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 2021 Réf : LVA-21-BL

Apprentissage d'un détecteur à classe paramétrable via modèle 3D

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

Les détecteurs d'objet ont connu une amélioration très significative grâce à la démocratisation des réseaux de neurones. Il est maintenant simple de détecter un type d'objet du moment que l'on dispose d'un ensemble d'images représentant cet objet dans diverses conditions. Ces ensembles d'images existent pour quelques classes d'objet (21 classes dans PascalVOC, 80 dans COCO, etc) mais en pratique, il est souvent nécessaire des créer des « datasets » spécifiques et de profiter des techniques de « fine-tuning » ou de « transfer learning » pour spécifier un réseau déjà appris. La constitution de ce « dataset » est d'une part excessivement fastidieuse et d'autre part la flexibilité de ce genre de modèle à l'ajout d'une classe est quasi nulle car il est nécessaire de repasser par une phase d'apprentissage complète sur le « dataset » augmenté.

Le challenge de cette thèse sera de s'affranchir de cette contrainte d'ensemble fixe de classe détectable. Formuler ainsi, le problème pourrait s'apparenter à un problème de « Zero-Shot Learning » [5], mais ce genre de modélisation cherche en général à s'appuyer sur une connaissance sémantique auxiliaire, comme par exemple une définition par attribut pour contrôler la description d'une image même lorsque l'image concerne un objet jamais vu a l'apprentissage. Ce genre d'approche repose donc sur la constitution d'une connaissance plus descriptive afin de pouvoir faire le lien avec les définitions sémantiques tel que nous les concevons dans les dictionnaires.

L'objet de cette thèse n'est pas d'aborder le problème sous cet angle, mais plutôt d'être en mesure de fournir en paramètre les types d'objet que l'on souhaite reconnaître par le biais de modèle 3D. Il ne s'agit pas de construire une connaissance sémantique représentative de l'ontologie qui nous entoure mais plutôt d'apprendre à un réseau comment retrouver dans une image, n'importe quelle forme 3D préalablement défini, à l'instar des anciens algorithmes de « template matching » [1] mais cette fois-ci avec tout l'apport des réseaux de neurones.

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

Positionnement :

La formulation du problème de détection comme un problème d'association de modèles 3D issu d'un ensemble d'objet paramétrable avec le contenu de l'image n'est à notre connaissance pas énormément étudié dans la littérature. Cependant des problèmes s'en approche, tel que :

- L'estimation de pose d'objet quelconque [6] couplée au détecteur d'objet agnostique qui vise à simplement définir la notion d'objet sans en déterminer la classe [3]
- La génération de description 3D dense des objets présents dans l'image [4] ou l'estimation de carte de profondeur par des approches de « deep learning » en vue d'une association 3D avec un modèle CAD quelconque.

Axe de recherche :

L'une des idées de cette thèse est d'exploiter les récentes architectures de type transformers pour répondre au problème d'association modèle 3D / position dans l'image 2D. Ces nouvelles architectures ont montré d'excellentes performances pour de nombreux problèmes tel que la traduction [2], la détection [8], le tracking [9], etc. C'est la formalisation du problème d'association à l'intérieur même de l'architecture de réseau de neurones et non à l'extérieur qui fait la différence.

Un autre axe de recherche important sera celui de la bonne modélisation de l'information 3D. En effet, les représentations 3D adaptées au réseaux de neurones sont variées, voxelisation, nuage de point, carte(s) de profondeur, etc [7]. Ces différentes représentations ont toutes des avantages et des inconvénients et leur utilisation adéquate pour l'objectif de cette thèse sera un axe de recherche a part entière. Les bases de données alignant des modèles 3D aux images ont été mis à disposition par la communauté comme ObjectNet3D, Pascal3D+, IKEA3D, Kitti, etc.

Enfin, la résolution de ce problème est intimement liée à celui de l'estimation de la pose de l'objet, voire de la segmentation des zones 3D de l'objet. La détermination de ces étapes sera à l'étude durant la thèse et pourra faire office d'extension. Comme autre voie d'extension possible, une fois la preuve de concept réalisée sur les objets rigides, sera de l'étendre aux objets articulés

Objectif :

L'objectif de cette thèse sera de proposer un nouveau paradigme de reconnaissance d'objet (détection ou segmentation d'instance) où la liste des objets à détecter sera défini comme paramètre et ne nécessitera pas de nouvel apprentissage en cas de changement. Une première approche sera d'entraîner un réseau de neurones à modéliser n'importe quelle forme 3D fourni en entrée, à comprendre comment elle peut être projeter dans l'image et à chercher cette reprojection dans des images « in the wild ».

La démonstration de la détection et l'estimation de la pose sur des objets jamais observés dans l'ensemble d'apprentissage sera un élément clé pour valider l'approche. De plus, montrer qu'il est possible de modifier la liste des objets à reconnaître à la demande sans avoir à repasser par une phase d'apprentissage sera une avancée significative pour les tâches de reconnaissance par vision.

Des utilisations directes dans le domaine industriel avec la détection de pièces quelconques définies uniquement par un modèle 3D ou bien l'adaptation rapide de la détection à tel ou tel type d'objet (armes, objets médicales, véhicules urbain, véhicule industriel, etc) selon le contexte applicatif, seront ainsi possibles.

Références:

[1] Template Matching Techniques in Computer Vision: Theory and Practice, Wiley, R. Brunelli, ISBN 978-

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

0-470-51706-2, 2009
[2] Attention is all you need, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L.Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, NIPS, 2017
[3] Class-agnostic Object Detection, A.Jaiswal,Y.Wu, P.Natarajan, P.Natarajan, arxiv, 2020.
[4] 3D Object Detection and Pose Estimation of Unseen Objects in Color Images with Local Surface Embeddings, G.Pitteri, A.Bugeau, S.Ilic, V.Lepetit, arxiv 2020.
[5] Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly, Y. Xian, C.H. Lampert, B. Schiele, Z. Akata, ArXiv, 2020
[6] PoseContrast: Class-Agnostic Object Viewpoint Estimation in the Wild with Pose-Aware Contrastive Learning, Y.Xiao, Y.Du, R.Marlet, arxiv, 2021.
[7] From Points to Multi-Object 3D Reconstruction, F. Engelmann, K. Rematas, B. Leibe, V. Ferrari, CVPR, 2021
[8] End-to-End Object Detection with Transformers, N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, ECCV, 2020
[9] TrackFormer : Multi-Object Tracking with Transformers, T. Meinhardt, A. Kirillov, L. Leal-Taixe, C. Feichtenhofer, ArXiv, 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.