Hes·so

RAPPORT

ADVITAM – ADAPTIVE DRIVER-VEHICLE INTERACTION TO MAKE FUTURE DRIVING SAFER

RAPPORT FINAL – APPEL À PROJETS OPEN DATA HES-SO 2021

Projet financé par Hasler Stiftung.

Auteur·es

Quentin Meteier, Elena Mugellini, Stefano Carrino

Hes·so

Table des matières

I/ RÉSUMÉ DE LA RECHERCHE

Le projet AdVitam (pour Adaptive Driver-Vehicle Interaction to Make future driving safer) a exploré comment l'interface homme-véhicule peut s'adapter en temps réel dans le cadre de la conduite conditionnellement automatisée (ou semi-autonome), afin de maintenir le conducteur conscient de son environnement et ainsi le soutenir au mieux lors des demandes de reprise de contrôle. Le but était de développer un nouveau modèle d'interactions collaboratives entre le conducteur et le véhicule, prenant en compte l'état physiologique du conducteur et la situation de conduite à tout instant. Pour cela, 3 doctorants avaient pour rôle de développer chacun un module du modèle AdVitam. Dans le cadre de ce projet, une collecte de données a été réalisée durant 6 expériences sur 2 simulateurs de conduite automatisée à base fixe. Ces tests ont d'abord permis d'entrainer les modèles d'intelligence artificielle permettant de prédire l'état psychophysiologique du conducteur, mais aussi la potentielle qualité de reprise de contrôle si un incident venait à se présenter. Une expérience finale a permis d'investiguer la validité du modèle en conditions réelles (prédictions continues) sur un échantillon de 35 personnes.

II/ DESCRIPTION DES DONNÉES

Décrire les données (types de données, format, quantité, degré de sensibilité, etc.)

Cet ensemble de données contient les données de 346 conducteurs recueillies au cours de six expériences menées dans deux simulateurs de conduite automatisée à base fixe. Cinq études ont simulé la conduite automatisée conditionnelle (L3-SAE), tandis que la dernière a simulé la conduite manuelle (L0-SAE). L'ensemble de données comprend des données physiologiques (électrocardiogramme (ECG), activité électrodermale (EDA) et respiration (RESP)), des données relatives à la conduite et au comportement (temps de réaction, angle du volant, ...), des données relatives à la performance de tâches non liées à la conduite et des réponses à des questionnaires. Parmi celles-ci, des mesures issues de questionnaires standardisés ont été collectées, soit pour contrôler la manipulation expérimentale de l'état du conducteur, soit pour mesurer des états psychophysiologiques liés aux facteurs humains et à la sécurité de la conduite (somnolence, charge mentale, état affectif, conscience de la situation, confiance situationnelle, expérience de l'utilisateur).

Dans l'ensemble de données fourni, certaines données brutes ont été traitées, notamment les données physiologiques à partir desquelles des indicateurs physiologiques (ou caractéristiques) ont été calculés. Ces derniers peuvent être utilisés comme données d'entrée pour des modèles d'apprentissage automatique afin de prédire divers états psychophysiologiques (manque de sommeil, charge mentale élevée, ...) qui peuvent être critiques pour la sécurité du conducteur. Les mesures subjectives autodéclarées peuvent également être utilisées comme vérité pour appliquer des techniques de régression. En outre, des analyses statistiques peuvent être réalisées à l'aide de l'ensemble de données, notamment pour analyser la conscience de la situation ou la qualité de la prise de contrôle des conducteurs, dans différents états et différents scénarios de conduite.

Le dataset est disponible sur Zenodo à l'adresse suivante : https://doi.org/10.5281/zenodo.7319612. Les données peuvent être téléchargées au format .zip et un fichier .zip a été créé pour chaque érience (6 au total). Un fichier README contient des informations générales sur l'expérience (méthode, mesures, références). Les données physiologiques brutes sont au format .acq (fichier brut venant du logiciel Biopac Student Lab) et au format .txt (une colonne pour chaque signal). Les indicateurs physiologiques (données traitées) et les autres mesures (performance à la tâche, réponses aux questionnaires, données de conduite) sont sauvegardées dans un ou plusieurs fichiers CSV, rangés dans des dossiers séparés. Un fichier de documentation au format .xslx est également disponible pour documenter la signification et le codage de chaque variable (metadonnées).

Le degré de sensibilité des données peut être considéré comme étant faible. Les données ont été collectées dans le cadre d'expériences en laboratoire. Les données sont pseudonymisées, ce qui signifie que chaque participant est identifié par un code mais son nom ne figure pas dans le dataset. Ainsi, les données socio-démographiques et physiologiques présentes dans le dataset ne peuvent pas être associées à une personne en particulier.

III/ ETAT DE L'ART

Présenter un bref état de l'art des bases de données (Data repositories) existantes pour la discipline en question et une argumentation justifiant le choix (visibilité, soutien existant, prix, sécurité, etc.)

Pour la prédiction de l'état physiologique du conducteur, un des principales bases de données hébergeant des ensembles de données physiologiques est PhysioNet (https://physionet.org/). Elle contient une centaine de jeux de données avec différents types de signaux (EEG, ECG, ..) collectés dans différents contextes. En particulier, plusieurs jeux de données contiennent des données recueillies auprès de conducteurs. L'un des ensembles de données collectés est une étude majeure sur la détection du stress chez les conducteurs (https://physionet.org/content/drivedb/1.0.0/). A côté de ça, un autre dataset (https://osf.io/sxuc3/) contenant plusieurs signaux de conducteurs distraits pendant la conduite a également été trouvé sur la base de données Open science Framework (OSF). Des analyses préliminaires ont été réalisées sur ce dernier.
Cependant, aucun d'entre eux ne simulait la conduite conditionnellement automatisée. C'est pour cela qu'il a été choisi de construire notre propre dataset dans ce contexte précis, afin de vérifier si l'état psychophysiologique du conducteur pouvait être prédit avec précision à partir des données physiologiques. Ces expériences ont également permis de collecter des données pour la qualité de reprise de contrôle du conducteur dans différents états, et de tester les interfaces humain-véhicule développées dans le cadre du projet (lumières ambiantes, application mobile, vibrations dans le siège).

Zenodo a été choisi pour mettre le dataset à disposition de la communauté scientifique. Ce choix s'explique par le fait que le dépôt est gratuit à hauteur de 50GB (ce qui était notre cas), qu'un DOI est associé au dataset, que le type de licence qui s'applique aux données est explicitement spécifié, et que cette plateforme fait partie des bases de données par le Fond National Suisse pour le Data Management Plan. Ce dernier est maintenant requis pour la demande de projet (voir références). C'est d'ailleurs pour certaines de ces raisons (DOI, licence, …) que Zenodo fait partie des 4 bases de données qui répondent aux exigences du FNS (permettent la publication de données FAIR, non-commerciales).

IV/ METHODE D'IMPLÉMENTATION DES PRINCIPES FAIR

Présenter la méthode d'implémentation des principes FAIR lors du dépôt, incluant une description détaillée (temps de travail, compétences nécessaires, etc) de toutes les tâches effectuées afin de rendre les données compatibles avec les principes FAIR.

Le but de ce travail était de documenter toutes les donnés collectées, dans le but qu'elles soient exploitables par n'importe quelle personne externe au projet. Par souci de clarté, les données relatives à chaque expérience ont été rangée dans un dossier spécifique, identifié par un code : Exp1, Exp2, Exp3, Exp4, ExpTOR, ExpFinal.

Pour toutes Les expériences, la structure des dossiers suit le même schéma, comme le montre la Figure 1 : deux sous-dossiers (Raw, Preprocessed) et un fichier README. Ce dernier contient un résumé de l'expérience, un résumé des méthodes et du matériel utilisés pour mener l'étude. Il contient également des informations sur la structure des fichiers et leur contenu. Les références scientifiques pertinentes sont également incluses.

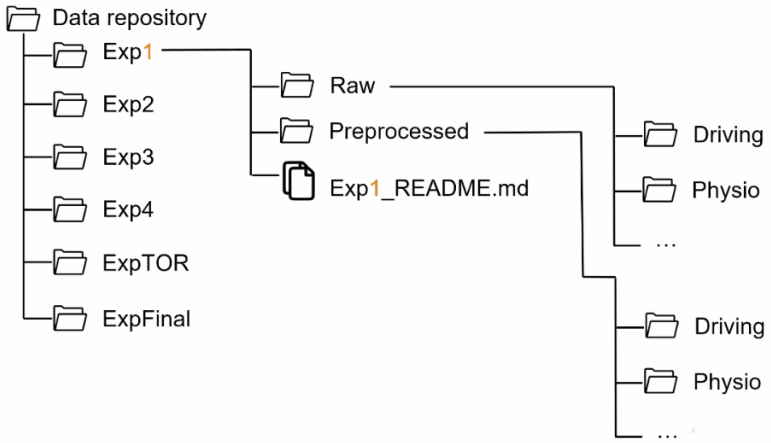

Figure 1: Vue d'ensemble de la structure du dataset.

Les données brutes sont contenues dans le dossier */Raw* incluant les données de conduite de chaque participant, les données physiologiques brutes des conducteurs et éventuellement d'autres données pour certaines expériences (performance de la tâche non liée à la conduite, questionnaires…). Enfin les données traitées sont dans le dossier */Preprocessed*, notamment les indicateurs physiologiques calculés à partir des données brutes, ou des indicateurs de qualité de reprise de contrôle du véhicule. Dans ce dossier on trouve aussi un fichier *ExpX_Database.csv*, contenant les données recueillies via le questionnaire, y compris les informations socio-démographiques des participants et d'éventuelles autres mesures. Pour documenter complètement cette database, un fichier *ExpX_Documentation.xslx* a été créé. Il comprend les termes et abréviations utilisés, les participants à exclure pour une analyse statistique (avec la raison), et les variables attribuées aux données et métadonnées, avec le nom de la variable, le type, la description, la plage et le codage.

Ce travail de rangement et d'organisation des données, ainsi que l'écriture du fichier README a été une étape conséquente du travail pour rendre les données FAIR. Ce travail a notamment été basé sur les recommandations d'un cours de Data Management donné dans le cadre du dispositif CUSO. C'est d'ailleurs ce cours qui a servi de base au cours de data management donné aux étudiant.e.s de 3ème de bachelor en psychologie à l'UniFR, avec qui nous avons collaboré pour la collecte de ces données. Il était demandé aux étudiant.e.s de rendre les données collectées dans leur travail de bachelor FAIR. Ceci a donc aussi servi de base au travail réalisé pour le projet.

De plus, un article présentant les données a été soumis dans le journal Data in Brief (Elsevier), dont le but principal est d'augmenter la visibilité du dataset tout en assurant l'application des principes FAIR, permettant la reproductibilité des expériences et l'exploitation totale des données collectées.

Au total, il a fallu environ 120h pour rendre les données FAIR, en plus du temps passé au cours des 4 ans du projet pour commencer ce travail. L'écriture de l'article a ensuite prit une cinquantaine d'heures. Enfin le dépôt des données sur Zenodo s'est fait sur une demi-journée (4h).

V/ DIFFICULTÉS JURIDIQUES ET ÉTHIQUES

Présenter un état des lieux des difficultés juridiques et éthiques posées lors du dépôt, accompagné d'une description des solutions pour les dépasser (anonymisation, pseudonymisation, consentement, fermeture partielle de l'accès aux données, etc.).

Globalement, nous avons rencontré assez peu de difficultés lors du dépôt en ligne du dataset sur Zenodo. Le consentement de tous les participants a été obtenu et les données ont été pseudonymisées, puisque chaque participant est identifié par un code, mais aucune information personnelle n'est disponible dans le dataset. La seule question que nous nous sommes posés au moment du dépôt était concernant la licence à appliquer aux données. Après consultations de personnes en interne à l'HEIA-FR (Ra&d, bibliothécaire), peu/aucune directive n'a été décidée concernant la licence à appliquer aux données de recherche Open Data. Pour le moment, la référence est semble-t-il l'université de Lausanne qui a travaillé sur le sujet. Ainsi, nous avons choisi par défaut la licence Creative Commons Attribution 4.0 International qui est la licence de base généralement appliquée. Nous avons hésité avec Open Data Commons Open Database License (ODbL), qui semblerait aussi convenir.

BIBLIOGRAPHIE

Bibliographie formée de documents de référence sur l'Open data dans la discipline en question qui ont été consultés.

What is Open Data and Open Licensing? : https://citizens-guide-open-data.github.io/guide/1-open-data

Open Data Commons Open Database License (ODbL) v1.0 : https://opendatacommons.org/licenses/odbl/1-0/

Open Data Commons Open Database License (ODbL) : https://opendatacommons.org/licenses/odbl/

Publisher's Guide to Open Data Licensing : https://theodi.org/article/publishers-guide-to-open-data-licensing/
Attribution 4.0 International (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/
HES-SO Open Data : https://www.hes-so.ch/recherche-innovation/open-science/open-data
Licence d'utilisation Open Research Data – UNIL :
https://www.unil.ch/openscience/home/menuinst/open-research-data/conformite--exigences/licences-dutilisation.html

Research data management for junior researchers – Essential knowledge and fundamental steps. Cours CUSO. Bellier, Blumer, Claivaz. 2018.

Explantions of the FAIR data principles – SNSf.
https://www.snf.ch/media/en/s1eRHVxx1Y41S5le/FAIR_principles_translation_SNSF_logo.pdf

Example of data repositories – SNSF.
https://www.snf.ch/media/en/k64VoMUfwKoUMMY5/FAIR_data_repositories_examples.pdf

ANNEXE

Doit figurer en annexe le rapport financier final extrait d'AGP et signé par le·la chef·fe de projet ainsi que par le·la responsable financier·ère de la haute école.