DetReco - Détection et reconnaissance d'objets dans les images : Rapport d'avancement

François Brémond francois.bremond@inria.fr
Inria, Université Côte d'Azur

Jonas Renault jonas.renault@inria.fr
Inria, Mission Défense & Sécurité

1. Introduction

L'objet de la convention DetReco est de mener une étude de l'état de l'art des algorithmes pouvant s'appliquer à la détection et à la reconnaissance, proche du temps réel, de véhicules militaires sur des images et vidéos, de proposer une implémentation de cet état de l'art, et d'en faire une évaluation.

Une première réunion de travail a eu lieu entre Inria et DGA Techniques Terrestres (DGA TT) à Bourges afin d'identifier les objectifs de la convention et les premières tâches à réaliser par les deux parties. Notamment, l'accent a été porté sur la nécessité de produire un jeu de données d'images d'entraînement pour des modèles de détection, la qualité des modèles de détection étant fortement dépendante du jeu de données sur lequel ils sont entraînés.

Dans ce rapport, nous présentons les travaux préliminaires réalisés par Inria suite à cette réunion de travail dans le cadre de la convention DetReco. Dans une première section, nous rappelons les éléments de la convention tels qu'ils ont été discutés lors de la réunion entre Inria et DGA TT. Puis nous présentons le travail réalisé pour la constitution d'un jeu de données d'entraînement, les premiers résultats obtenus avec l'entraînement d'un algorithme de détection automatique, et les pistes d'amélioration explorées pour améliorer la qualité des résultats.

2. Réunion DGA TT - Inria

La réunion de travail DGA TT - Inria s'est tenue le 5 avril 2023 à Bourges, en présence de Frédérique Segond (Inria), Jonas Renault (Inria), François Brémond (Inria), Caroline Moritz (DGA TT), Hanond Nong (DGA TT) & Philippe Chevalier (DGA TT). Trois tâches ont été discutées en particulier :

1. constitution d'une base de données pour l'entraînement de modèles de détection automatique de véhicules militaires.
2. définition d'une plateforme d'évaluation des algorithmes de détection
3. fourniture d'algorithmes de détection pour valider la plateforme d'évaluation

Nous rappelons également les éléments sur lesquels porte la convention : détection, reconnaissance & identification.

2.1 Détection

La détection est la capacité à distinguer un objet du fond. Le critère d'évaluation est la distance maximale à laquelle se perçoit l'objet. La détection à longue distance ne permet pas la reconnaissance de la cible, seulement d'en détecter la présence.

Pour entraîner un algorithme de détection, il faut fournir des images présentant l'objet cible à la distance maximale visée. On définit la taille des cibles en pixels minimum nécessaires pour la détection. Ainsi, pour la détection, il faudra au minimum **4x8 pixels**.

2.2 Reconnaissance

La reconnaissance est la capacité à classifier un objet dans une catégorie. Le critère d'évaluation est la distance maximale à laquelle l'objet peut être reconnu correctement. La taille minimale nécessaire pour la reconnaissance est d'au moins **20x40 pixels**.

La reconnaissance nécessite la définition d'une classification des véhicules cibles. Plusieurs pistes ont été proposées par la DGA TT : classification des véhicules par catégories (véhicules blindés, moyens anti-char, artillerie sol-sol, etc.) ou par fonction (matériels de l'avant, matériels d'appui ou de soutien, etc.). La difficulté pour l'établissement d'une classification vient de l'ambiguïté des classes assignables à certains véhicules. Notamment, un véhicule d'une même catégorie peut avoir différentes fonctions, raison pour laquelle une classification par catégorie semble être à privilégier.

2.3 Identification

L'identification est la capacité à identifier précisément un objet cible. L'identification par un expert s'appuie sur des clés d'identification spécifiques pour chaque matériel. Dans le cadre de la convention DetReco, l'identification est laissée hors périmètre.

3. Constitution d'une base de données d'entraînement

La première tâche identifiée consiste à créer une base de données d'images permettant l'entraînement d'un algorithme de détection et reconnaissance de véhicules militaires à l'état de l'art.

Pour la DGA TT, cela consiste à identifier les sources de données existantes pouvant être transférées à Inria: ROC-V, SAFARI, MATTER, CRT, etc. (y compris, le besoin d'annoter ces données pour l'entraînement d'un algorithme de détection, ainsi que de déclassifier ces données pour qu'elles puissent être manipulées par Inria).

Pour Inria, il s'agit d'identifier un algorithme cible de l'état de l'art pouvant servir de référence pour l'évaluation de la qualité du jeu de données, et d'identifier les caractéristiques du jeu de données pertinentes pour la qualité du modèle entraîné. En effet, les modèles de deep-learning visés par la convention DetReco sont fortement sensibles aux jeux de données utilisés pour les entraîner. Les caractéristiques qui peuvent jouer sur la qualité du modèle entraîné sont :

- dimensions et résolutions des images
- luminosité

- dimensions et nombre des objets cibles à détecter dans l'image (gros plan, arrière plan, etc.)
- visibilité de l'objet cible (occulté, ou bien peu visible en raison des conditions de la mise en scène telles que brouillard, pluie, neige, fumée, obstacles, etc.)
- diversité et représentativité des classes

Un algorithme qui n'aurait été entraîné que sur un sous-ensemble des données possibles serait inapte à détecter les véhicules dans des conditions qu'il ne connaît pas. Ainsi, pour pouvoir détecter des objets dans des conditions de terrain réelles, il est nécessaire de constituer un jeu de données au plus proche possible de ces conditions.

3.1 Entraînement d'un algorithme de l'état de l'art pour la détection et la reconnaissance

Les algorithmes de l'état de l'art proposés dans le cadre de la convention sont :

- Pour la détection: YoloVx (Jocher et al., 2023), FasterRCNN (Ren et al., 2016), GroundingDINO (Liu et al., 2023)
- Pour le suivi d'objets (tracking): DeepSort (Wojke et al., 2017), ByteTrack (Zhang et al., 2022)

Pour la première phase d'étude du projet, nous avons implémenté un algorithme qui s'appuie sur le modèle YoloV8 pour la détection et l'algorithme DeepOCsort (Maggiolino et al., 2023) pour le tracking de véhicules.

3.2 Jeu de données d'entraînement

En complément des données fournies par DGA TT, Inria a identifié des sources de données pouvant servir à l'entraînement d'algorithmes de détection. Notamment, des jeux de données standards utilisés pour l'entraînement d'algorithmes de visualisation, ou des images disponibles librement sur Internet.

Nous décrivons dans cette section les jeux de données identifiés et utilisés pour l'entraînement de l'algorithme YoloV8 de détection.

3.2.1 Open Images v7

Open Images est un jeu de données d'environ 9 million d'images annotées, avec notamment 16 million d'annotations *bounding boxes* pour 600 classes. Parmi ces classes, une classe *Tank* contient 1248 images annotées de véhicules militaires.

3.2.2 ImageNet

Le jeu de données ImageNet, créé pour le challenge de détection ILSVRC2012, a été au centre des avancées récentes dans le domaine de la reconnaissance automatique d'objets dans des images par deep-learning. Ce jeu de données sert aujourd'hui de jeu de données

Figure 1: Image annotée d'un char, extraite du jeu de données Open Images v7.

standard pour l'entraînement d'algorithmes de transfert learning. Il contient plus de 14 million d'images annotées, divisées en 21841 classes.

Parmi ces 21841 classes, plusieurs peuvent contenir des images pertinentes pour la reconnaissance de véhicules militaires. Par exemple, les classes

- n02739889 (armored car, armoured car)
- n02740061 (armored car, armoured car)
- n02740300 (armored personnel carrier, armoured personnel carrier, APC)
- n02740533 (armored vehicle, armoured vehicle)
- n04389033 (tank, army tank, armored combat vehicle, armoured combat vehicle)

Néanmoins, parmi celles-ci seules les images de la classe n04389033 (tank, army tank, armored combat vehicle, armoured combat vehicle) contiennent des annotations pour la détection d'objets. Cette classe contient 378 images annotées.

3.2.3 Données disponibles librement

Une recherche d'images sur Internet permet de trouver aisément des exemples pouvant servir à l'entraînement d'un algorithme de détection de véhicules. Plusieurs facteurs sont cependant à prendre en compte. Premièrement, les images que l'on peut trouver ne sont

pas annotées, et demandent donc un travail supplémentaire d'annotation manuelle, qui est souvent très coûteux (à minima en temps).

Figure 2: Annotation manuelle d'une image d'engin blindé du génie.

Deuxièmement, ces images servent principalement à répondre aux attentes des utilisateurs d'un moteur de recherche, et ne correspondent pas nécessairement aux critères attendus pour l'entraînement d'un algorithme de détection d'objets. Ainsi, une recherche d'images de char Leclerc donnera des résultats qui présentent l'objet en gros plan, clairement identifiable. Il est difficile d'obtenir des images qui présentent l'objet « en situation réelle de combat », ce qui limitera la qualité du modèle entraîné.

Nous avons tout de même pu constituer, à partir d'une recherche limitée pour certains modèles de véhicules militaires, un premier corpus composé de 669 images. Nous avons annoté manuellement ces 669 images avec quatre classes.

3.3 Définition d'une classification de véhicules militaires

Comme discuté lors de la réunion Inria - DGA TT, la définition d'une classification pose problème en raison de l'ambiguïté des véhicules pouvant se retrouver dans plusieurs classes. Une première classification a été proposée pour regrouper les véhicules dans des catégories larges :

- Armoured Fighting Vehicle (**AFV**) : les véhicules blindés de combat. Ex. Figure 3
- Armoured Personnel Carrier (**APC**) : les véhicules blindés de transport de troupes. Ex. Figure 4
- Military Engineering Vehicle (**MEV**) : les engins blindés du génie. Ex. Figure 5
- Light armoured vehicle (**LAV**) : les véhicules blindés légers. Ex. Figure 6

Cette classification pourra être amendée en fonction des résultats obtenus par l'algorithme de détection et de l'expertise de la DGA TT. D'autre part, la classification pourra être étendue sur plusieurs niveaux, afin de définir des classes génériques qui pourront ensuite être précisées en sous-classes en fonction de la granularité souhaitée (i.e. *véhicule militaire* > *véhicule blindé* > *véhicule blindé de combat*).

Figure 3: **AFV** - char M1 Abrams

Figure 4: **APC** - Panhard VCR

Figure 5: **MEV** - Kodiak Wisent

Figure 6: **LAV** - Panhard VBL

4. Résultats préliminaires

L'entraînement du modèle YoloV8 sur le jeu de données constitué ne pose pas de problèmes. Le transfert learning permet de spécifier une tâche générique (détection d'objet dans des images) pour un besoin spécifique (reconnaissance de véhicules militaires) avec un jeu de données de taille limitée.

Néanmoins, nous constatons les inconvénients liés à la qualité du jeu de données et qui étaient prévisibles : le modèle est entraîné sur des images qui présentent majoritairement des véhicules de face ou de profil, en gros plan, et sans occultation. Pour ces raisons, le modèle n'offre pas encore les qualités requises pour une utilisation en conditions réelles, i.e. pour la détection de véhicules à distance, par faible visibilité.

Les Figures 7, 8 & 9 montrent que le modèle reconnaît aisément un véhicule blindé de transport de troupes lorsque celui-ci se présente de profil, mais il ne reconnaît plus ce même véhicule lorsque celui-ci se présente de face ou bien lorsqu'il est partiellement occulté par des projections.

Figure 7: Lorsque le véhicule est de face au loin, il n'est pas reconnu.

Figure 8: Lorsqu'il est de profil, il est bien reconnu.

Figure 9: Les projections d'eau empêchent la reconnaissance.

5. Amélioration du jeu de données d'entraînement

Plusieurs pistes d'amélioration du jeu de données d'entraînement sont possibles avant de travailler spécifiquement sur les paramètres du modèle. Nous détaillons dans cette section les pistes étudiées.

5.1 Guide d'annotation pour la constitution du jeu de données

Une tâche identifiée lors de la réunion Inria - DGA TT consiste à produire un guide d'annotation qui permettra à la DGA TT de réunir des images qui répondent aux besoins pour l'entraînement efficace d'un algorithme de détection. Ce travail doit être réalisé conjointement entre Inria et la DGA TT pour prendre en compte à la fois les possibilités offertes par la DGA TT et les besoins en images diverses d'Inria. Il permettra d'augmenter considérablement la taille du jeu de données, ainsi que sa qualité, en fournissant des images qui répondent aux critères d'exigence pour un entraînement efficace. Une première version du guide d'annotation est en cours de rédaction.

5.2 Utilisation d'images de synthèse

Une possibilité offerte par DGA TT est la génération d'images de synthèses à l'aide de son moteur de simulation et des modèles des nombreux véhicules militaires dont ils disposent. Ainsi, il est possible de générer des images de synthèse réalistes qui mettent en scène les

modèles, selon des prises de vue variées. Cette possibilité permettrai de répondre au besoin d'images de véhicules dans des mises en scène particulières.

Figure 10: Image de synthèse d'un véhicule militaire en opération.

5.3 Génération de données d'entraînement par modèles de diffusion

Les modèles de diffusion *text-to-images* permettent de générer une image à partir d'un *prompt* textuel. Certaines méthodes récentes permettent de personnaliser des modèles de diffusion à partir de seulement quelques images d'un sujet particulier. Ainsi, la méthode *Dreambooth* (Ruiz et al., 2023) permet au modèle de diffusion de générer des images d'un sujet particulier dans un contexte variant les poses, la mise en scène, etc. à partir d'une description textuelle. Cette méthode est actuellement étudiée pour augmenter les jeux de données d'entraînement dont la taille est limitée en générant de nouvelles images dont la mise en scène variée apportera de la diversité au jeu de données.

Nous avons entrepris d'utiliser cette méthode *Dreambooth* afin d'entraîner un modèle de diffusion à générer des images pour un modèle particulier de véhicule militaire. Nous avons entraîné le modèle à partir de quinze images d'un char Leclerc. Une fois le modèle entraîné, il est possible de générer de nouvelles images d'un char Leclerc dans différentes mises en scène. La Figure 11 illustre ces résultats préliminaires.

Nous travaillons à améliorer la configuration du modèle de diffusion afin de corriger certains artefacts présents sur les images générées afin de produire des images de synthèse utilisables pour l'entraînement du modèle de détection.

6. Poursuite des travaux

Les premiers résultats obtenus par l'équipe sont encourageants. Ils montrent notamment que les modèles de détection de l'état de l'art sont aptes à répondre aux besoins de détection de véhicules militaires, tout en mettant en avant la sensibilité de ces modèles aux données d'entraînement. Il apparaît que la qualité du modèle de détection est directement liée à la qualité ainsi qu'à la diversité du jeu de données utilisé pour son entraînement.

Les premières pistes de travail explorées offrent des solutions prometteuses pour améliorer la base de données d'entraînement, en particulier lorsque celle-ci contiendra des images fournies par la DGA TT. Dans cette optique, le travail d'écriture d'un guide d'annotation et de création d'une base de données d'entraînement constitue une étape importante pour

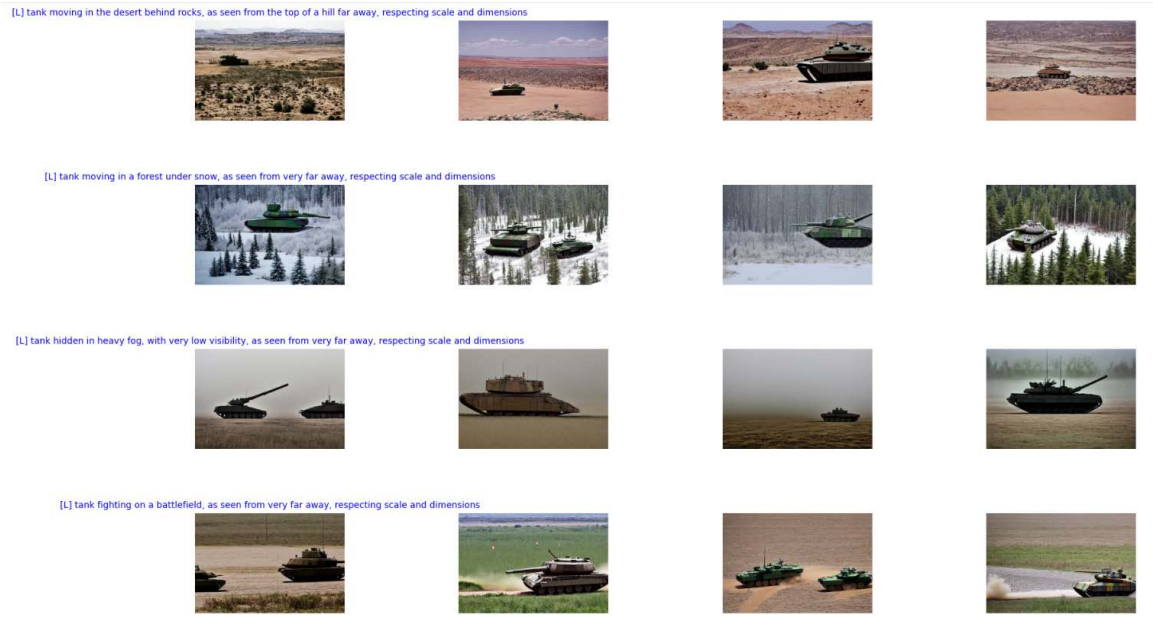

Figure 11: Génération d'images d'un char Leclerc à partir de prompts textuels.

l'avancée du projet. Une fois cette étape réalisée, le travail de production d'une méthodologie d'évaluation pourra également commencer. Le recrutement par Inria d'un post-doc dans le domaine de la détection automatique d'objets en vision par ordinateur permettra d'avancer sereinement vers ces jalons. Une offre de poste a été publiée avant la période estivale et nous sommes toujours à la recherche d'un candidat adéquat.

Références

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. URL `https://github.com/ultralytics/ultralytics`.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification, 2023.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022.