CEA List
Laboratoire de Vision et Apprentissage pour l'analyse de scène
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Bertrand LUVISON
Tél +33 (0)1 69 08 01 37
E-mail bertrand.luvison@cea.fr

THESE 2022 Réf : LVA-22-T4

Exploitation de données faiblement voire non annotées pour la détection d'interaction

Présentation du laboratoire d'accueil

Basé à Paris-Saclay, le CEA List est l'un des quatre instituts de recherche technologique de CEA Tech, direction de la recherche technologique du CEA. Dédié aux systèmes numériques intelligents, il contribue au développement de la compétitivité des entreprises par le développement et le transfert de technologies.

L'expertise et les compétences développées par les 800 ingénieurs-chercheurs et techniciens du CEA List permettent à l'Institut d'accompagner chaque année plus de 200 entreprises françaises et étrangères sur des projets de recherche appliquée s'appuyant sur 4 programmes et 9 plateformes technologiques. 21 start-ups ont été créées depuis 2003.

Labellisé Institut Carnot depuis 2006, le CEA List est aujourd'hui l'institut Carnot Technologies Numériques

Le Laboratoire de Vision et Apprentissage pour l'analyse de scène (LVA) mène ses recherches dans le domaine de la Vision par Ordinateur (Computer Vision) selon quatre axes principaux :

- La reconnaissance visuelle (détection et/ou segmentation d'objets, de personnes, de patterns; détection d'anomalies; caractérisation)
- L'analyse du comportement (reconnaissance de gestes, d'actions, d'activités, de comportements anormaux ou spécifiques pour des individus, un groupe, une foule)
- L'annotation intelligente (annotation à grande échelle de données visuelles 2D/3D de manière semi-automatique)
- La perception et la décision (processus de décision markovien, navigation)

Description de la thèse

La détection d'interaction homme-objet par vision est un problème visant à déterminer pour toutes les personnes présentes dans une image les relations d'interaction qu'elle a avec son environnement. Cela se traduit par un ensemble de triplet <sujet, verbe, cible> où le sujet est déterminé par une boîte, le verbe par un label et la cible par une autre boîte délimitant l'objet en interaction. Les études sur ce problème sont bien plus récentes, d'une part car la complexité sémantique est supérieure et d'autre part car les bases de données ne sont apparues que récemment (V-COCO [Gupta2015], HICO-DET [Chao2018], etc).

Grâce à ces données, des approches ont montré qu'il était possible de répondre à cette problématique en analysant tout ou partie des couples personne-objet présent dans l'image [Gkioxari2018, Li2019] avec en conséquence une complexité quadratique avec le nombre d'objets présents dans l'image. D'autres approches plus récentes, résolvent ce problème de complexité tout en améliorant les performances globales [Chafik2020], notamment avec des architectures "transformers" [Tamura2021, Chen2021].

CEA List
Laboratoire de Vision et Apprentissage pour l'analyse de scène
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Bertrand LUVISON
Tél +33 (0)1 69 08 01 37
E-mail bertrand.luvison@cea.fr

Exemples d'interactions détectées par l'algorithme du LVA

Bien que les performances ne cessent de s'améliorer, au vu de la complexité du problème, le problème est loin d'être résolu. L'une des causes de cette faiblesse, sont les bases de données elles-mêmes. Elles sont en effet, pauvres par rapport à celles de détection, pauvres en nombre d'image par type d'interaction (nb de verbe) et encore plus pauvres par type d'interaction sur un type d'objet donné (couple verbe+type objet target). C'est un constat assez paradoxal lorsque l'on sait que le niveau sémantique à interpréter est plus élevé.

Ce problème du manque de données annotées est assez classique en vision par ordinateur. Pour le pallier sans passer systématiquement par une annotation manuelle extrêmement chronophage et non scalable, des approches capables d'utiliser d'autres données que celles initialement prévues ont été proposées. C'est le cas en classification, détection, classification d'action, etc. L'exploitation de ces données tierce peut se faire par différemment moyens :

- par des méthodes de pré apprentissages avec des tâches prétextes [Radford2021] dont les propriétés semblent adaptées au problème que l'on cherche à résoudre
- avec des techniques dites de distillation de connaissance [Liu2021]
- avec des supervisions faibles lorsque les données ont des annotations d'une autre nature desquelles il est possible de faire des hypothèses en rapport avec le problème donné [Kirthi2021].

Dans les contextes de l'analyse des interactions, ces méthodes ne sont pas encore très utilisées. C'est sur cette hypothèse que repose cette thèse. L'objectif sera d'étudier les différents procédés permettant d'exploiter d'autres données que les bases d'images d'interaction. La nature de ces nouvelles données ne sera pas nécessairement des images, l'exploitation de vidéos ou de données textuelles pourra être à l'étude.

Références:

[Gupta2015] S. Gupta and J. Malik. Visual semantic role labeling. arXiv preprint arXiv:1505.04474, 2015
[Chao2018] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object Interactions. WACV, 2018.
[Gkioxari2018] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. CVPR, 2018.
[Li2019] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, and C. Lu. Transferable interactiveness knowledge for human-object interaction detection. CVPR, 2019.
[Chafik2020] S. Chafik, A. Orcesi, R. Audigier, and B. Luvison. Classifying all interacting pairs in a single shot. WACV, 2020.
[Tamura2021] Masato Tamura, H. Ohashi, T. Yoshinaga. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. CVPR, 2021
[Liu2021] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, P. Vajda. Unbiased Teacher for Semi-Supervised Object Detection. ICLR, 2021
[Kirthi2021], S. Kirthi Kumaraswam, E. Kijak. Detecting Human-Object Interaction with Mixed Supervision. WACV, 2021

CEA List
Laboratoire de Vision et Apprentissage pour l'analyse de scène
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Bertrand LUVISON
Tél +33 (0)1 69 08 01 37
E-mail bertrand.luvison@cea.fr

[Sugimoto2021] M. Sugimoto, R. Furuta, Y. Taniguchi. Weakly-supervised Human-object Interaction Detection. VISAPP, 2021
[Chen2021] J. Chen, K. Yanai. QAHOI: Query-Based Anchors for Human-Object Interaction Detection. arXiv preprint arXiv:2112.08647, 2021
[Radford2021] A. Radford et al. Learning Transferable Visual Models from Natural Language Supervision. 2021

Niveau demandé :	Ingénieur, Master 2
Durée :	3 ans
Rémunération :	entre 1800 € et 2000 €.

Compétences requises :

- Vision par ordinateur
- Apprentissage automatique (deep learning)
- Reconnaissance de formes
- C/C++, Python
- La maîtrise d'un framework d'apprentissage profond (en particulier Tensorflow ou PyTorch) est un plus**.**