INF6804

Vision par ordinateur

TP 3

Détection et suivi d’un objet d’intérêt

Félix Auger (2067772)

Arnaud Dalbec-Constant (2014288)

15 avril 2024

Introduction

Le suivi d'objets dans le domaine de l'analyse vidéo est une tâche critique qui permet non seulement d'identifier l'emplacement et la classe des objets dans la séquence, mais aussi de maintenir un identifiant unique pour chaque objet détecté au fur et à mesure que la vidéo progresse. Les applications sont multiples, allant de la surveillance et la sécurité à l'analyse sportive en temps réel.

Le but de ce TP est de proposer une méthode capable d'effectuer la détection et le suivi de multiples objets d'intérêt, en l'occurrence des tasses, à travers une séquence vidéo, en conservant l'identifiant unique de chaque objet détecté.

Description de la solution

L'équipe a choisi d'opter pour une solution déjà existante et de l'adapter aux besoins du TP. La solution choisie est basée sur YOLOv8, un nouveau modèle de vision par ordinateur développé par Ultralytics. [1] Le modèle YOLOv8 contient une prise en charge intégrée de la détection d'objets, de la classification et des tâches de segmentation. Le modèle est donc parfaitement adapté pour effectuer du suivi multi-objets avec l'ajout d'un algorithme de suivi comme ByteTrack ou BoT-SORT (également offerts par Ultralytics). Le code de YOLOv8 est disponible en open source et il est possible d'installer le package directement dans Python. [2]

Le modèle YOLO d'Ultralytics se distingue sur de nombreux plans. D'abord, YOLO est hautement efficace pour effectuer du suivi d'objets en permettant de traiter les flux vidéo en temps réel sans compromettre la précision. Dans YOLOv8, il y a différentes tailles de modèles telles que yolov8n (nano), yolov8s (petit), yolov8m (moyen), yolov8l (grand) et yolov8x (très grand). Un modèle plus grand a l'avantage d'être plus robuste et précis, mais en étant plus lourd et moins rapide. YOLO est également très flexible en prenant en charge plusieurs algorithmes de suivi et configurations. Son API Python simple facilite l'utilisation et permet une intégration et un déploiement rapides. Il est également personnalisable et est facile à utiliser avec des modèles YOLO entraînés sur mesure, permettant une intégration dans des applications spécifiques à un domaine. [1]

Le modèle présente toutefois quelques inconvénients. Tout d'abord, il peut avoir du mal à détecter précisément les objets de petite taille en raison de ses caractéristiques de résolution plus faible, rendant ainsi leur localisation plus difficile et pouvant entraîner une précision de détection inférieure par rapport à d'autres méthodes spécialisées dans cette tâche. De plus, étant donné que YOLO prédit des boîtes englobantes sur une grille, il peut rencontrer des défis pour localiser avec précision des objets présentant des formes complexes ou ceux qui sont proches les uns des autres, ce qui peut conduire à des prédictions de boîtes englobantes légèrement imprécises. [3]

Identification des difficultés de la séquence

La séquence vidéo fournie présente plusieurs difficultés, tant pour la détection que pour le suivi des objets d'intérêt, soit les tasses.

Détection :

- Le contraste fluctuant représente un défi majeur. De nombreuses tasses sont de couleur blanche, tout comme l'arrière-plan dans certains cas, rendant les contours difficiles à discerner. De plus, la proximité de plusieurs tasses blanches complique la détection individuelle et la distinction entre elles.
- Les tasses sont fréquemment partiellement occultées. Par exemple, une tasse peut être dissimulée derrière une autre à plus de 50 %, ou partiellement cachée par la main de la personne qui les manipule. Des fois, une partie de la tasse peut même sortir du champ de vision de la caméra, ne laissant apparaître qu'une fraction de sa forme. Ces occultations partielles ou totales nuisent à la détection.

- Les variations de netteté, notamment lorsque les tasses sont hors de focus pendant leur déplacement, compliquent la détection. Par exemple, lorsque le focus est sur une tasse proche de la caméra, les autres tasses en arrière-plan apparaissent floues, rendant leur détection plus difficile.
- La présence de verres dans la vidéo ajoute à la complexité. Leur forme similaire à celle des tasses peut entraîner des fausses détections, tandis que le phénomène de distorsion des formes peut altérer l'apparence des tasses en arrière-plan, les rendant difficiles à détecter.
- La présence de petites tasses blanches placées l'une sur l'autre vers la fin de la séquence constitue un défi supplémentaire. Leur petite taille, leur couleur similaire à celle de l'arrière-plan et leur proximité rendent leur détection délicate.
- L'orientation des tasses peut aussi grandement influencer la catégorie de la prédiction, surtout pour des objets asymétriques qui sont fréquemment retrouvés à la verticale.
- Enfin, lorsqu'une tasse blanche est filmée de très près, l'entièreté de ses contours n'est pas visible, ce qui peut compromettre la précision de la détection.

Suivi :

- L'attribution cohérente d'un même ID à chaque tasse fait face à plusieurs obstacles. Tout d'abord, le nombre fluctuant de tasses dans une seule image, pouvant aller jusqu'à 7, pose un défi pour le modèle, qui doit gérer efficacement cette variabilité tout en maintenant l'association correcte entre chaque tasse et son ID.
- De plus, lorsque les tasses sortent complètement du champ de vision de la caméra, comme lorsque la tasse rouge est soulevée et redéposée, le modèle doit se rappeler leur dernière position et leurs caractéristiques pour les reconnaître à leur réapparition, plusieurs images plus tard.

Justification de la méthode

Le choix de YOLOv8 pour la détection et le suivi des tasses dans notre séquence vidéo se base avant tout sur une prise en charge intégrée de la détection d'objets et du suivi multi-objets, ce qui répond directement aux besoins du TP. Ensuite, YOLOv8 à la capacité à répondre efficacement à plusieurs des défis identifiés, tout en reconnaissant certains obstacles qui resteront complexes à surmonter.

Problèmes faciles à résoudre:

- Contraste et occultations partielles: Grâce à des configurations spécifiques et une optimisation des paramètres de détection, YOLOv8 peut efficacement gérer les variations de contraste et les occultations partielles. Même si des tasses blanches proches avec des contours difficiles à distinguer pourraient induire des erreurs de détection, la robustesse de YOLOv8 devrait minimiser ces incidents.
- Présence de verres: YOLOv8 a été entraîné sur une large gamme d'objets, y compris des tasses et des verres, donc il devrait être en mesure de les distinguer correctement ces objets similaires.
- Nombre élevé de tasses: YOLOv8 est conçu pour identifier et faire le suivi de nombreux objets simultanément, ce qui le rend idéal pour gérer la présence de plusieurs tasses. Sa capacité à détecter plusieurs dizaines de piétons en même temps dans certaines applications assure que le suivi de 7 tasses devrait se faire sans encombre.
- Tasses filmées de près: La détection de tasses filmées de près devrait être efficace avec YOLOv8, grâce à la visibilité de caractéristiques distinctives telles que la poignée.

Problèmes difficiles à traiter:

- Variation de netteté: Les contours très flous dus à une forte variation de netteté pourraient poser problème à YOLOv8, car cela affecte directement la capacité du modèle à identifier clairement les objets.

- Distorsion des formes: Les variations extrêmes de netteté et le phénomène de distorsion des formes, en particulier pour les tasses visibles au travers les verres, posent un défi significatif. Ces facteurs altèrent beaucoup l'apparence des tasses, les rendant difficiles à détecter de manière fiable.
- Petites tasses lointaines: La détection de petites tasses éloignées reste un défi, en particulier lorsqu'elles sont difficiles à distinguer du fond. Leur petite taille réduit la quantité d'informations visuelles disponibles pour la détection.
- Identification après sortie et retour: Le suivi d'une tasse qui sort puis revient dans le champ de vision peut s'avérer complexe. YOLOv8 peut avoir du mal à maintenir l'identité constante de l'objet à travers ces interruptions, surtout puisque l'architecture concentre son attention sur la trame actuelle et ne possède pas un réel mécanisme de mémoire.

Description de l'implémentation utilisée

Tout d'abord, le modèle YOLOv8x pré-entrainé est chargé à partir de la libraire Ultralytics. Le modèle est testé sur les bases de données MOT dans lesquels se trouvent un fichier de vérité similaire à celui fournie par le professeur pour ce travail, mais avec une syntaxe différente. [4] Les données sont passées dans le modèle, une image à la fois et en extrait les identifiants ainsi que les classes de tous les objets détectés. Ces résultats sont comparés au fichier de vérité (groundtruth) avec le calcul du métrique HOTA (Higher Order Tracking Accuracy) et HOTA(0) qui a été implémenté manuellement. La valeur de seuil de localisation *alpha* de 0.05 est utilisée par défaut.

Le modèle YOLOv8n a été choisi puisqu'il est le plus simple de la gamme YOLOv8, donc le temps de calcul est plus court. Malgré que ce modèle offre de moins bonnes performances en termes de score HOTA, l'économie de temps était d'ailleurs préconisée pour la réalisation de ce travail.

La configuration de l'algorithme de suivi peut se faire grâce à plusieurs arguments dans la fonction «model.track», comme CONF et IOU. CONF définit le seuil de confiance (CONF) minimum pour les détections, i.e. les objets détectés avec une confiance inférieure à ce seuil sont ignorés. L'ajustement de cette valeur peut aider à réduire les faux positifs. Ensuite, des valeurs plus faibles pour le seuil de l'intersection sur l'union (IOU) entraînent moins de détections en éliminant les boîtes qui se chevauchent, ce qui est utile pour réduire les doublons. Les valeurs par défaut de CONF et IOU sont respectivement de 0.25 et 0.7.

Pour tester les performances de la méthode, la jeu de données public MOT20 a été utilisé. Ce *dataset* est utilisé pour un défi qui consiste à faire la détection et le suivi de piétons. Ce benchmark contient des vidéos dans des environnements non contraints, et le suivi et l'évaluation sont effectuée en coordonnées d'image. Dans ce travail, les vidéos MOT20-01, MOT20-02 et MOT20-03 sont principalement utilisées.

Présentation des résultats de validation

Algorithme de suivi

Le modèle YOLO de Ultralytics supporte les algorithmes de suivi ByteTrack et BoT-SORT. Le Tableau 1 compare les performances de ces deux *trackers* sur la vidéo MOT20-01.

Tableau 1: Comparaison des performances de deux trackers

Tracker	Temps de calcul (sec.)	Score HOTA
ByteTrack	56	0.151
BoT-SORT	57	0.152

On peut voir que le temps de calcul ainsi que le score HOTA sont quasiment identiques. Le *tracker* BoT-SORT a été choisi puisqu'il est l'algorithme de suivi par défaut.

Taille du modèle

Le Tableau 2 compare les performances et le temps de calcul de différentes tailles du modèle YOLOv8, soit YOLOv8n, YOLOv8m et YOLOv8x.

Tableau 2: Comparaison des performances des différentes versions de YOLOv8

Modèle	Temps de calcul (sec.)	Score HOTA
YOLOv8n	71	0.154
YOLOv8m	172	0.162
YOLOv8x	320	0.168

Comme prévu, les modèles plus grands prennent plus de temps d'inférence pour détecter et suivre les objets avec une précision plus élevée. Le modèle YOLOv8n a été retenu, car il permet d'obtenir un score HOTA seulement 8% plus faible que YOLOv8x, pour un temps de calcul 78% moindre.

Paramètres

Le Tableau 3 montre le score HOTA sur la vidéo MOT20-01 pour différentes combinaisons de valeurs de paramètres entourant les paramètres par défauts (

Tableau 3: Scores HOTA pour différents paramètres de suivi

CONF \ IOU	0.6	0.7 (default)	0.8
0.2	0.15003	0.1501	0.14995
0.25 (default)	0.15129	0.15129	0.15104
0.3	0.15023	0.15035	0.14999

On voit bien que les paramètres CONF et IOU ont dans ce cas-ci très peu d'impacts (moins de 1%) sur la performance du modèle. En conséquence, les paramètres par défaut ont été conservés.

Seuil de détection pour calcul de HOTA

Le métrique HOTA utilise un *threshold* qui permet d'ajuster la sensibilité de la localisation. Le Tableau 4 montre le scores HOTA obtenus pour différentes vidéos de la banque de données MOT20 avec un *threshold* de 0.05 et 0.

Tableau 4: Scores HOTA(0.05) et HOTA(0) pour différentes vidéos

Vidéo	Score HOTA(0.05)	Score HOTA(0)
MOT20-01	0.154	0.246
MOT20-02	0.381	0.502
MOT20-03	0.133	0.183

Détection des tasses

Les Figures 1 à 6 montrent la performance du modèle pour la détection des tasses. Ces captures d'écran seront analysées à la section « Discussion des résultats » pour expliquer les forces et les faiblesses de la méthode.

Figure 1: Détection et suivi des tasses

Figure 2: Détection et suivi des tasses

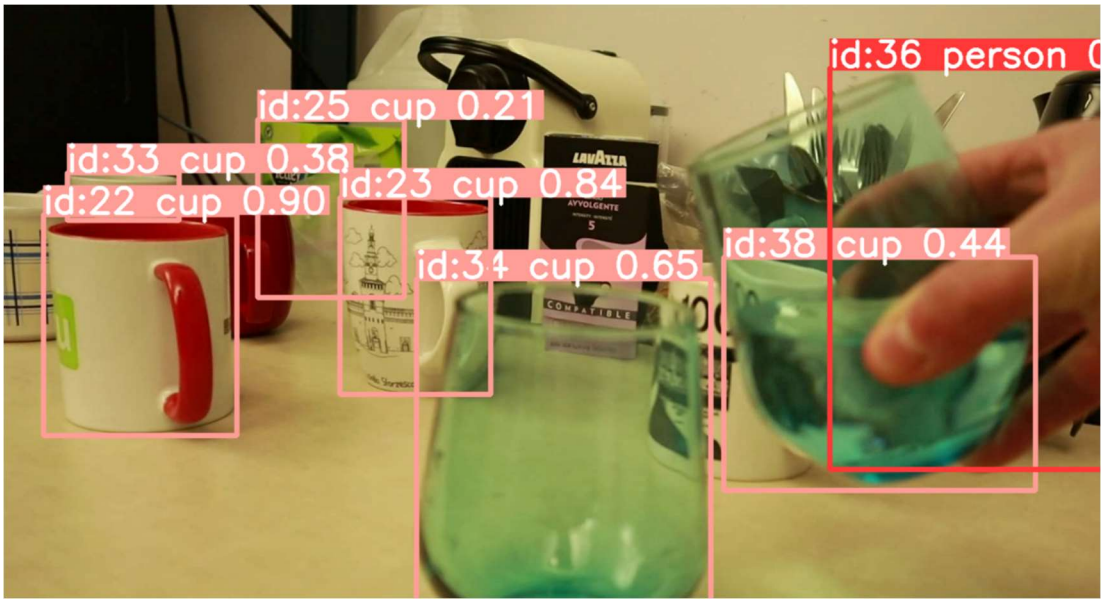

Figure 3: Détection et suivi des tasses

Figure 4: Détection et suivi des tasses

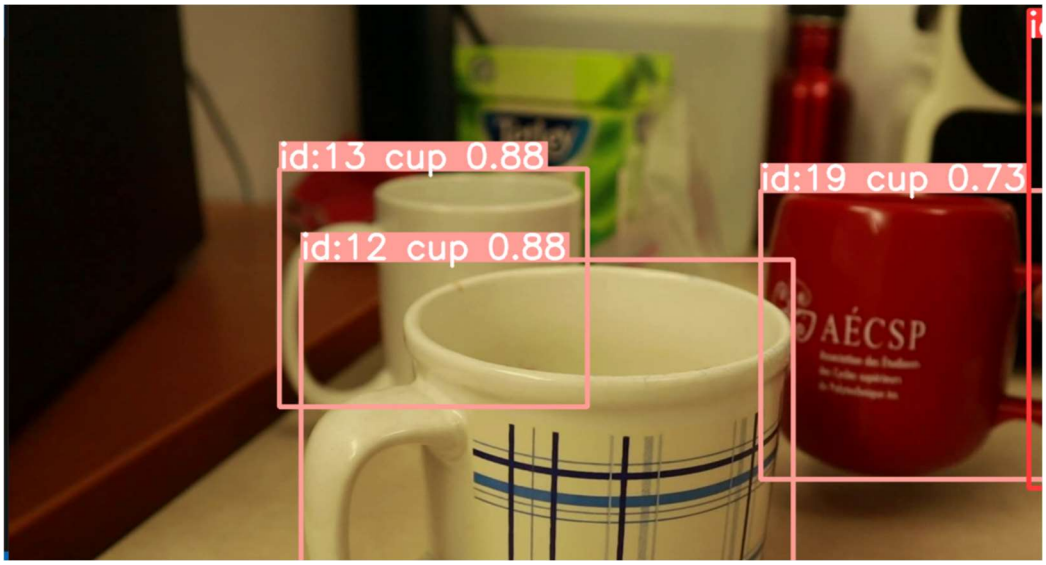

Figure 5: Détection et suivi des tasses

Figure 6: Détection et suivi des tasses

Discussion des résultats

HOTA vs HOTA(0)

La différence entre HOTA et HOTA(0) réside dans la valeur du seuil de localisation α (alpha) utilisé pour déterminer les correspondances entre les détections prédites et les détections de référence. HOTA utilise un seuil α de 0.05. Cela signifie que pour qu'une détection prédite soit considérée comme une correspondance positive avec une détection du *groundtruth*, l'Intersection sur Union (IOU) entre ces deux détections doit être supérieure à 0.05. Ce seuil permet d'éliminer les correspondances de faible qualité qui pourraient ne pas être pertinentes pour une évaluation précise du suivi. HOTA(0), d'autre part, utilise un seuil α de 0. Cela signifie qu'il n'y a pas de seuil minimal pour l'IOU, permettant ainsi toutes les correspondances potentielles entre les détections prédites et du *groundtruth*, indépendamment de la qualité de l'alignement spatial.

Dans le cas de ce travail, on peut tirer du Tableau 4 que le score HOTA moyen obtenu sur les vidéos de MOT20 est de 0.223, tandis qu'il s'élève à 0.310 en moyenne pour HOTA(0). Plusieurs facteurs peuvent expliquer que les scores de HOTA(0) soient meilleurs que ceux obtenus avec HOTA, mais la différence est principalement dû à la sensibilité aux correspondances de faible qualité. Effectivement, si le système de suivi génère de nombreuses détections avec des IOU très bas qui ne seraient normalement pas considérées comme des correspondances à un seuil plus élevé, ces détections sont incluses dans l'évaluation avec HOTA(0). Cela peut artificiellement augmenter le nombre de correspondances positives (true positives), améliorant ainsi le score global de suivi.

Forces du modèle

- Comme on peut le voir à la Figure 1, les sept tasses présentes sur l'image sont correctement détectées et identifiées. Le modèle n'a donc pas de difficulté à détecter et suivre de nombreux objets simultanément. Les scores de détection et de localisation en sont avantagés.
- Le modèle réussit à détecter les tasses qui sont partiellement occultés. En effet, on peut voir sur la Figure 1 que la tasse id:33 est cachée en partie par les tasses id:22 et id:35.
- Le modèle réussit également à détecter les tasses dont on ne distingue pas bien tous les contours. En effet, on voit sur la Figure 2 que la tasse id:2 est très près de la caméra et on ne voit ni la pognée ni le bas. Malgré tout, la tasse est correctement détectée.

Faiblesses du modèle

- La détection de tasses qui sont hors focus est plus difficile. Sur la Figure 2, on voit que le focus est sur la tasse id:2, alors que les trois autres tasses à gauche sont perçues avec moins de netteté. Conséquemment, une seule de ces trois tasses est correctement détectée. Cette erreur pourrait également être attribuée au fait que les deux tasses non détectées sont blanches comme l'arrière-plan, donc le manque de contraste peu nuire aux performances du modèle.
- La score de détection est également affecté négativement par la présence de faux positifs. En effet, on peut voir sur la Figure 3 que les deux verres sont identifiés incorrectement comme étant des tasses.
- La détection est limitée pour les objets qui sont petits ou éloignées. Par exemple, sur la Figure 6, les deux petites tasses déposées sur la tablette ne sont pas du tout détectées.
- Les scores de suivi sont plutôt faibles avec le modèle et les paramètres choisis. Effectivement, sur la Figure 4 a le id 11 tandis que sur la Figure 5, le id est passé à 19. Entre temps, la tasse a été soulevée hors du champ de vision de la caméra avant d'être redéposée. Le modèle n'a pas été en mesure de réidentifier correctement cette tasse. Cette limitation a également été observé pour les piétons sur les vidéos MOT20, et c'est un facteur qui nuit beaucoup au score HOTA.

Conclusion

Pour conclure, ce travail a exploré le suivi d'objets avec le modèle YOLOv8n et le tracker BoT-SORT, mettant en avant l'importance des seuils de détection dans le score HOTA. L'expérimentation a révélé que les scores HOTA(0) étaient supérieurs, indiquant une meilleure prise en compte des correspondances de faible qualité. Malgré une détection plutôt efficace, des défis demeurent avec les détections hors focus et les faux positifs. De plus, le système a montré des faiblesses dans la réidentification d'objets après leur disparition du champ de vision. Le modèle actuel, malgré ses limitations, offre un bon équilibre entre précision et efficacité, suggérant un potentiel pour des améliorations futures pour optimiser les performances en suivi d'objets en temps réel. Finalement, une amélioration au niveau de l'architecture serait d'ajouter un mécanisme de mémoire des caractéristiques des objets identifiés, afin de mieux les réidentifier.

Bibliographie

[1] Ultralytics, « Track ». Consulté le: 8 avril 2024. [En ligne]. Disponible à: https://docs.ultralytics.com/modes/track

[2] « YOLOv8: A New State-of-the-Art Computer Vision Model ». Consulté le: 8 avril 2024. [En ligne]. Disponible à: https://yolov8.com

[3] K. KALRA, « YOLO (You Only Look Onc », Medium. Consulté le: 8 avril 2024. [En ligne]. Disponible à: https://medium.com/@khwabkalra1/yolo-you-only-look-onc-523b01ec4f4d

[4] « MOT Challenge - Data ». Consulté le: 10 avril 2024. [En ligne]. Disponible à: https://motchallenge.net/data/MOT20/