DESCRIPTIF DU SUJET ET ARGUMENTAIRE DU DIRECTEUR DE THESE

Noms et prénoms des co-directeurs de thèse : Julien Moreau et Franck Davoine.

Intitulé du sujet de thèse : Apprentissage automatique pour l'auto-calibrage de capteurs et l'analyse sémantique de scènes de conduite.

Résumé du sujet de thèse

Les nouveaux véhicules bas carbone, plus légers, avec de nouvelles motorisations, peuvent être un précieux outil au service d'une approche systémique de la mobilité (la technologie, les pratiques, les infrastructures et la réglementation vues comme des outils d'égale importance, et qui doivent être coordonnés) [1]. Ce projet de thèse vise à développer des systèmes de perception de scènes économes et robustes, pour des applications d'assistance à la conduite de véhicules légers. Nous prévoyons pour cela d'utiliser des capteurs en nombre réduit, peu coûteux et à faible consommation d'énergie (caméras RGB, caméras à événements, et/ou Lidars bas coût). Dans ce contexte, le calibrage des capteurs est très important d'une part pour compenser leurs déformations (pouvant être causées par exemple par les vibrations plus importantes sur des véhicules légers) et tirer le maximum de leurs possibilités, d'autre part pour aligner correctement leurs mesures spatiales pour la fusion de données et la perception sémantique de la scène. Le calibrage est contraignant, car il doit être adapté au cas par cas à chaque exemplaire de capteur. L'objectif de la thèse est de prospecter les méthodes d'apprentissage machine de l'état de l'art, intégrant notamment les mécanismes d'auto-attention. Notre intention est d'arriver à un processus multitâche d'analyse sémantique multi-capteurs et de calibrage et recalibrage en ligne afin de toujours bénéficier de la vision la plus fidèle possible du monde.

DESCRIPTIF DU SUJET

1) Le sujet de recherche choisi et son contexte scientifique et économique :

L'analyse de scène pour la robotique mobile repose sur différentes modalités de perception (caméra couleur, caméra à événements, lidar, etc.). Pour une interprétation optimale du monde, ces capteurs doivent être calibrés. De manière individuelle à chaque capteur, la projection du monde réel en données perçues doit être connue, ce sont les paramètres internes tels les distorsions. De manière collective pour la fusion entre capteurs, leurs points de vue doivent être connus relativement les uns aux autres, ce sont leurs poses relatives, paramètres externes du système de perception. Or, chaque capteur est unique car il existe toujours des variations entre exemplaires d'un même capteur. De plus, des éléments peuvent bouger suite aux vibrations du robot, ce qui ajoute des variations supplémentaires. Le calibrage ne devrait donc théoriquement pas n'être effectué qu'une seule fois pour un exemplaire, mais régulièrement pour chaque exemplaire des systèmes de perception. Cette observation est de toute première importance lorsqu'il s'agit d'assurer la sécurité d'un système de robot autonome visant à terme à être dupliqué à grande échelle.

Traditionnellement, le calibrage repose sur l'observation de structures connues, des mires, par les capteurs à calibrer. Une heuristique de détection de la structure attendue est appliquée sur les données capteur, puis une minimisation des différences entre l'observation réelle et l'observation attendue est faite de sorte à estimer les paramètres de calibrage. C'est ce que l'on appelle le calibrage fort, précis mais lourd à mettre en œuvre. Un tel processus ne peut être déployé à grande échelle, à plus forte raison s'il faut en permanence optimiser le calibrage pour pallier aux décalages liés à l'usure des capteurs. Ceci impose de développer un calibrage ne reposant pas sur l'utilisation de mires, mais sur l'observation directe du monde réel : c'est le calibrage faible, ou auto-calibrage [2]. Ces méthodes visent à comparer au moins deux acquisitions successives de la même scène, c'est-à-dire avec un léger décalage entre points de vue, pour arriver à trouver les paramètres de calibrage. La difficulté est d'arriver à la même précision qu'un calibrage fort.

Récemment, des travaux prometteurs ont exploité l'apprentissage profond pour le calibrage automatique de caméras. Le réseau de neurones [3] nécessite seulement une fraction de seconde pour estimer les paramètres de calibrage, aussi bien sur GPU (1ms) que sur CPU (0,33s). Ces résultats ouvrent la porte à des méthodes de calibrage efficaces en direct pendant le fonctionnement du robot, ou recalibrage.

Concernant le calibrage extrinsèque, quelques tentatives de méthodes de recalibrage ont été présentées entre capteurs caméra et lidar [4]. Ces travaux reposent sur l'analyse de la géométrie de la scène par détection d'éléments saillants pouvant être comparés entre les modalités. Nous pensons qu'une approche basée apprentissage profond pourrait arriver à un recalibrage multi-capteurs rapide et précis. Dans le contexte de la robotique mobile, ceci est d'autant plus intéressant que nous pourrions garantir des résultats optimaux également depuis de la perception issue de capteurs bas coût de produits du grand public et arriver à une démocratisation de certains systèmes robotiques.

Récemment, des solutions ont été présentées dans la littérature à base de mécanismes d'auto-attention [5] et de réseaux de neurones récurrents pour le traitement du langage naturel. On citera par exemple les architectures d'analyse sémantique telles que BERT ou GPT-2. Le mécanisme d'auto-attention a également été exploité pour des tâches de segmentation d'images par réseaux de neurones profonds, et encore plus récemment dans des systèmes de détection d'objets visuels. Ces méthodes sont à même de capturer des dépendances contextuelles (relations entre caractéristiques locales et globales) dans les images. D'autres propositions, toujours à base de mécanismes d'auto-attention, visent à se passer des réseaux profonds à convolution pour la classification d'images [6]. Enfin, des travaux récents à base de mécanismes d'attention cherchent à modéliser des données multimodales telles que des images et leurs annotations textuelles [7]. Des méthodes encore moins gourmandes en ressources pour des résultats à l'état de l'art sont actuellement à l'étude et semblent prometteuses pour pouvoir être rapidement appliquées à des systèmes embarqués.
En plus d'arriver à des algorithmes d'apprentissage profond plus légers, les mécanismes d'auto-attention permettent de traiter la scène de manière locale et globale, ils semblent ainsi tout indiqués pour estimer les paramètres de calibrage quelles qu'en soient leurs valeurs initiales données avant optimisation. Ils permettraient par ailleurs d'améliorer les résultats en traitant les données sur un laps de temps en profitant du déplacement du robot (de l'ensemble des capteurs).

En s'inspirant notamment de ces travaux, l'objectif du projet de thèse est de prospecter des solutions innovantes à des problématiques jointes de calibrage et d'analyse sémantique. L'analyse des images peut être par exemple de la détection d'objets ou d'éléments d'intérêt, de la segmentation d'instances, de l'estimation de la profondeur, etc. À la manière de la réflexion sur les avantages des réseaux multitâches menée dans [8], cette analyse pourrait renforcer les contraintes pour estimer le calibrage et vice-versa, l'estimation précise du calibrage pourrait aider à une bonne analyse de la scène. Le calibrage concernera conjointement les paramètres intrinsèques mono capteur et extrinsèques relatifs entre plusieurs capteurs de modalités variées. Une étude devra aussi être menée sur la combinaison de données multimodales issues de capteurs commerciaux de type caméras RGB, caméras à événement, et lidar ; ces derniers fournissant respectivement des informations sur l'apparence des images, sur le mouvement des régions et objets qui les composent, et sur les distances des éléments de la scène au système de perception.

2) L'état du sujet dans le laboratoire d'accueil :

Le laboratoire Heudiasyc est très actif dans les domaines de la localisation pour la navigation du véhicule autonome et de la communication v2v (véhicules à véhicules) et v2x (véhicules à infrastructures). Le positionnement GNSS a besoin d'être complété par des capteurs proprioceptifs et extéroceptifs et par une cartographie numérique pour améliorer les performances de localisation et l'intégrité de l'estimation. La communication pour l'échange et la collecte d'informations nécessite une adaptation des infrastructures, et dans le cas v2v, les véhicules doivent de toute manière être capables d'analyser eux-mêmes les situations. L'analyse de la scène passe par la perception, et elle doit être étudiée en profondeur pour garantir la meilleure fiabilité possible dans les estimations, pour plus de sécurité. Les systèmes de perception que nous développons au laboratoire utilisent principlalement des Lidars et des caméras, et ont besoin d'améliorations pour l'analyse de scène et l'extraction des zones navigables.
Heudiasyc développe des algorithmes de traitement de données Lidar, vidéo et événements pour détecter et classifier des obstacles, piétons ou la chaussée praticable et ses marquages [9, 10]. La perception visuelle était jusqu'à présent obtenue à partir de solutions toutes faites et fermées du marché. Notre objectif avec cette thèse est de développer de nouveaux algorithmes de traitement de la vision multimodale, avec nos capteurs calibrés entre eux. Nous nous focaliserons sur les méthodes basées sur l'apprentissage profond, avec pourquoi pas des projets futurs dédiés à la localisation et cartographie, l'amélioration de la navigation, l'enrichissement de cartes, etc.

Heudiasyc dispose de plateformes et d'infrastructures expérimentales : plusieurs véhicules instrumentés et à énergie électrique (Renault Zoe), ainsi qu'une piste d'essais (Séville, à côté du centre d'innovation de l'UTC). Trois véhicules de type Zoe seront équipés de modules matériels et logiciels pour le traitement de la perception et la conduite autonome : serveur embarqué avec 4 GPU Nvidia de dernière génération et de disques SSD pour l'enregistrement des données en temps réel. Pour l'entraînement de réseaux de neurones profonds, le laboratoire Heudiasyc dispose de plusieurs machines équipées de GPU, de deux stations Nvidia DGX-1 (8 GPUs Tesla V100 avec NVLink), deux calculateurs IBM POWER8 Minsky tous deux équipés en GPUs (2 x 4 GPUs Tesla P100 avec NVLink), et très prochainement de moyens de calcul embarqués Nvidia Drive. Dans le cadre d'essais pour des systèmes embarqués légers, le laboratoire dispose en outre de cartes Nvidia Jetson avec GPU réduit intégré.

3) Les objectifs visés, les résultats escomptés

L'objectif de la thèse est d'arriver à procéder conjointement à une analyse sémantique de la scène et à un recalibrage des capteurs. Ceci implique de gérer la fusion multi-modalités à l'intérieur du réseau. Plusieurs solutions pourront être testées : réseau multi-branches, encodage des *features* issues des différents capteurs pour les uniformiser selon un modèle donné, utilisation d'*embeddings*, ou autre. Ce travail vise à développer un réseau multitâche pour tirer bénéfice des tâches entre elles

en termes d'analyse et compréhension de la scène. L'originalité est d'intégrer le calibrage comme l'une de ces tâches, permettant, d'une part, d'exploiter des capteurs bas coût et d'en optimiser la durabilité grâce au recalibrage, et d'autre part, d'améliorer la précision de l'analyse sémantique et de mesurer et contrôler l'alignement des données multimodales. C'est-à-dire calculer aussi une estimation de la qualité et de la disponibilité du calibrage et de la fusion.

4) Les collaborations prévues *(préciser le cadre, la nature des collaborations, l'ancrage national, international, la transdisciplinarité éventuellement)*

Une collaboration est envisagée avec la société Prophesee, implantée à Paris et leader international dans le domaine des caméras à événements HD. Ceci à plus forte raison que, nous collaborons déjà étroitement avec cette société dans le cadre de la thèse de Vincent Brebion, UTC : « Perception multimodale des vulnérables pour la conduite autonome en environnement urbain ». Les caméras Prophesee sont en voie de production à grande échelle et de miniaturisation suite à un partenariat avec Sony. Ceci implique que ces capteurs seront plus abordables dans un futur proche. Un domaine qui intéresse Prophesee tout particulièrement est le multimédia avec l'intégration de caméras à événements dans des tablettes et smartphones. Les caméras à événements ont de nombreux avantages : très forte dynamique de perception, latence extrêmement faible, basse consommation d'énergie. Elles sont idéales pour être embarquées sur des robots mobiles.

Une collaboration est également possible avec le laboratoire commun SIVALab entre le laboratoire Heudiasyc et le département de recherche du Groupe Renault. Il existe déjà plusieurs thèses communes entre Renault et Heudiasyc (comme la thèse PERVAU mentionnée plus haut). Des *workshops* SIVALab ont lieu chaque mois et regroupent des chercheurs de Renault et du laboratoire Heudiasyc. Nous avons des échanges réguliers et le thème du calibrage de la perception pour le véhicule intéresse au plus haut point les chercheurs de Renault, car un recalibrage *online* faciliterait grandement un déploiement en grande série de systèmes de perception multimodaux complexes.

Ce projet sa place dans le cadre de la nouvelle infrastructure nationale TIRREX (*Technological Infrastructure for Robotics Research of Excellence*) pour la recherche en robotique. L'équipe SyRI du laboratoire Heudiasyc contribuera en effet aux travaux de recherche sur la reprise en main en conduite de véhicule semi-autonome avec interaction multimodale entre le système et l'humain. Enfin, ce projet de thèse trouvera très bien sa place dans les thématiques traitées par le nouveau projet RITMEA (Recherche et Innovation en Transports et Mobilité Écoresponsables et Autonomes) financé par le contrat de plan État-région Hauts-de-France. Enfin, des collaborations pourront être envisagées sur la durée de la thèse avec certains de nos partenaires académiques internationaux (Allemagne, Espagne, Japon, Chine).

5) Publications

[1] Étude comparative de l'impact carbone de l'offre de véhicules. Étude pilotée par Nicolas Raillard pour le think tank The Shift Project, Février 2020 (V1).

[2] Construction de modèles 3D à partir de données vidéo fisheye : application à la localisation en milieu urbain. Julien Moreau, Thèse de doctorat de l'Université de Technologie de Belfort-Montbéliard, 2016.

[3] Deep Single Image Camera Calibration with Radial Distortion, Lopez, Manuel and Mari, Roger and Gargallo, Pau and Kuang, Yubin and Gonzalez-Jimenez, Javier and Haro, Gloria, CVPR 2019.

[4] Online Intelligent Calibration of Cameras and LiDARs for Autonomous Driving Systems, Xu, H. and Lan, G. and Wu, S. and Hao, Q., ITSC 2019.

[5] Attention Is All You Need, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, NIPS 2017.

[6] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, submitted to ICLR 2021.

[7] Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers, Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, Jianlong Fu, arXiv:2004.00849, 2020.

[8] Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, Kendall, Alex and Gal, Yarin and Cipolla, Roberto, CVPR 2018.

[9] Fusion of neural networks for LIDAR-based evidential road mapping, Édouard Capellier, Franck Davoine, Véronique Cherfaoui, You Li. Journal of Field Robotics, Willey, 2021;1–32. arxiv.org/abs/2102.03326.

[10] Transfer Learning in Computer Vision Tasks: Remember Where You Come From, Xuhong Li, Yves Grandvalet, Franck Davoine, Jingchun Cheng, Yin Cui, Hang Zhang, Serge Belongie, Yi-Hsuan Tsai, Ming-Hsuan Yang. Image and Vision Computing, Vol. 93, Elsevier, January 2020.