed and trained on nine drivers' right foot orientation angle sequences to predict foot activities including accelerating and braking. The authors tested the neural network on the time-series data of one new driver and had an overall accuracy of 99.8%. This system requires sensors to be attached on driver's body thus is not practically feasible for the application on real road.

Recently, the ubiquitous use of commercial depth cameras has led to a growing interest in the recognition of driver upper body posture from a depth image (Zhao et al. 2018; Xing et al. 2017; Jegham et al. 2020; Borges et al. 2021). Motivated by the 3D measurement, exploring the possibility of using a depth camera to monitor the driver's lower limbs can be an interesting topic. In addition, it has been observed that driver's feet movements could cause pressure distribution variations in the AutoConduct dataset (Section 4.2). Is it possible to detect the feet positions by using pressure sensors on driver's seat?

The goal of this chapter is to explore effective methods based on a depth camera and/or pressure sensors that contribute to improving the driving safety by detecting driver feet positions. The rest of this Chapter is organized as follows. Section 6.1 presents a method to predict driver's feet positions by using pressure sensors. The prediction method based on a depth camera is presented in Section 6.2. Sensor fusion methods are investigated in Section 6.3. This Chapter is summarized in Section 6.4.

6.1 Classification of feet positions based on pressure sensors

6.1.1 Method

This Section is an extension of Section 4.2 in Chapter 4, where 200 pressure feature candidates were constructed and selected to classify five different trunk postures. In this Section, the same pressure feature candidates are evaluated and selected to classify the drivers' typical feet positions that cover the posture space of driver's lower body in both manual and automated driving modes. The feet positions are represented by the foot center positions of the ground

truth posture. For the right foot, three classes (RFP0—right foot on accelerator pedal, RFP1—right foot on brake pedal, and RFP2—right foot on floor) were defined, while two classes (LFP0—left foot on floor and LFP1—left foot on clutch) were defined for the left foot. The class definition scheme is illustrated in **Figure 6.1**. The RFP0 and LFP0 were regarded as the reference position for right foot and left foot, recpectively.

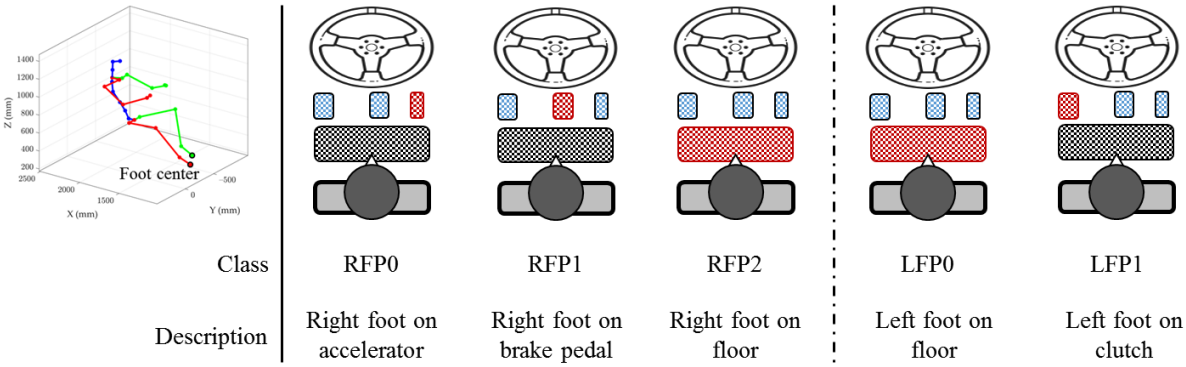


Figure 6.1. Definition of classes of driver's feet positions.

To reduce data redundancy and overfitting risk of the classifiers, an intra- and inter-motion filter was applied to remove similar postures of the same participant. Two right-foot positions were regarded as different if the Euclidean distance between the foot centers is bigger than 2 cm. The same rule applies to the left foot. The filtering process was independently performed for each foot. Finally, 8024 right-foot postures, and 5216 left-foot postures were extracted from the experimental dataset.

Similar to Section 4.2, the Random Forest OOB error estimation method was used to evaluate the feature importance for the classification of left/right positions, and the Random Forest classifiers were iteratively trained through all the important relative feature combinations. The F1 Score was again used to quantify the classification accuracy of the classifiers.

6.1.2 Results and discussion

Figure 6.2 gives the classification performance of the classifier for left foot or right foot along with the number of important features, and the confusion matrixes corresponding to the classifiers with best feature combinations. As shown in **Figure 6.2**, the average F1 Score of left foot positions plateaued (0.93) at around 24 important features and remained stable thereafter. The potential best average F1 Score (around 0.74) for the classification of right foot positions is approached for the first time when 22 important features were selected. Note that the use of more features did not necessarily improve posture recognition scores. The selected important features are given in **Table S3**.

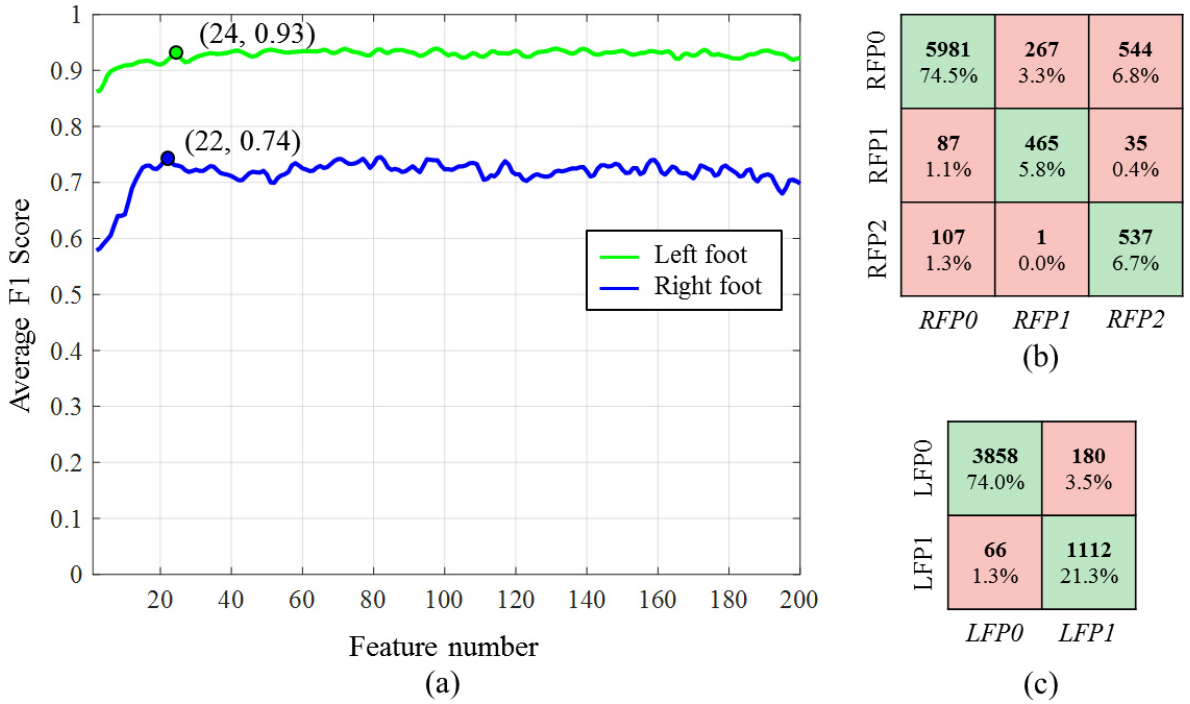


Figure 6.2. (a) Average F1 Score vs feature number. (b) Confusion matrix of the classification of right foot position. (c) Confusion matrix of the classification of left foot position.

The main errors are related to the recognition of two right foot positions, RFP1 (right foot on brake pedal) and RFP2 (right foot on floor), which are frequently misclassified as RFP0 (right foot on accelerator). One reason could be the high inter-individual differences in terms of the right foot position when performing related tasks. When braking, some participants just slightly rotated the right shank to reach the brake pedal, while others first shifted the right foot to left before pressing. Likewise, for relaxed position with the right foot on the floor, big differences in foot positions between drivers were observed. Consequently, if the test driver did not perform the tasks in a similar way as in the training data, the position could be misclassified. To overcome these problems, more data need to be collected to refine foot position classes, and the use of additional pressure sensors on the floor could also be of help. Another alternative is to explore the potential of using a depth camera to predict the feet positions as explained in the following Section.

In terms of the influence of the driver's BMI on posture classification as discussed in Section 4.2, the results for the classification of feet positions were in line with that for the classification of trunk position, i.e., the classifiers trained on the relative pressure features have robust performance across drivers of different body sizes.

6.2 Posture recognition based on a depth camera

Due to the limited operating range, the depth camera DS325 was placed to look at the driver shanks instead of the feet. This is because the footwell in a car is very small, it is challenging to find a desirable position and view angle for the depth camera to look at both feet directly. **Figure 6.3** illustrates the placement of the DS325.

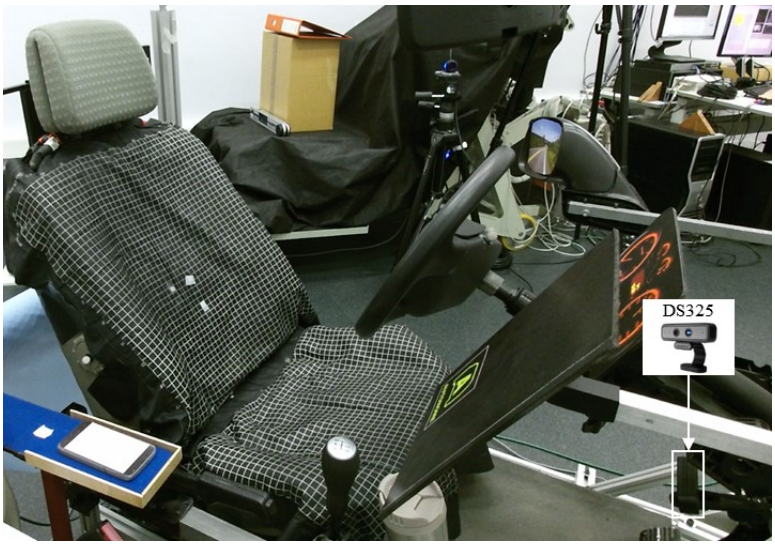


Figure 6.3. Depth camera DS325 is mounted above the pedals and right below the steering wheel to scan the legroom.

Under this circumstance, driver's feet were not always visible in the depth image, as shown in **Figure 6.4**. To predict the feet positions, the shank postures needed to be first described with the expectation that the feet positions could be inferred based on the recognized shank postural information.

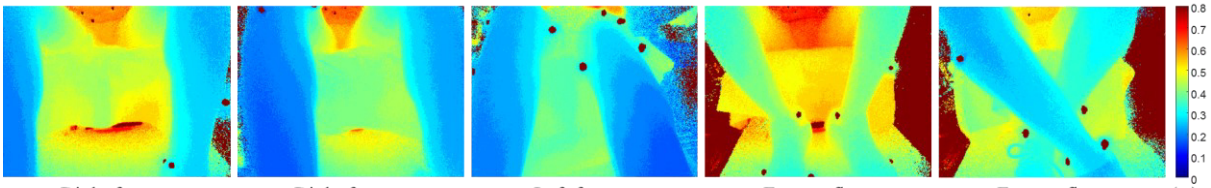


Figure 6.4. Depth view of DS325 when the drivers are performing different tasks. The default positions for left foot and right foot are on the floor and on the accelerator, respectively, unless otherwise specified.

6.2.1 Identification of shank related keypoints and postural features

Based on the intrinsic parameters of the depth camera, the depth image can be converted to a point cloud in the camera coordinate system as shown in **Figure 6.5**. To describe the shank position and orientation, an idea was to extract several keypoints from the point cloud.

Unfortunately, the shanks lack distinctive visual features on the depth image or in the point cloud. Aiming at specific regions or landmarks fixed on the shanks thus was not a good idea. Instead, we proposed a method to extract keypoints by analyzing the point cloud of the shanks at specific regions fixed in the legroom space. Note that the drivers in our experiment were not normally dressed because of the use of motion capture system. We thus assume that drivers on real road are dressed with shorts or tight trousers, which generate similar depth image patterns to the ones shown in **Figure 6.4**.

First, one vertical plane (XY-plane), right in front of the driver seat pan and behind the driver shanks, was used to extract the point cloud belonging to the shanks. The Z value of the background plane was estimated by inspecting the depth of the front edge of the seat pan in the middle of the depth image, as shown in **Figure 6.5**.

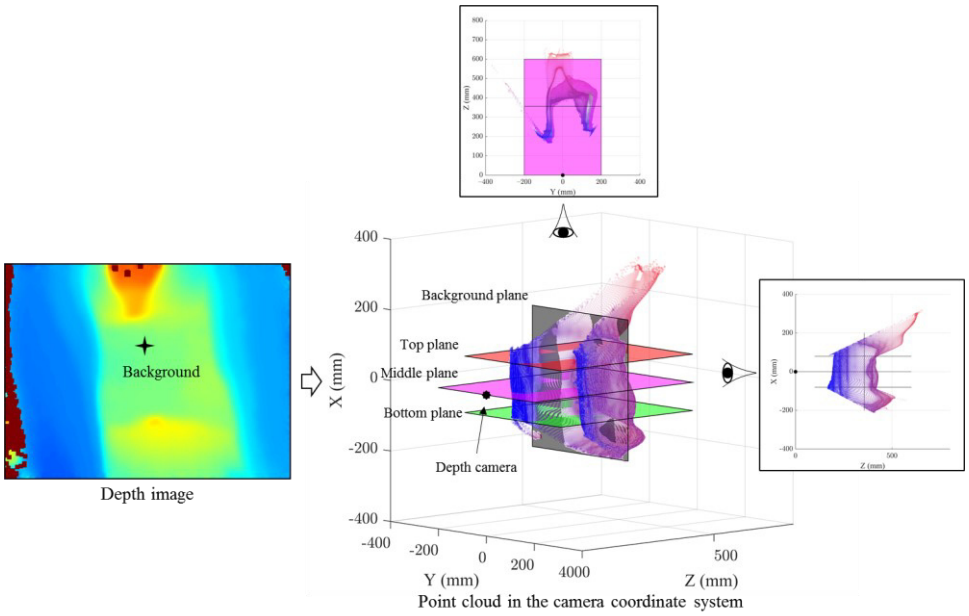


Figure 6.5. Legroom segmentation and three cutting planes used to slice the shanks.

Once the point cloud in the legroom was obtained, the first region that came into our mind was the lower section of the legroom. Because this section corresponds to the lower part of the shanks which was close to the foot and a keypoint extracted from this section may serve as a strong indicator for predicting the foot position. The foot position is also determined by the shank orientation. Therefore, a second region corresponding to the upper part of the shank was needed. However, with complex shank postures, such as legs crossed, the lower or upper section of the shank may not be visible in the depth image. With these considerations, we proposed to extract keypoints from three regions in the point cloud for each shank. The three regions here were represented by profiles on three YZ-planes, as shown in **Figure 6.5**.

The middle plane simply passed through the camera optical center aiming at the middle part of the legroom section. The top plane was used to inspect the upper part of the legroom space and the bottom plane for the lower part. Given the limited field of view, a higher top plane and a lower bottom plane will be of help for a holistic inspection of the legroom space. However, when the shank gets closer to the camera (e.g., when the participant extended the legs to reach pedals), a top plane and a bottom plane far away from the middle plane may not return valid profiles of the shank surface. This is because when the object gets closer to a camera, the visible area becomes smaller, as shown by the left view in **Figure 6.5**. Based on the observation of the experiment data, the top plane and the bottom plane were empirically set at X = 80 mm and X = –80 mm, respectively. Note that, there were almost no points falling exactly on these YZ-planes. To obtain complete profiles from the point cloud, a slight difference (1 mm) was allowed, i.e., the points in the vicinity of the YZ-plane within a range of 1 mm were extracted to form the profile.

Once the profiles in the three cutting planes were retrieved, keypoints were identified from the profiles to represent the shank postures, as illustrated in **Figure 6.6**. The recognition of shank keypoints was accomplished by using a three-step procedure as explained below.

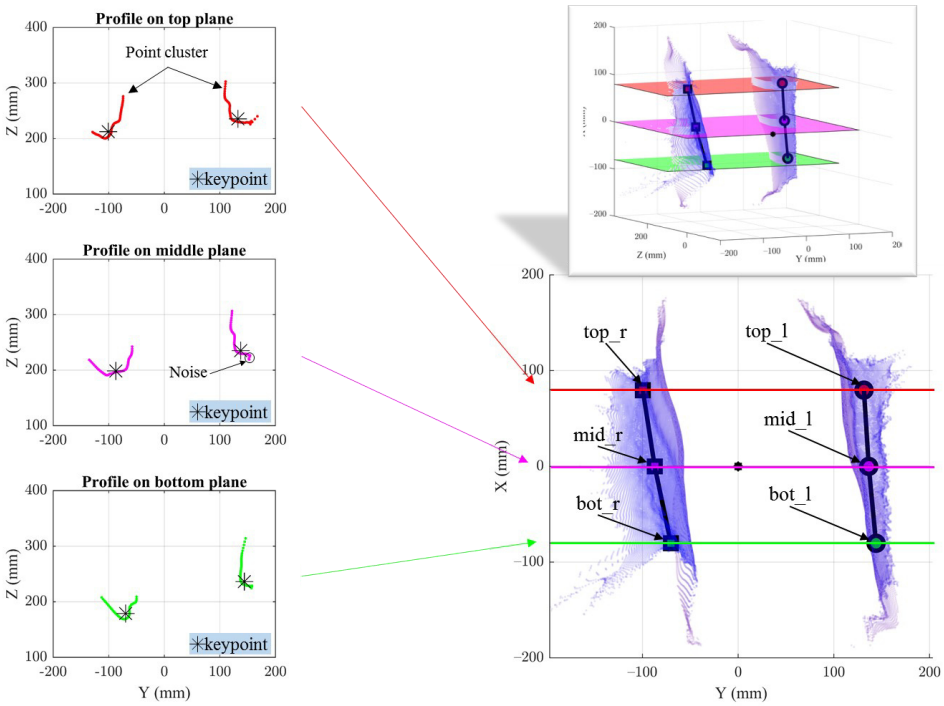


Figure 6.6. Shank posture recognition.

The first step was to identify point cluster number in each profile and to compute keypoint from each cluster. If two shanks both were visible on one plane, there will be two distinct point

clusters included in the profile, which was the case in the left part of **Figure 6.6**. However, when driver crosses the legs or the lower part of the leg stuck into the space underneath the seat pan (**Figure 6.7**), there may be only one cluster in the profile. Thus, a method automatically grouping the points in each profile was needed. Once the number of clusters was determined, we chose to compute one keypoint from the cluster to represent the position of the upper/middle/lower shank on the plane. A simple option was to select the point closest to the camera from the cluster. However, using such a point from the surface of the shank was prone to inaccuracy due to occasional local measurement noise (e.g., the outlier at the border of the profile on the middle plane in **Figure 6.6**). In this work, an automatic clustering method based on mean-shift (Comaniciu and Meer 2002) was employed to identify the number of point clusters, meanwhile to find the keypoint for each cluster.

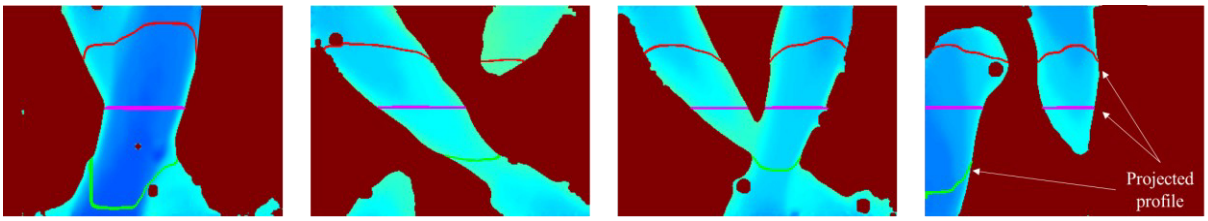


Figure 6.7. Challenging shank postures in the segmented depth image.

Mean-shift is a non-parametric density estimation technique widely used for cluster analysis in computer vision. The algorithm starts at a subset of points from the profile on each plane, and iteratively computes the mean-shift vector to approach the modes (the maxima) of the density function implied by the discrete data points. The points that eventually converge to the same point are grouped as one cluster, and each cluster represents the shape of one shank on the plane. We used a weighted Gaussian kernel function to find the converged point which can be regarded as a weighted center of the cluster with higher weights for the points closer to the camera, and the converged point is adopted as the keypoint required by this work. As opposed to a point lying on the shank surface, this keypoint is less sensitive to local measurement noise.

The second step was to determine the identities of the keypoints available on the three planes and connect the keypoints belonging to the same shank. In most cases (e.g., accelerating, braking and switching gear), this was an easy task for the profile on each plane witnessed two keypoints and the left keypoints in the plot belong to the right shank while the right keypoints in the plot belong to the left shank. This task turned out to be tricky when driver crosses the legs or the lower part of the shank is away from the camera fading into the background (**Figure 6.7**). In this case, there was only one keypoint on the plane. This keypoint may belong to either right shank or left shank, depending on the postural context. These rare circumstances usually

implied that driver's feet were in abnormal positions which may threaten driving safety, therefore should be specifically accounted.

To tackle these issues, two latent variables were introduced during shank posture recognition. The first latent variable (*crossed*) was used to indicate if the two legs could be judged as crossed (*crossed* = 1) or not (*crossed* = 0). The second latent variable was the angle ($angle\_b$) between two shanks in the XY plane.

$$angle\_b = angle\_l + angle\_r \tag{6.4}$$

where $angle\_l$ and $angle\_r$ are the rotation angles of the left shank and the right shank relative to a vertical line in the XY plane, respectively. For both shanks, inward turning about the top key point (top_r or top_l) relative to a vertical line had a positive value. If the bottom key point of one shank was missing, the middle key point will be connected to the top one to calculate the rotation angle. If $angle\_b$ was bigger than a threshold (40°), the two legs will be judged as crossed.

These two latent variables on one hand were used during the determination of keypoints' identities. On the other hand, they were part of the output of shank postural information. In addition, it is observed that if driver's legs were crossed, the keypoint of the front shank was closer to the camera than the other if available, which was consistent across all the planes. With help of the latent variables and this empirical knowledge, the keypoints can be related to the correct shank by inspecting the number of keypoints from the profile plane by plane in a top-down order.

Finally, if both the top and middle keypoints of one shank were available while the bottom one is missing, the linear interpolation method was used to complement the missing point, which is calculated as the intersection of the bottom plane and the line determined by the top and middle key points.

6.2.2 Classification of feet positions

The postural information of the shanks is summarized in **Table 6.1**. They consist of 22 features including latent variables and 3D positions of the key points. Similar to Section 6.1, the labelled data was collected to include 8024 samples for right foot and 5216 for left foot to train and test the prediction model.

Intuition indicates that foot position is largely related to the position of shank's lower part. However, the lower key point in this work was not fixed on the shank rather restrained by the bottom plane. Therefore, instead of relying on the lower key points alone, we used the whole features from the shanks to predict feet positions with the expectation that some implicit

knowledge can be learned from the global shank posture. To this end, a feature evaluation procedure based on Random Forest (RF) Out-Of-Bag (OOB) error estimation method was applied. Since the vertical positions (X values) of the key points were constants across all drivers, they were excluded from the feature set. The remaining 16 features were ranked by their importance for distinguishing different positions of right foot and left foot respectively. The RF classifier was again used to evaluate the impact of different feature numbers on classification accuracy, starting from the most important two features until all the features are involved with an increment of a less important one each time. Likewise, LOO cross-validation was performed to evaluate the generalization capability of the classifier and the F1 Score (a harmonic mean of Precision and Recall) was adopted as the evaluation metric.

Table 6.1. Shank postural information.

Type		**Feature**
Latent variables		*crossed, angle_l, angle_r and angle_b*
Positions of key points	Left shank	*top_l(x,y,z)*
		mid_l(x,y,z)
		bot_l(x,y,z)
	Right shank	*top_r(x,y,z)*
		mid_r(x,y,z)
		bot_r(x,y,z)

6.2.3 Results and discussions

We first examined the performance of the proposed shank posture recognition method during all the driver motions. A few representatives are given in **Figure 6.8**. The results show that the proposed method performed well when both shanks were clearly visible in the depth image (**Figure 6.8** a, b, c & d). Even for the challenging shank postures (**Figure 6.8** e, f, g & h), the proposed method could successfully relate the keypoints to the correct shank and recover the missing key point due to occlusion or background excessive segmentation. Particularly when crossing legs, the shanks of short obese drivers are very close to each other in the depth image and one of their shanks may be too close to the camera leading to invalid measurement (**Figure 6.8** g), the mean-shift algorithm is still able to find out the desirable keypoints from the cluttered profile.

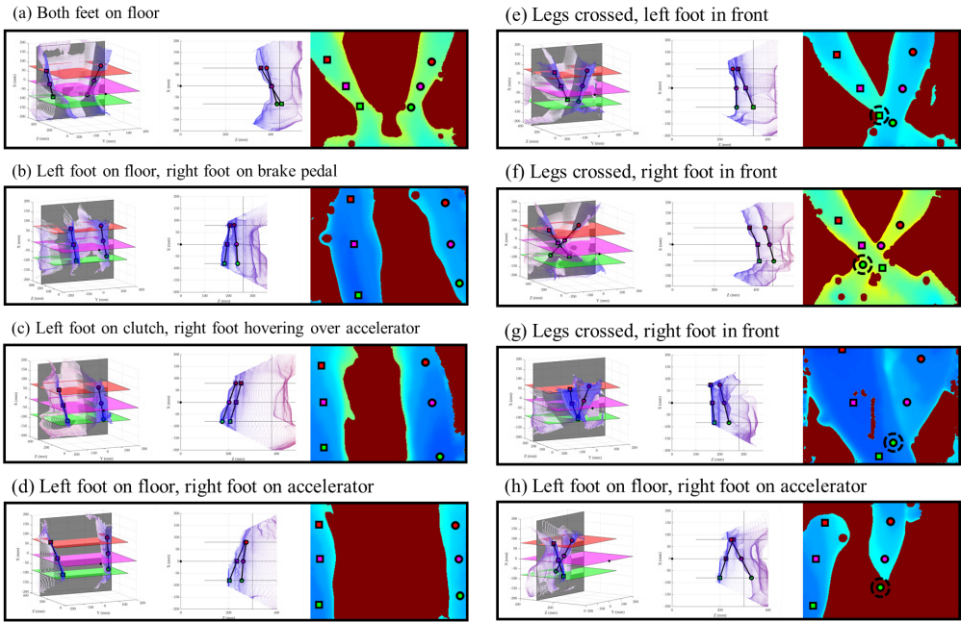


Figure 6.8. Shank posture recognition results. Each individual figure gives the perspective view in the left, driver left view in the middle and depth image view in the right. The key points surrounded by black dashed circles are inferred by interpolation.

Figure 6.9 (a, c) shows the estimated feature importance by RF OOB errors and the prediction performance of feet positions as a function of different important feature combinations. It can be observed that the first two important features for predicting left foot positions are related to the left lower keypoint (bot_lz and bot_ly). Likewise, the first two important features for predicting right foot position are related to the right lower keypoint (bot_rz and bot_ry). With help of additional information from right shank (e.g., bot_ry), the possible highest average F1 score of 0.94 could be achieved for predicting left foot positions. Regarding the prediction of right foot positions, the highest average F1 score of 0.88 could be achieved at the first 13 important features, including features from both shanks and latent variables. The use of more features thereafter does not necessarily improve the classification performance further. The Confusion matrixes of the best classification results are given in **Figure 6.9** (b, d). Relative lower F1 Score (0.84) is determined for the classification of two right foot positions: RFP1 and RFP2. This was because RFP1 was occasionally recognized as RFP0 and vice versa. Regarding RFP2, it was also occasionally recognized as RFP0.

To further reveal the classification errors of the feet positions, we performed a similar study using the ground truth posture instead of the measurement from the depth camera. We found that the classifiers trained on the ankle positions only could lead to an average F1 score of 0.97 for three right foot positions and 0.99 for two left foot positions, although the degree of freedom

of the foot extension relative to the ankle and the shank rotation relative to the knee are not considered. The results thus suggest that the classification errors come from the uncertainty of the shank keypoints recognized by the proposed method.

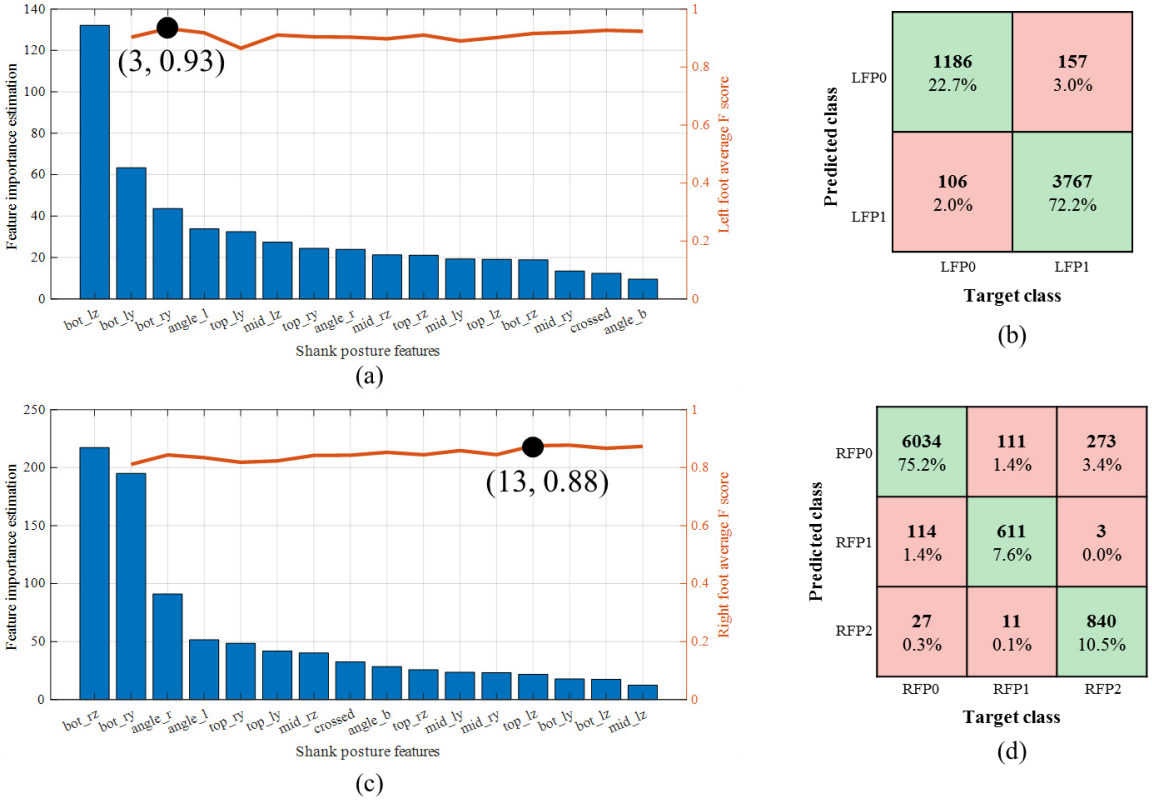


Figure 6.9. (a) feature importance estimation for classifying two left foot positions. (b) Confusion matrix of the best classification of left foot positions. (c) feature importance estimation for classifying three right foot positions. (d) Confusion matrix of the best classification of right foot positions.

In contrast to the classification method based on pressure sensors, the average F1 Score of two left foot positions by using the depth camera was not changed (0.93), but the average F1 Score of the three right foot positions is much higher (0.74 vs 0.88).

The biggest limitation of the proposed method is concerned with the assumption that the drivers in this study were not normally dressed because of the use of motion capture system for acquiring ground truth motion data. In real driving conditions for which people may wear trousers or skirts, cloth deformation will certainly challenge the robustness of the proposed method. Therefore, further validation is needed.

6.3 Sensor fusion

To test if the classification of feet positions could benefit from both the pressure sensors and the depth camera, two sensor fusion methods were investigated. The first method was based on the fusion of representative features from different sensors, including the selected pressure features and the selected shank postural features. The second method involved the fusion of preliminary decisions (class scores) made by each sensor. The final classification results are given in **Table 6.2**. Results show that sensor fusion can improve the classification accuracy of left foot positions, regardless of which fusion method was used. Regarding the classification accuracy or right foot positions, the result of the feature level based fusion is bit lower than using depth camera alone (average F1 Score 0.87 vs 0.88), but the decision level based fusion (average F1 Score 0.92) brings improvement with respect to using either depth camera (average F1 Score 0.74) or pressure sensors alone (average F1 Score 0.88). The results suggest that pressure sensors could be used as a good supplementary to the depth camera.

Table 6.2. F1 Scores of the classification of feet positions using different fusion methods.

Fusion method	Right foot (N=8024)				Left foot (N=5216)		
	RFP0	RFP1	RFP2	Avg	LFP0	LFP1	Avg
Only depth camera	0.96	0.84	0.84	0.88	0.89	0.96	0.93
Only pressure sensors	0.92	0.70	0.61	0.74	0.96	0.90	0.93
Feature level	0.95	0.87	0.79	0.87	0.93	0.98	0.96
Decision level	0.97	0.92	0.87	0.92	0.96	0.98	0.97

6.4 Summary

This Chapter proposed three methods to predict the positions of driver's feet. By using the pressure sensors, relevant pressure features were extracted from the pressure distributions on seat pan and backrest and their importance was evaluated by the Random Forest Out-Of-Bag error estimations. Random Forest classifiers then were trained on the best combination of relative pressure feature values to independently distinguish three right foot positions (on accelerator, on brake pedal and on floor) and two left foot positions (on floor and on clutch). By LOO validation across 23 drivers, an average F1 score of 74% and 93% were achieved for the classification of right foot positions and left foot positions, respectively.

By using a depth camera, three shank section profiles at three levels were first extracted by inspecting the shank point cloud in the legroom space. Mean-shift algorithm was performed to cluster the profiles in order to obtain keypoints. The keypoints were associated with the shank using a top-down approach and the missing key point was recovered by interpolation. Finally,

shank postural features were evaluated and then used to train Random Forest classifiers to predict driver's feet positions. LOO cross-validation showed an average F1 score of 88% and 93% for the classification of right foot positions and left foot positions, respectively.

By using sensor fusion, the results show that the fusion of pressure sensors and the depth camera on decision level outperformed the methods based on either pressure sensor or depth camera alone.

Chapter 7

7 Conclusion and future works

The ultimate goal of this work is to develop a reliable driver posture monitoring system for improving the road traffic safety. As an initial effort towards this goal, we have created a large-scale driver posture dataset (AutoConduct) including data collected from an experiment as well as synthesized data generated from a data augmentation pipeline. Based on this dataset, the relationship between driver postures and body pressure distributions were investigated and different classifiers were proposed and evaluated. Posture recognition methods of driver's upper-body, head and lower-body based on a depth camera were proposed. Particular attention was paid to reduce the posture recognition errors of driver's upper-body by using our motion databases. The potential of information fusion from a depth camera and pressure sensors were investigated. This Chapter provides a brief summary of the main results, limitations and perspective for the future work.

7.1 Summary of main results

Real driver postural data of AutoConduct was collected from 23 drivers performing 42 activities. The real data featured almost full-body coverage including images of multimodalities from three depth cameras, pressure distributions on driver's seat and ground truth driver motions recorded by an optical Mocap system. The postural measurement from different sensors were temporally synchronized and spatially aligned. The driver motions were reconstructed using an articulated model to accurately estimate the joint centers and joint angles. The real data can be used as a good benchmark to evaluate a series of monitoring functions, ranging from head pose estimation, upper-body pose estimation and lower-body pose estimation.

A data augmentation pipeline based on the techniques developed in computer graphics has been established to generate a large number of synthetic yet realistic annotated body images. In contrast to the conventional data augmentation approaches, this method enables the automatic collection of image annotations and ground truth 3D postures while enriching the posture

samples at the same time. The synthetic data of AutoConduct is useful for the domain adaptation of existing generic posture recognition models.

Posture classification methods based on pressure sensors were proposed. Results showed that the classifiers trained on the best combination of manually constructed pressure features could predict five trunk postures with an average F1 Score of 0.91 and outperformed the existing methods. Regarding the feet positions, the pressure sensor based method could predict two left foot positions and three right foot positions with an average F1 Score of 0.93 and 0.74 respectively. As the classifiers were trained on the relative pressure features with respect to a reference posture, the classifiers were robust to the anthropometric variations across different drivers.

Taking the advantages of existing posture recognition algorithms, a posture estimation method consisting of body part localization and joint offset regression was proposed to predict the upper-body posture in 3D using a depth camera. Results showed that the convolutional neural network trained on our in-vehicle driver posture dataset led to smaller prediction errors compared to the original models pre-trained on the generic human posture datasets, showing the utility of our in-vehicle driver posture dataset.

To reduce the posture recognition errors of driver's upper-body based on a depth camera, a correction framework based on pre-built motion databases were proposed. Results showed that the frequently occurred posture recognition errors such as confusion and miss detection caused by the uncertainty of the algorithms or body occlusions could be reduced. By integrating the additional trunk postural information from pressure sensors, the percentage of the data frames with a predicted joint within 5 mm from the ground truth was improved from 91% to 93% on average across the seven joints.

Based on the recognized facial key points from the adapted convolutional neural network, two baseline methods respectively based on Rigid Body Matching (RBM) and Random Forest Regression (RFR) were proposed to estimate the head orientation and position. Results showed that the RFR method was less sensitive to the incompleteness and noise of the head key points than the RBM method. Using the RFR method, the balanced mean errors were less than 11° and 2 cm respectively for the head orientation and position in 96.3% of the cases including the easy and moderate sub-datasets.

A posture recognition method based on a depth camera was proposed to predict the feet positions. The results showed that the shank postural information extracted from the point cloud could serve as useful cues for the prediction of feet positions. With the best combination of relevant features, an average F1 Score of 0.93 and 0.88 was achieved for the classification of

two left foot positions and three right foot positions, respectively. By fusing the information from pressure sensors, it was observed that better results could be obtained.

7.2 Limitations

The main limitation of the present work is that the proposed posture recognition methods were only evaluated on the existing experimental data, without considering the real road driving conditions. To ensure the reliability of motion capture, the drivers were not normally dressed and the mockup was simplified. Therefore, the images of driver's body were less realistic. Consequently, the evaluation of vision-based methods may be biased particularly for the posture estimation methods of driver's upper-body and lower-body. In addition, the inevitable vibration from the engine, the inertia caused by vehicle acceleration and the unevenness of road surface may compromise the effectiveness of pressure measurement from the driver seat, leaving the performance of the classifiers somewhat unknown.

Concerning the data augmentation pipeline, it is subject to several limitations. First, the driver seat and steering wheel in the car model is not fully rigged. Consequently, they cannot be adjusted to suit the driver's seating preference. Second, the motion constraints imposed by the vehicle interior was not considered in the motion retargeting process. Furthermore, challenging camouflage such as the occlusions caused by objects like handbags, smartphones, tablets or bottles etc., were not considered in the current framework.

Concerning the parameters used in upper body posture correction using a motion database, such as the number of filtered Nodes, the number of filtered Edges, the number of similar postures to be selected, etc., were not systematically investigated. Therefore, the performance achieved in this work may be suboptimal.

7.3 Future work

Based on the results and the limitations of the current work, further efforts are needed to bring the driver posture monitoring systems closer to real applications.

(1) Improvement of the proposed data augmentation pipeline. To generate more realistic in-vehicle postures, the motion retargeting process will be refined by considering the interior of different car models. To generate challenging posture samples, objects such as hand-held cell phone, books and tablets will be modeled and incorporated into the animation. With help of the improved pipeline, more synthetic image data and pressure data will be generated and they will be shared with the researchers in the community to facilitate the progress into accurate driver posture monitoring.

(2) Validation of the proposed monitoring functions in different experimental settings. The proposed posture recognition methods based on different sensors and the posture correction framework of driver's upper-body were only tested on the data collected from a vehicle mock-up under lab conditions. To objectively evaluate the performance of these functions, cross dataset validation is needed and the test on real road is needed.

(3) Investigation of more performant posture recognition methods based on AutoConduct dataset. This work focused on the adaptation of a few existing general-purpose posture recognition algorithms for driver posture monitoring. In the future, more existing algorithms will be adapted and new algorithms better suited to in-vehicle monitoring will be explored. Furthermore, the optimum camera placement for monitoring driver's upper-body and head will be investigated.

(4) Applications of the monitoring system for the development of human-centered active or passive safety systems. The scope of this thesis was limited to the development of posture estimation algorithms in order to obtain fundamental postural information from raw input. In the future, the proposed system will be used to identify the crucial postures or critical postural indicators useful for evaluating driver's state, which will be integrated into the development of advanced driver assistance systems or smart restraint systems.

(5) Fusion with the information from vehicle and environment. This thesis only focused on one element of the driving process, i.e., the driver. For a well-designed active/passive safety system, the information fusion of the driver, the vehicle and the driving environment are necessary.

List of publications

Journal papers

Zhao, M., Beurier, G., Wang, H., & Wang, X. (2021). Exploration of Driver Posture Monitoring Using Pressure Sensors with Lower Resolution. Sensors, 21(10), 3346.

Zhao, M., Beurier, G., Wang, H., & Wang, X. (2021). Driver posture monitoring in highly automated vehicles using pressure measurement. Traffic injury prevention, 22(4), 278-283.

Wang, X., Beurier, G., Zhao, M., & Obadia, J. M. (2021). Objective and subjective evaluation of a new airplane seat with an optimally pre-shaped foam support. Work, (Preprint), 1-14.

Wang, H. Y., Zhao, M. M., Beurier, G., & Wang, X. G. (2019). Automobile driver posture monitoring systems: A review. China J. Highw. Transp, 32(2), 1-18.

Conference papers

Hanson, L. (2020, September). A Pipeline for Creating In-Vehicle Posture Database for Developing Driver Posture Monitoring Systems. In DHM2020: Proceedings of the 6th International Digital Human Modeling Symposium, August 31-September 2, 2020 (Vol. 11, p. 187). IOS Press.

Zhao, M., Beurier, G., Wang, H., & Wang, X. (2020, August). Extraction of pressure features for predicting driver posture. In Proceedings of the International Research Conference on the Biomechanics of Impact, Munich, Germany (pp. 398-409).

Zhao, M., Beurier, G., Wang, H., & Wang, X. (2020, May). Driver Posture Prediction Using Pressure Measurement and Deep Learning, In Proceedings of the International Research Conference on the Biomechanics of Impact, Beijing, China (pp. 102-105).

Zhao, M., Beurier, G., Wang, H., & Wang, X. (2019, September). Detection of Driver Posture Change by Seat Pressure Measurement. In Proceedings of the International Research Conference on the Biomechanics of Impact, Florence, Italy (p. 84-85).

Zhao, M., Beurier, G., Wang, H., & Wang, X. (2018). In vehicle diver postural monitoring using a depth camera kinect (No. 2018-01-0505). SAE world congress 2018.

Papers to be submitted

AutoConduct: A multi-modal and multi-view in-vehicle driver posture dataset for postural monitoring. (Signal Processing: Image Communication)

Monitoring of Driver Lower Limbs Using a Depth Camera. (Sensors/IEEE Sensors)

A data-driven approach for driver upper-body posture monitoring. (IEEE Sensors)

Bibliography

Altman, Naomi S. 1992. 'An introduction to kernel and nearest-neighbor nonparametric regression', *Am. Stat.*, 46: 175-85.

Ansari, Shahzeb, Haiping Du, and Fazel Naghdy. 2020. "Driver's Foot Trajectory Tracking for Safe Maneuverability Using New Modified reLU-BiLSTM Deep Neural Network." In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 4392-97. IEEE.

Baltrusaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. "Openface 2.0: Facial behavior analysis toolkit." In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 59-66. IEEE.

Beanland, Vanessa, Michael Fitzharris, Kristie L Young, and Michael G Lenné. 2013. 'Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study', *Accident Anal Prev.*, 54: 99-107.

Berretti, Stefano, Mohamed Daoudi, Pavan Turaga, and Anup Basu. 2018. 'Representation, analysis, and recognition of 3D humans: A survey', *ACM Transactions on Multimedia Computing, Communications, Applications*, 14: 1-36.

Billah, Tashrif, SM Mahbubur Rahman, M Omair Ahmad, and MNS Swamy. 2018. 'Recognizing distractions for assistive driving by tracking body parts', *IEEE Transactions on Circuits and Systems for Video Technology*, 29: 1048-62.

Bishop, Gary, and Greg Welch. 2001. "An introduction to the kalman filter." In *SIGGRAPH 2001, Course 8*, 41. Los Angeles, CA, USA.

Borges, João, Sandro Queirós, Bruno Oliveira, Helena Torres, Nelson Rodrigues, Victor Coelho, Johannes Pallauf, José Henrique Brito, José Mendes, and Jaime C Fonseca. 2021. 'A system for the generation of in-car human body pose datasets', *Machine Vision and Applications*, 32: 1-15.

Borghi, Guido, Elia Frigieri, Roberto Vezzani, and Rita Cucchiara. 2018. "Hands on the wheel: a dataset for driver hand detection and tracking." In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 564-70. IEEE.

Borghi, Guido, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. 2017. "Poseidon: Face-from-depth for driver pose estimation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4661-70.

Bourahmoune, Katia, and Toshiyuki Amagasa. 2019. "AI-powered posture training: application of machine learning in sitting posture recognition using the lifechair smart cushion." In *28th International Joint Conference on Artificial Intelligence*, 5808-14. Macao, China: AAAI Press.

Breiman, Leo. 1996. 'Out-of-bag estimation'.

Breiman, Leo. 2001. 'Random forests', *Machine Learning.*, 45: 5-32.

Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. "Realtime multi-person 2d pose estimation using part affinity fields." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291-99.

Chai, Jinxiang, and Jessica K Hodgins. 2005. 'Performance animation from low-dimensional control signals.' in, *ACM SIGGRAPH 2005 Papers*.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. 'LIBSVM: A library for support vector machines', *ACM T Intel Syst Tec.*, 2: 1-27.

Chen, Wenzheng, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2016. "Synthesizing training images for boosting human 3d pose estimation." In *2016 Fourth International Conference on 3D Vision (3DV)*, 479-88. IEEE.

Cheng, Shinko Y, and Mohan M Trivedi. 2004. "Human posture estimation using voxel data for" smart" airbag systems: issues and framework." In *IEEE Intelligent Vehicles Symposium*, edited by IEEE, 84-89. Parma, Italy: IEEE.

Cheng, Shinko Y, and Mohan M Trivedi. 2010. 'Vision-based infotainment user determination by hand recognition for driver assistance', *IEEE Transactions on Intelligent Transportation Systems*, 11: 759-64.

Cheng, Shinko Yuanhsien, and Mohan M Trivedi. 2006. 'Turn-intent analysis using body pose for intelligent driver assistance', *IEEE Pervasive Computing*, 5: 28-37.

Choi, Minho, Gyogwon Koo, Minseok Seo, and Sang Woo Kim. 2018. 'Wearable Device-Based System to Monitor a Driver's Stress, Fatigue, and Drowsiness', *IEEE Transactions on Instrumentation and Measurement*, 67: 634-45.

Collins, William, Larry Evans, and Rick Hughes. 2015. "Driver brake and accelerator controls and pedal misapplication rates in North Carolina." In.

Comaniciu, Dorin, and Peter Meer. 2002. 'Mean shift: A robust approach toward feature space analysis', *IEEE Transactions on Pattern Analysis & Machine Intelligence*: 603-19.

Craye, Céline, and Fakhri Karray. 2015. 'Driver distraction detection and recognition using RGB-D sensor', *arXiv preprint arXiv:.00250*.

Cruz, Steve Dias Da, Oliver Wasenmuller, Hans-Peter Beise, Thomas Stifter, and Didier Stricker. 2020. "Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 973-82.

Das, Nikhil, Eshed Ohn-Bar, and Mohan M Trivedi. 2015. "On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics." In *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*, edited by IEEE, 2953-58. Las Palmas, Spain: IEEE.

Deo, Nachiket, and Mohan M Trivedi. 2019. 'Looking at the driver/rider in autonomous vehicles to predict take-over readiness', *IEEE Transactions on Intelligent Vehicles*, 5: 41-52.

Ding, Ming, Tatsuya Suzuki, and Tsukasa Ogasawara. 2017. "Estimation of driver's posture using pressure distribution sensors in driving simulator and on-road experiment." In *IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 215-20. Beijing, China: IEEE.

Dingus, Thomas A, Sheila G Klauer, Vicki L Neale, Andy Petersen, Suzanne E Lee, JD Sudweeks, Miguel A Perez, Jonathan Hankey, DJ Ramsey, and Santosh Gupta. 2006. "The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment." In. Virginia, Minnesota.

Eraqi, Hesham M, Yehya Abouelnaga, Mohamed H Saad, and Mohamed N Moustafa. 2019. 'Driver distraction identification with an ensemble of convolutional neural networks', *Journal of Advanced Transportation*, 2019.

Feld, Hartmut, Bruno Mirbach, Jigyasa Katrolia, Mohamed Selim, Oliver Wasenmüller, and Didier Stricker. 2021. 'Dfki cabin simulator: A test platform for visual in-cabin monitoring functions.' in, *Commercial Vehicle Technology 2020/2021* (Springer).

Filatov, Anton, John M Scanlon, Alexander Bruno, Sri Sai Kameshwari Danthurthi, and Jacob Fisher. 2019. "Effects of Innovation in Automated Vehicles on Occupant Compartment Designs, Evaluation, and Safety: A Review of Public Marketing, Literature, and Standards." In.: SAE Technical Paper.

Frank, Sebastian, and Arjan Kuijper. 2019. 'Robust driver foot tracking and foot gesture recognition using capacitive proximity sensing', *Journal of Ambient Intelligence and Smart Environments*, 11: 221-35.

Fu, Xianping, Guan Xiao, Eli Peli, Hongbo Liu, and Gang Luo. 2013. 'Automatic Calibration Method for Driver's Head Orientation in Natural Driving Environment', *IEEE Transactions on Intelligent Transportation Systems*, 14: 303-12.

Girshick, Ross, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. 2011. "Efficient regression of general-activity human poses from depth images." In *2011 International Conference on Computer Vision*, 415-22. IEEE.

Guo, Zhibo, Huajun Liu, Qiong Wang, and Jingyu Yang. 2007. "A Fast Algorithm Face Detection and Head Pose Estimation for Driver Assistant System." In *8th International Conference on Signal Processing*, edited by IEEE, 1733-37. Beijing, China: IEEE.

Han, Fei, Brian Reily, William Hoff, and Hao Zhang. 2017. 'Space-time representation of people based on 3D skeletal data: A review', *Computer Vision Image Understanding*, 158: 85-105.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media).

Hu, Tiancheng, Sumit Jha, and Carlos Busso. 2020. "Robust driver head pose estimation in naturalistic conditions from point-cloud data." In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 1176-82. IEEE.

Jegham, Imen, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2020. 'A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD', *Signal Processing: Image Communication*, 88: 115960.

Jiang, Binhui, Hongze Ren, Feng Zhu, Clifford Chou, and Zhonghao Bai. 2020. "A Preliminary Study on the Restraint System of Self-Driving Car." In.: SAE Technical Paper.

Jonas, Rachel, Caroline Crump, Robyn Brinkerhoff, Audra Krake, Heather Watson, and Douglas Young. 2018. "Variability in Circumstances Underlying Pedal Errors: An Investigation Using the National Motor Vehicle Crash Causation Survey." In.: SAE Technical Paper.

Kalman, R. E. 1960. 'A New Approach to Linear Filtering and Prediction Problems', *Journal of Basic Engineering*, 82: 35-45.

Kato, T., T. Fujii, and M. Tanimoto. 2004. "Detection of driver's posture in the car by using far infrared camera." In *Intelligent Vehicle Symposium*, edited by IEEE, 339-44. Parma, Italy: IEEE.

Kondyli, A, A Barmpoutis, VP Sisiopiku, L Zhang, L Zhao, MM Islam, SS Patil, and S Rostami Hosuri. 2018. 'A 3D body posture analysis framework during merging and lane-changing maneuvers', *Journal of Transportation Safety, Security*, 10: 411-28.

Kondyli, Alexandra, Virginia P. Sisiopiku, Liangke Zhao, and Angelos Barmpoutis. 2015. 'Computer Assisted Analysis of Drivers' Body Activity Using a Range Camera', *IEEE Intelligent Transportation Systems Magazine*, 7: 18-28.

Krotosky, S. J, S. Y Cheng, and M. M Trivedi. 2008. "Real-time stereo-based head detection using size, shape and disparity constraints." In *Intelligent Vehicles Symposium*, edited by IEEE, 550-56. Las Vegas, NV, USA: IEEE.

Large, David R, Gary Burnett, Andrew Morris, Arun Muthumani, and Rebecca Matthias. 2017. "A longitudinal simulator study to explore drivers' behaviour during highly-automated driving." In *International Conference on Applied Human Factors and Ergonomics*, 583-94. Los Angeles, California, USA: Springer.

Lee, Dae Seok, Teak Wei Chong, and Boon Giin Lee. 2017. 'Stress Events Detection of Driver by Wearable Glove System', *IEEE Sensors Journal*, 17: 194-204.

Lee, Seul Chan, Sol Hee Yoon, and Yong Gu Ji. 2021. 'Effects of non-driving-related task attributes on takeover quality in automated vehicles', *International Journal of Human–Computer Interaction*, 37: 211-19.

Lee, Sung Joo, Jaeik Jo, Ho Gi Jung, Ryoung Park Kang, and Jaihie Kim. 2011. 'Real-Time Gaze Estimator Based on Driver's Head Orientation for Forward Collision Warning System', *IEEE Transactions on Intelligent Transportation Systems*, 12: 254-67.

Leledakis, Alexandros, Jonas Östh, Johan Davidsson, and Lotta Jakobsson. 2021. 'The influence of car passengers' sitting postures in intersection crashes', *Accident Analysis & Prevention*, 157: 106170.

Li, Peng, Meiqi Lu, Zhiwei Zhang, Donghui Shan, and Yang Yang. 2019. "A Novel Spatial-Temporal Graph for Skeleton-based Driver Action Recognition." In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 3243-48. IEEE.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. "Microsoft coco: Common objects in context." In *European conference on computer vision*, 740-55. Springer.

Liu, Long, Zhelong Wang, and Sen Qiu. 2020. 'Driving Behavior Tracking and Recognition Based on Multi-Sensors Data Fusion', *IEEE Sensors Journal*.

Lu, Zhenji, Riender Happee, Christopher DD Cabrall, Miltos Kyriakidis, and Joost de Winter. 2016. 'Human factors of transitions in automated driving: A general framework and literature survey', *Transport Res F-Traf.*, 43: 183-98.

Ma, Congcong, Wenfeng Li, Raffaele Gravina, and Giancarlo Fortino. 2017. 'Posture detection based on smart cushion for wheelchair users', *Sensors*, 17: 719.

Martin, Manuel, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. 2019. "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2801-10.

Martin, Sujitha, Ashish Tawari, Erik Murphy-Chutorian, Shinko Y. Cheng, and Mohan Trivedi. 2012. "On the design and evaluation of robust head pose for visual user interfaces:algorithms, databases, and comparisons." In *4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, edited by Bastian Pfleging and Marc Kurz, 149-54. Portsmouth, New Hampshire, USA: ACM.

Martin, Sujitha, Kevan Yuen, and Mohan M Trivedi. 2016. "Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge." In *2016 IEEE Intelligent Vehicles Symposium (IV)*, 1010-14. IEEE.

Martínez-González, Angel, Michael Villamizar, Olivier Canévet, and Jean-Marc Odobez. 2019. 'Efficient convolutional neural networks for depth-based multi-person pose estimation', *IEEE Transactions on Circuits and Systems for Video Technology*, 30: 4207-21.

McGehee, Daniel V, Cheryl A Roe, Linda Ng Boyle, Yuqing Wu, Kazutoshi Ebe, James Foley, and Linda Angell. 2016. 'The wagging foot of uncertainty: data collection and reduction methods for examining foot pedal behavior in naturalistic driving', *SAE International journal of transportation safety*, 4: 289-94.

Mergl, Christian, Margit Klendauer, Claude Mangen, and Heiner Bubb. 2005. "Predicting long term riding comfort in cars by contact forces between human and seat." In.: SAE Technical Paper.

Monnier, G, X Wang, J Trasbot, and Human Factors Engineering. 2009. 'A motion simulation tool for automotive interior design', *Handbook of Digital Human Modeling: Research for Applied Ergonomics*.

Murphy-Chutorian, Erik, Anup Doshi, and Mohan Manubhai Trivedi. 2007. "Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation." In *Intelligent Transportation Systems Conference*, edited by IEEE, 709-14. Seattle, WA, USA: IEEE.

Naujoks, Frederik, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. 2017. "A review of non-driving-related tasks used in studies on automated driving." In *International Conference on Applied Human Factors and Ergonomics*, 525-37. Springer.

Née, Mélanie, Benjamin Contrand, Ludivine Orriols, Cédric Gil-Jardiné, Cedric Galéra, and Emmanuel Lagarde. 2019. 'Road safety and distraction, results from a responsibility case-control study among a sample of road users interviewed at the emergency room', *Accident Anal Prev.*, 122: 19-24.

Niu, Jianwei, Xiai Wang, Xingguo Liu, Dan Wang, Hua Qin, and Yunhong Zhang. 2019. 'Effects of mobile phone use on driving performance in a multiresource workload scenario', *Traffic injury prevention*, 20: 37-44.

Ohn-Bar, Eshed, Sujitha Martin, and Mohan Manubhai Trivedi. 2013. 'Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies', *Journal of Electronic Imaging*, 22: 04111901-10.

Okabe, Kenta, Keiichi Watanuki, Kazunori Kaede, and Keiichi Muramatsu. 2018. "Study on estimation of driver's state during automatic driving using seat pressure." In *International Conference on Intelligent Human Systems Integration*, 35-41. Dubai, United Arab Emirates: Springer.

Pagliari, Diana , and Livio Pinto. 2015. 'Calibration of Kinect for Xbox One and Comparison between the Two Generations of Microsoft Sensors', *Sensors*, 15: 27569-89.

Pan, Sinno Jialin, and Qiang Yang. 2009. 'A survey on transfer learning', *IEEE Transactions on knowledge and data engineering*, 22: 1345-59.

Park, Byoung-Keon D, Sheila Ebert, Carl Miller, and Matthew P Reed. 2020. "Robust Markerless 3D Head Tracking of a Vehicle Occupant Using OpenPose." In *International Conference on Applied Human Factors and Ergonomics*, 336-42. Springer.

Park, Jangwoon, Sheila M Ebert, Matthew P Reed, and Jason Hallman. 2016. 'Statistical models for predicting automobile driving postures for men and women including effects of age', *Human Factors*, 58: 261-78.

Peng, Junfeng, Jules Panda, Serge Van Sint Jan, and Xuguang Wang. 2015. 'Methods for determining hip and lumbosacral joint centers in a seated position from external anatomical landmarks', *Journal of biomechanics*, 48: 396-400.

Plantard, Pierre, Edouard Auvinet, Anne Sophie Le Pierres, and Franck Multon. 2015. 'Pose Estimation with a Kinect for Ergonomic Studies: Evaluation of the Accuracy Using a Virtual Mannequin', *Sensors*, 15: 1785-803.

Plantard, Pierre, Hubert P. H. Shum, and Franck Multon. 2017. 'Filtered pose graph for efficient kinect pose reconstruction', *Multimedia Tools and Applications*, 76: 4291-312.

Rangesh, Akshay, and Mohan Trivedi. 2019. "Forced spatial attention for driver foot activity classification." In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0-0.

Reed, Matthew P, Miriam A Manary, and Lawrence W Schneider. 1999. "Methods for measuring and representing automobile occupant posture." In.: SAE Technical Paper.

Ren, Jindong, Xiaoming Du, Tao Liu, Honghao Liu, Meng Hua, and Qun Liu. 2017. "An Integrated Method for Evaluation of Seat Comfort Based on Virtual Simulation of the Interface Pressures of Driver with Different Body Sizes." In.: SAE Technical Paper.

Roth, Markus, and Dariu M Gavrila. 2019. "Dd-pose-a large-scale driver head pose benchmark." In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 927-34. IEEE.

Sathyanarayana, A, S Nageswaren, H Ghasemzadeh, and R Jafari. 2008. "Body sensor networks for driver distraction identification." In *International Conference on Vehicular Electronics and Safety*, edited by IEEE, 120-25. Columbus, OH, USA: IEEE.

Schwarz, Anke, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. 2017. "Driveahead-a large-scale driver head pose dataset." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1-10.

Selim, Mohamed, Ahmet Firintepe, Alain Pagani, and Didier Stricker. 2020. "AutoPOSE: Large-scale Automotive Driver Head Pose and Gaze Dataset with Deep Head Orientation Baseline." In *VISIGRAPP (4: VISAPP)*, 599-606.

Shia, Victor A, Yiqi Gao, Ramanarayan Vasudevan, Katherine Driggs Campbell, Theresa Lin, Francesco Borrelli, and Ruzena Bajcsy. 2014. 'Semiautonomous vehicular control using driver modeling', *IEEE Transactions on Intelligent Transportation Systems*, 15: 2696-709.

Shin, Hangsik, Kwanghyun Kim, Jeawon Son, and Miri Kim. 2015. 'Preliminary study for driver's posture correction support system based on seat-embedded pressure sensing platform.' in, *Computer Science and Its Applications* (Springer).

Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. "Real-time human pose recognition in parts from single depth images." In *Computer Vision and Pattern Recognition*, edited by IEEE, 1297-304. Colorado Springs, CO, USA: IEEE.

Shotton, Jamie, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Richard Moore, Richard Moore, Pushmeet Kohli, Antonio Criminisi, and Alex Kipman. 2013. 'Efficient Human Pose Estimation from Single Depth Images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35: 2821-40.

Shum, H. P., E. S. Ho, Y. Jiang, and S Takagi. 2013. 'Real-time posture reconstruction for Microsoft Kinect', *IEEE Transactions on Cybernetics*, 43: 1357-69.

Sigal, Leonid, Alexandru O Balan, and Michael J Black. 2010. 'Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion', *International journal of computer vision*, 87: 4.

Singh, Santokh. 2015. "Critical reasons for crashes investigated in the national motor vehicle crash causation survey." In. Washington, DC, USA.

StateFarm. 2016. 'State farm distracted driver detection', Accessed January 15, 2021. https://www.kaggle.com/c/state-farm-distracted-driver-detection/.

Tawari, Ashish, Sujitha Martin, and Mohan Manubhai Trivedi. 2014. 'Continuous Head Movement Estimator for Driver Assistance: Issues, Algorithms, and On-Road Evaluations', *IEEE Transactions on Intelligent Transportation Systems*, 15: 818-30.

Toma, Madalina Ioana, Leon J. M Rothkrantz, and Csaba Antonya. 2012. "Car driver skills assessment based on driving postures recognition." In *IEEE International Conference on Cognitive Infocommunications*, edited by IEEE, 439-46. Kosice, Slovakia: IEEE.

Torres, Helena R, Bruno Oliveira, Jaime Fonseca, Sandro Queirós, João Borges, Nélson Rodrigues, Victor Coelho, Johannes Pallauf, José Brito, and José Mendes. 2019. "Real-Time Human Body Pose Estimation for In-Car Depth Images." In *Doctoral Conference on Computing, Electrical and Industrial Systems*, 169-82. Springer.

Tran, Cuong, Anup Doshi, and Mohan M Trivedi. 2011. "Pedal error prediction by driver foot gesture analysis: A vision-based inquiry." In *2011 IEEE Intelligent Vehicles Symposium (IV)*, 577-82. IEEE.

Tran, Cuong, Anup Doshi, and Mohan Manubhai Trivedi. 2012. 'Modeling and prediction of driver behavior by foot gesture analysis', *Computer vision and image understanding*, 116: 435-45.

Tran, Cuong, and Mohan M. Trivedi. 2009. "Introducing "XMOB": Extremity Movement Observation Framework for Upper Body Pose Tracking in 3D." In *IEEE International Symposium on Multimedia*, edited by IEEE, 446-47. San Diego, CA, USA: IEEE.

Trivedi, M. M., Shinko Yuanhsien Cheng, E. M. C. Childers, and S. J. Krotosky. 2004. 'Occupant posture analysis with stereo and thermal infrared video: algorithms and experimental evaluation', *IEEE Transactions on Vehicular Technology*, 53: 1698-712.

Veeraraghavan, H., S. Atev, N. Bird, and P. Schrater. 2005. "Driver activity monitoring through supervised and unsupervised learning." In *International IEEE Conference on Intelligent Transportation Systems*, edited by IEEE, 580-85. Vienna, Austria: IEEE.

Venturelli, Marco, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. 2017. 'From depth data to head pose estimation: a siamese approach', *arXiv preprint arXiv:.03624*.

Vergnano, Alberto, and Francesco Leali. 2019. "Out of position driver monitoring from seat pressure in dynamic maneuvers." In *International Conference on Intelligent Human Systems Integration*, 76-81. San Diego, CA, USA: Springer.

Wang, Hongyan, Mingming Zhao, Georges Beurier, and Xuguang Wang. 2019. 'Automobile Driver Posture Monitoring Systems: A Review', *China J. Highw. Transp.*, 32: 1-18.

Wang, Pichao, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. 2018. 'RGB-D-based human motion recognition with deep learning: A survey', *Computer vision and image understanding*, 171: 118-39.

Wu, Hequan, Haibin Hou, Ming Shen, King H Yang, and Xin Jin. 2020. 'Occupant kinematics and biomechanics during frontal collision in autonomous vehicles—can rotatable seat provides additional protection?', *Comput Method Biomec.*: 1-10.

Wu, Yuqing, Linda Ng Boyle, Daniel McGehee, Cheryl A Roe, Kazutoshi Ebe, and James Foley. 2017. 'Foot placement during error and pedal applications in naturalistic driving', *Accident Analysis & Prevention*, 99: 102-09.

Xing, Yang, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang. 2019. 'Driver activity recognition for intelligent vehicles: A deep learning approach', *IEEE Transactions on Vehicular Technology*, 68: 5379-90.

Xing, Yang, Chen Lv, Zhaozhong Zhang, Huaji Wang, Xiaoxiang Na, Dongpu Cao, Efstathios Velenis, and Fei Yue Wang. 2017. 'Identification and Analysis of Driver Postures for In-Vehicle Driving Activities and Secondary Tasks Recognition', *IEEE Transactions on Computational Social Systems*, PP: 1-14.

Yamada, Takahiro, Hidetsugu Irie, Masahiro Kunitake, Eiji Nagano, and Shuichi Sakai. 2018. "Estimating driver's readiness by understanding driving posture." In *2018 IEEE International Conference on Consumer Electronics (ICCE)*, 1-4. IEEE.

Yan, Chao, Frans Coenen, and Bai Ling Zhang. 2014. 'Driving Posture Recognition by Joint Application of Motion History Image and Pyramid Histogram of Oriented Gradients', *Advanced Materials Research*, 846-847: 1102-05.

Yang, CY David, and Donald L Fisher. 2021. "Safety impacts and benefits of connected and automated vehicles: How real are they?" In.: Taylor & Francis.

Yoon, Sol Hee, and Yong Gu Ji. 2019. 'Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts', *Transport Res F-Traf.*, 60: 620-31.

Yuen, Kevan, and Mohan Manubhai Trivedi. 2019. 'Looking at hands in autonomous vehicles: A ConvNet approach using part affinity fields', *IEEE Transactions on Intelligent Vehicles*, 5: 361-71.
Zhang, Zhengyou 2000. 'A flexible new technique for camera calibration', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22: 1330-34.
Zhao, C. H, B. L Zhang, J He, and J Lian. 2012. 'Recognition of driving postures by contourlet transform and random forests', *Intelligent Transport Systems*, 6: 161-68.
Zhao, Mingming, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2018. "In vehicle diver postural monitoring using a depth camera kinect." In *WCX™ World Congress Experience*, 1-9. Detroit, Michigan: SAE International.
Zhao, Mingming, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2020a. "Driver Posture Prediction Using Pressure Measurement and Deep Learning." In *IRCOBI Asia Conference*. Beijing, China. (to be published).
Zhao, Mingming, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2020b. "Extraction of pressure features for predicting driver posture." In *International Research Conference on the Biomechanics of Impact*, 398-409. Munich, Germany.
Zhao, Mingming, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2020c. 'A Pipeline for Creating In-Vehicle Posture Database for Developing Driver Posture Monitoring Systems.' in L Hanson (ed.), *DHM2020: Proceedings of the 6th International Digital Human Modeling Symposium* (IOS Press: Skövde, Sweden).
Zhao, Mingming, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2021a. 'Driver posture monitoring in highly automated vehicles using pressure measurement', *TRAFFIC INJ PREV*: 1-6.
Zhao, Mingming, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2021b. 'Exploration of Driver Posture Monitoring Using Pressure Sensors with Lower Resolution', *Sensors*, 21: 3346.
Zhou, Qing, Peijun JI, Yi Huang, Bingbing Nie, and Yuan Huang. 2017. 'Challenges and opportunities of smart occupant protection against motor vehicle collision accidents in future traffic environment', *Journal of Automotive Safety and Energy*, 8: 333.

Appendices

Table S1. Driver task list.

ID	Task	Description	Start/end posture
(12*)	Primary Driving Task		
1	Standard driving	Driver activities that are frequently involved in normal driving.	Hands on steering wheel, eye on the road, trunk upright, right foot on accelerator pedal, left foot on floor
2	Accelerating		
3	Braking		
4	Changing gear		
5	Checking center rear-view mirror		
6	Checking left rear-view mirror		
7	Checking right rear-view mirror		
8	Adjusting navigation system		
9	Turning steering wheel 90° to left		
10	Turning steering wheel 90° to right		
11	Turning steering wheel 360° to left		
12	Turning steering wheel 360° to right		
(12)	Secondary Driving Task		
13	Bringing coffee to mouth	Secondary activities driver may perform while driving. During these activities, at least one hand is on steering wheel and right foot stays on pedal.	Hands on steering wheel, looking forward, trunk upright, right foot on accelerator pedal, left foot on floor
14	Bringing phone to left ear from left door		
15	Looking center control panel		
16	Checking dashboard		
17	Looking over left shoulder		
18	Looking over right shoulder		
19	Looking out left window		
20	Looking out right window		
21	Reaching an object from passenger seat		
22	Picking up an object from floor		
23	Reading text from hand-held cell phone		
24	Dialing hand-held cell phone		
(9)	Non-Driving-Related Task		
25	Reading book	Activities driver may be engaged in when automated driving systems are in charge of vehicle control. These activities take driver's hands off the wheel and feet off the pedals.	Hands on thighs, looking forward, trunk upright, right foot on accelerator pedal, left foot on floor
26	Texting on hand-held cell phone		
27	Adjusting infotainment panel		
28	Searching bag for an object		
29	Putting hands on legs		
30	Crossing arms		
31	Crossing legs		
32	Resting		
33	Sleeping		
(9)	General Driver Body Action		
34	Right arm abduction	Body movements that are frequently involved in in-vehicle driver activities. These movements were intended for more driver postural variations.	Hands on thighs or steering wheel, looking forward, trunk upright, right foot on accelerator pedal, left foot on floor
35	Right arm flexion		
36	Right elbow flexion		
37	Bending head		
38	Rotating head		
39	Trunk flexion		
40	Trunk rotation		
41	Releasing right foot from pedal		
42	Sitting up straight		

(*) The number of tasks belonging to the same type.

Table S2. Pressure features extracted from segmented pressure mats.

Sensing area	Contact Area Proportion (CAP)	Center of Pressure (COP)	Pressure Ratio pair (PR, area 1 / area 2)
Backrest pressure mat			
{Bi}(i $= 1,2,\dots,12$)	$\frac{\sum_{area} n(i,j) = 1}{N}$ Where $n(i,j) = 1$ if the cell at position (i,j) within a sensing area on the backrest is occupied, 0 otherwise. $N = 44 \times 42$	The COP of a sensing area in both up-down and left-right directions on the backrest	Bi/B
B			--
B1B2			B1B2/B
B3B4			B3B4/B
B5B6			B5B6/B
B7B8			B7B8/B
B9B10			B11B12/B
B11B12			B11B12/B
B1B5B9			B1B5B9/B
B2B6B10			B2B6B10/B
B3B7B11			B3B7B11/B
B4B8B12			B4B8B12/B
B1B2B3B4			B1B2B3B4/B
B5B6B7B8			B5B6B7B8/B
B9B10B11B12			B9B10B11B12/B
B1B2B5B6B9B10			B1B2B5B6B9B10/B
B3B4B7B8B11B12			B3B4B7B8B11B12/B
Seat pan pressure mat			
{Si}(i = 1,2, … , 8)	$\frac{\sum_{area} n(i,j) = 1}{N}$ Where $n(i,j) = 1$ if the cell at position (i,j) within a sensing area on the seat pan is occupied, 0 otherwise. $N = 44 \times 42$	The COP of a sensing area in both fore-aft and left-right directions on the seat pan	Si/S
S			--
S1S2			S1S2/S S1S2/S1S2S3S4
S3S4			S3S4/S
S5S6			S5S6/S
S7S8			S7S8/S S7S8/S5S6S7S8
S1S3			S1S3/S S1S3/S1S2S3S4 S1S3/S1S3S5S7
S2S4			S2S4/S
S5S7			S5S7/S S5S7/S5S6S7S8
S6S8			S6S8/S S6S8/S2S4S6S8
S1S2S3S4			S1S2S3S4/S
S5S6S7S8			S5S6S7S8/S
S1S3S5S7			S1S3S5S7/S
S2S4S6S8			S2S4S6S8/S
Note. If one sensing area (individual subarea or combine subareas) has no contact with driver's body, the pressure features related to this sensing area will be given NaN.			

Table S3. Best feature combinations for the classification of trunk postures and feet positions. X_COP_U (V) stands for the COP position of area X in left-right (up-down) direction on backrest and fore-aft (left-right) direction on seat pan. X_CAP is the contact proportion within area X. X_Y_PR is referred to as the ratio of the pressure sum between area X and area Y. For each body part, the features are ranked according to their importance estimated by OOB errors.

Important features used for the classification of trunk postures							
ID	Feature		ID	Feature		ID	Feature
1	B_CAP		10	B2_B_PR		19	S5S6S7S8_COP_V
2	B2B6B10_B_PR		11	S_CAP		20	B5B6B7B8_B_PR
3	B5_B_PR		12	S4_S_PR		21	S5S6S7S8_COP_U
4	B1B5B9_B_PR		13	S1S2S3S4_S_PR		22	S_COP_V
5	B_COP_U		14	B6_B_PR		23	S4_COP_V
6	S_COP_U		15	S7S8_S5S6S7S8_PR		24	B11_B_PR
7	B9_B_PR		16	B4B8B12_B_PR		25	B7_B_PR
8	S1S2S3S4_COP_U		17	S5_S_PR		26	S6S8_S2S4S6S8_PR
9	B_COP_V		18	S1_S_PR		27	S8_S_PR
Important features used for the classification of left foot positions							
ID	Feature		ID	Feature		ID	Feature
1	S4_CAP		9	S6_COP_V		17	B1B2B3B4_CAP
2	B9B10_CAP		10	B_CAP		18	B6_CAP
3	B10_CAP		11	B5B6_COP_U		19	B1_CAP
4	B4B8B12_B_PR		12	S_CAP		20	B1_COP_U
5	B1B2B5B6B9B10_CAP		13	S3S4_CAP		21	B9_CAP
6	B9B10B11B12_CAP		14	B1_B_PR		22	S7_CAP
7	S2_S_PR		15	S1S3_S1S2S3S4_PR		23	S4_COP_U
8	B5_CAP		16	B7B8_COP_U		24	B5_COP_V
Important features used for the classification of right foot positions							
ID	Feature		ID	Feature		ID	Feature
1	S1_S_PR		9	S8_CAP		17	S2S4S6S8_CAP
2	S8_COP_V		10	B10_CAP		18	S2_COP_V
3	B_CAP		11	S6_CAP		19	B3_COP_U
4	S5_COP_V		12	B1_B_PR		20	B2_CAP
5	B9_CAP		13	B5_CAP		21	B6_CAP
6	S5S7_COP_V		14	S7_COP_V		22	B12_B_PR
7	S7S8_COP_V		15	B12_CAP			
8	B11_COP_V		16	B1_CAP			

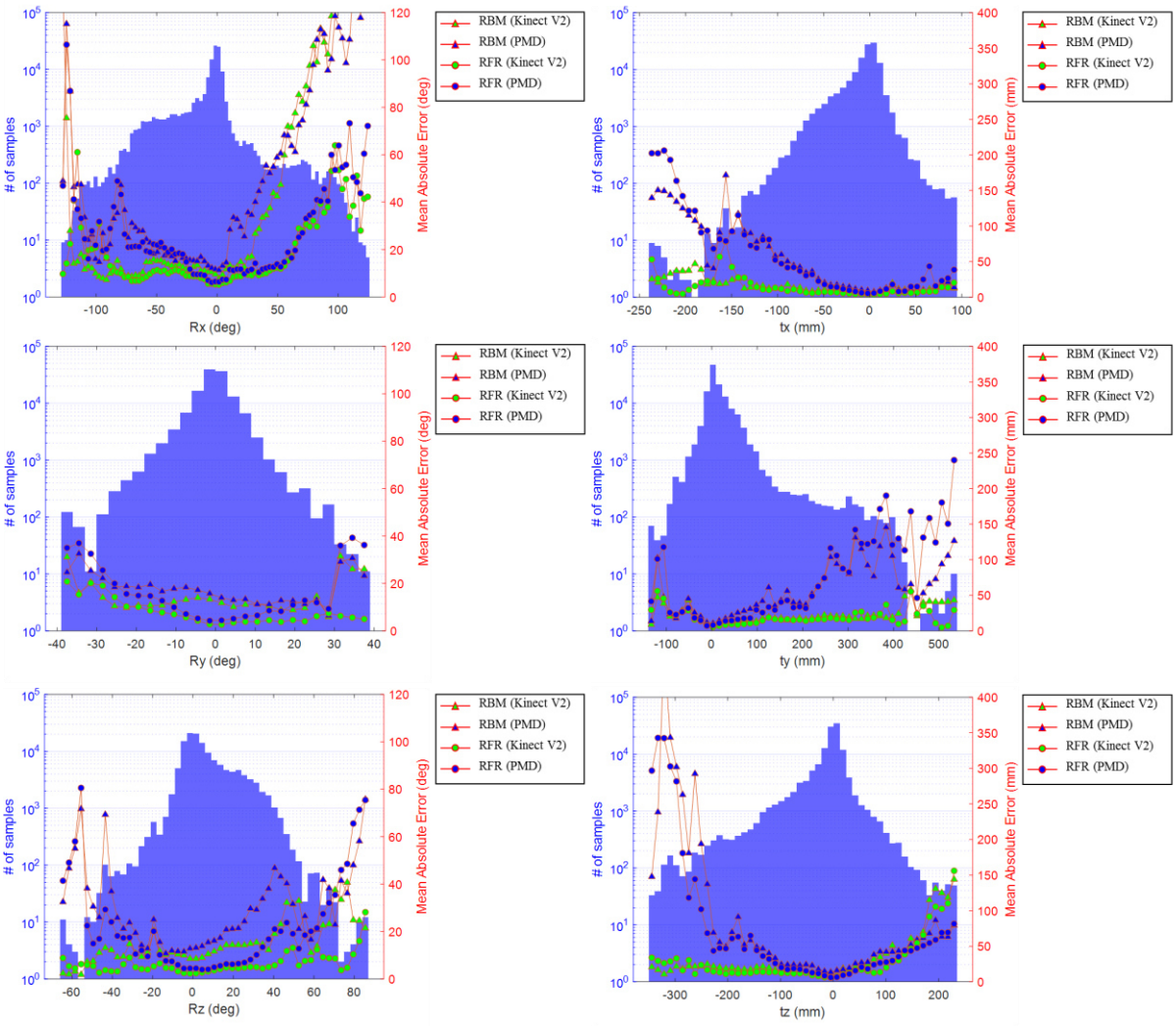


Figure S1. Mean Absolute Error (MAE) of different HPE methods evaluated on the overall dataset of different cameras. The left y-axis represents the number of data samples in different ranges of the ground truth.