Dossier Année de Recherche Pré-doctorale à l'Étranger

Sujet et Motivations

Solal Nathan

27 Février 2020

Table des matières

1 Information sur le laboratoire **2**
1.1 Contact du Laboratoire 2
1.2 Directeur de l'institut 2
1.3 Maitre de Stage 2

2 Sujet de l'ARPE **2**

3 Motivation **4**

Références **4**

1 Information sur le laboratoire

Le laboratoire qui m'accueille l'an prochain est le « Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung » ou « Fraunhofer Institute of Optronics, System Technologies and Image Exploitation », souvent abrégé IOSB. Dans la suite de ce document nous le nommerons IOSB. Le Fraunhofer est un institut allemand regroupant 67 laboratoires en science appliqué.

La personne avec qui je suis rentré en contact dans l'IOSB est Thomas Pollok. Il apparait comme mon futur de maitre de stage. J'ai notamment discuté avec lui par visio-conférence et il s'agit de la personne ayant proposé mon sujet de recherche.

1.1 Contact du Laboratoire

- Fraunhofer IOSB
- Fraunhoferstraße 1
- 76131 Karlsruhe
- Téléphone : +49 721 6091-0

Source : Le site de IOSB dans l'onglet Karlsruhe.

1.2 Directeur de l'institut

- Head of institute / managing director
- Prof. Dr.-Ing. habil. Jürgen Beyerer
- Téléphone : +49 721 6091-210
- juergen.beyerer@iosb.fraunhofer.de

1.3 Maitre de Stage

- Thomas Pollok
- thomas.pollok@iosb.fraunhofer.de
- VID Departement, VSA Group
- Research assistant pursuing a PhD

2 Sujet de l'ARPE

Dans les années récentes le domaine de la vision par ordinateur à grandement été modifié par les approches nouvelles basés sur l'apprentissage statistique (*machine learning* en anglais). Ces techniques se basent la plupart du temps sur des réseaux de neurones et notamment leurs variantes convolutives capable d'exploiter pleinement les informations spatiales en deux dimensions (voir plus pour des images qui ne sont pas niveau de gris). Certaines techniques

à la frontière avec l'infographie (imagerie 3D) tentent d'appliquer ces méthodes à des jeux de données en trois dimensions.

Le but de ce sujet de recherche est de créer un nuage de point en 4D (trois dimensions spatiales et une temporelles) à partir de plusieurs sources vidéos non synchronisés. Les capteurs des différentes caméras peuvent être très différents, les flux vidéo peuvent commencer et finir à des instants distincts.

Ce problème est difficile pour plusieurs raisons et peut en pratique être découpé en une multitude de sous problèmes dans le domaine de la vision par ordinateur et de l'apprentissage statistique. En effet, il faut en premier lieu être capable de reconstituer un panorama en trois dimensions à partir d'une ensemble d'images provenant d'une même scène. Il faut ensuite estimer la position des différentes caméras dans la scène reconstituée. Cela n'est pas toujours aisé avec des méthodes conventionnels comme Harris, FAST ou SIFT car les descripteurs ne sont pas parfaitement invariants pour les différentes perspectives si on considère une caméra à 6 dégrées de libertés (**6 DOF camera robust pose estimation**) [1, 2]. De plus l'occlusion de certains des points d'intérêts au cours du temps (personne passant devant une caméra par exemple) rend la tâche encore plus compliqué. Finalement certaines sources sont fixes (caméra de surveillance), mais d'autres peuvent être mobile dans la scène (téléphone portable). Il s'agira alors de correctement estimer la position de la caméra à chaque instant. Ensuite, il sera utile de détecter les objets et de segmenter la scène (**object detection and segmentation**). Cela permettra par exemple de suivre une personne ou un objet en mouvement. La méthode utilisée pour l'instant pour cette étape est un **Mask R-CNN** [3] mais d'autre méthodes tels que **YOLO** [4, 5] pourrons être envisagées. Finalement, il est intéressant d'estimer la position des personnes dans la scène. Une fois les personnes détectées on applique un algorithme d'estimation de posture comme **OpenPose** [6] pour cela.

La solution existante développé par l'IOSB [1] peut encore être améliorée sur un certains nombre de points. On peut par exemple chercher à détecter les plans (**plane detection**) pour mieux reconstituer la scène en trois dimensions et par exemple associer un ensemble de points à des surfaces munies de textures. Les solutions existantes dans la littératures sont le plus souvent *monocular*, mais sachant que l'on possède plus d'information grâce à aux différents angles de vues il doit être possible de développer de meilleurs algorithmes.

De plus, on peut réaliser de la **Depth Estimation** [7, 8] sur chaque image, ou sur un ensemble d'image pour s'aider à construire le nuage de point.

Nous pourrions également s'intéresser à des techniques récentes d'estimation de posture comme **DensePose** [9] ou des techniques d'estimation de position pour des objets que l'on trouve communément dans les scènes étudiés (voitures, lampadaires, meubles...).

Nous pourrions aussi effectuer un **UV mapping** complet des personnes ou objets dans la scène pour en extraire une texture détaillée.

La plupart du travail sera réalisé en se basant sur le travail existant déjà réalisé par l'équipe de recherche dont Thomas Pollok fait partie à l'IOSB. À cela s'ajoute les bibliothèques et outils classiques du traitement d'image et de la vision par ordinateur comme **OpenCV** [10]

1. Pour un rapide présentation de ce travail je vous renvoie à la vidéo de présentation réalisée par Thomas Pollok à l'occasion de la 9th International Conference on Imaging for Crime Detection and Prevention

ou l'**Intel Inference engine** [11]. À ceci s'ajoute les outils de *machine learning* du type **Tensorflow/PyTorch**. Pour effectuer un prototypage facile et rapide on utilisera **Python** pour écrire et manipuler les algorithmes de l'état de l'art. Finalement, le résultat prendra la forme d'une bibliothèque écrite en **C++** pour mieux s'intégrer au travail existant.

3 Motivation

Mon parcours académique suit le domaine dans lequel je souhaiterais effectuer mes recherches plus tard. Je souhaite devenir chercheur dans le domaine de l'apprentissage statistique et plus précisément dans le domaine de la vision par ordinateur. En effet, après avoir expérimenté dans plusieurs de ces domaines, celui-ci s'avère être celui qui m'intéresse le plus, bien que j'apprécie également plusieurs autre domaines connexes comme l'apprentissage par renforcement ou la robotique. J'ai eu la chance d'effectuer des travaux de recherche encadrés par un chercheur dans le domaine (Justin Carpentier, INRIA) grâce à mon TER. De plus j'ai suivi cette année en tant qu'auditeur libre certains des cours du parcours IA; ceux de Nicolas Vayatis apportant les notions de fondement l'apprentissage statistique et ceux de Gabriele Facciolo et Pablo Arias dans le domaine de la vision (Tous trois sont professeurs au MVA). Une grande partie de mon travail et de mon intérêt pour le domaine vient de mes recherches personnelles. Je me tiens au courant de l'actualité et de l'état de l'art dans plusieurs domaines du machine learning comme la vision, le reinforcement learning, le NLP et en AI Safety.

Le projet académique que je souhaite poursuivre est alors :

- Stage sur le crowd couting au SATIE avec Emanuel Aldea
- ARPE sur la reconstruction de scène 4D à l'IOSB avec Thomas Pollok
- M2R dans le domaine du machine learning : MVA
- Thèse dans le domaine du machine learning (assez probablement en vision)

À priori il ne devrait pas y avoir de soucis d'annulation de mon stage ou de ma ARPE pour cause de *COVID-19*, mais si cela est le cas alors je souhaiterais effectuer une année de césure pour travailler sur des projets personnels, continuer de me tenir à jour sur l'état de l'art des domaines qui m'intéresses et assister à des cours ouverts (comme ceux du collège de France) et des conférences (les IEEE, les ACM, AI Summit, CAIML, ICCV, NeurIPS, etc. . .). Ce projet n'est pas très mature contrairement à mon projet de ARPE mais pourrait servir de solution de secours.

Références

[1] Kendall, A., Grimes, M. and Cipolla, R. (2015). Convolutional networks for real-time 6-dof camera relocalization. *CoRR* **abs/1505.07427**.

[2] Kendall, A. and Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. *2016 IEEE International Conference on Robotics and Automation (ICRA).*

[3] He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017). Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV).*

[4] Redmon, J., Divvala, S. K., Girshick, R. B. and Farhadi, A. (2015). You only look once: Unified, real-time object detection. *CoRR* **abs/1506.02640**.

[5] Gonzalez, M., Kacete, A., Murienne, A. and Marchand. (2020). YOLOff: You only learn offsets for robust 6DoF object pose estimation. *ArXiv* **abs/2002.00911**.

[6] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[7] Bhoi, A. (2019). Monocular depth estimation: A survey.

[8] Scharstein, S., D. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision 47, 7–42.*

[9] Guler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S. and Kokkinos, I. (2017). DenseReg: Fully convolutional dense shape regression in-the-wild (densepose). *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[10] Itseez. (2015). Open source computer vision library.

[11] Intel. OpenVINO toolkit.