Contribution à la caractérisation de l'affordance d'un environnement de travail industriel : une approche basée sur l'apprentissage profond combinant données réelles et synthétiques.

Sarah Ouarab[1,2], David Garcia[1], Nicolas Ragot[3], Yohan Dupuis[4]

[1] CESI LINEACT, Lyon, France
[2] École Nationale Supérieure des Arts et Métiers, Paris, France
[3] CESI LINEACT, Rouen, France
[4] CESI LINEACT, Paris, France

souarab@cesi.fr

Résumé

Ce travail s'inscrit dans le cadre du projet "École De La Batterie", dont l'un des objectifs concerne l'optimisation de la conception des postes de travail manuel dans le but d'améliorer leur ergonomie. Nos travaux s'inscrivent dans cette démarche et visent à caractériser l'affordance des éléments de ces environnements avec lesquels les opérateurs interagissent (outils, composants, etc.) lors des opérations qu'ils réalisent. Un des verrous liés à cette problématique concerne la détection des éléments mobilisés par l'opérateur au cours de son activité pour aboutir à la caractérisation de leurs affordances. Les approches basées sur l'apprentissage profond fournissent de très bons résultats, mais nécessitent des bases de données d'apprentissage importantes. Dans le contexte industriel ces bases de connaissances labellisées n'existent pas ou sont en quantité très limitées pour ce type d'application. La méthode proposée repose sur un processus d'apprentissage automatique supervisé qui mobilise à la fois la génération de données réelles et synthétiques, qui sont respectivement issues de l'expérimentation et du jumeau numérique d'un poste de travail. Nous questionnons notamment la proportion de données réelles requises pour obtenir un modèle performant avec un effort de labellisation minimal, pour atteindre des performances de détection des outils cohérentes pour notre objectif de caractérisation des affordances du poste de travail.

Mots-clés

Affordance, environnement de travail, industrie, apprentissage profond, jumeau numérique, base de données réelle et synthétique.

Abstract

This work is part of the "École De La Batterie" project, one of whose objectives is to optimize the design of manual workstations in order to improve their ergonomics. Our contribution aligns with this goal by seeking to characterize the affordances of the elements in these environments (tools, components, workspace, etc.) with which operators interact during their tasks. One of the main challenges here is detecting the elements mobilized by the operator over the course of their work, in order to characterize their affordances. Deep learning–based approaches provide excellent results but require large training datasets. In industrial settings, however, such labeled datasets either do not exist or are available only in very limited quantities for this type of application. The proposed method relies on a supervised machine learning process that uses both real and synthetic data, derived from experiments and the digital twin of a workstation, respectively. In particular, we investigate the proportion of real data required to develop an effective model with minimal labelling effort, aiming to achieve consistent tool detection performance for our goal of characterising workstation affordances.

Keywords

Affordance, work environment, industry, deep learning, digital twin, real and synthetic dataset.

1 Introduction

Le concept d'affordance est né et a été popularisé en 1977 par Gibson [3], qui le définit comme les possibilités d'action ou d'utilisation offertes par un objet. En 1988, Norman reprend ce concept pour affirmer que l'affordance résulte de la relation entre les propriétés de l'objet et les capacités de l'agent [7]. En 2020, Simonian étend ce concept pour intégrer les inférences du sujet (l'humain dans notre cas) en situation réelle, en considérant que les affordances sont dépendantes de ce qu'il perçoit des propriétés de l'objet, pour agir [9]. Issu des champs disciplinaires de la psychologie cognitive et de la perception, le concept d'affordance a été progressivement investi par de nombreuses autres thématiques, notamment celles des sciences de l'ingénieur et de l'information et plus précisément celles de la robotique (sociale) et des interactions humains-systèmes [4]. Par ailleurs, la dynamique de l'industrie 5.0, qui repositionne l'humain au centre des systèmes de production, fait émerger de nombreux défis, notamment l'optimisation des inter-

actions entre les opérateurs et leur environnement de travail, ainsi que l'adaptation des outils et équipements aux besoins spécifiques des utilisateurs, ce qui implique une amélioration continue de l'ergonomie des postes de travail. Dans ce travail, nous cherchons à contribuer à cette problématique en caractérisant l'affordance des outils utilisés lors d'une tâche d'assemblage. Cela passe notamment par la détection des outils mobilisés par l'opérateur (cf. Figure 1). De nombreux travaux ont été menés sur les algorithmes d'apprentissage profond pour la détection d'objets. Cependant, ces méthodes étant particulièrement gourmandes en données d'entraînement, leur transposition au contexte industriel reste complexe en raison du manque de données annotées. Pour pallier ce problème, l'utilisation du jumeau numérique est intéressant, car il permet de générer des données synthétiques annotées massivement. Cependant, des modèles entraînés uniquement sur des données synthétiques n'offrent pas des performances acceptables [10]. Nos travaux visent à étudier une alternative qui repose sur des jeux de données d'entraînement mixtes puisqu'ils combinent des données réelles et synthétiques. En particulier, nous questionnons la proportion optimale de données réelles annotées nécessaires dans un jeu de données d'entraînement pour atteindre de bonnes performances en détection d'objets. Ici, le terme optimal renvoie au nombre minimal de données réelles nécessaire afin de minimiser le processus d'annotation tout en maximisant les performances du modèle. L'article est organisé de la façon suivante : nous présentons d'abord un bref état de l'art, suivi de notre approche méthodologique afin de créer les jeux de données et d'étudier les performances du modèle. Enfin, nous discutons des résultats obtenus avant de conclure et d'ouvrir sur des perspectives.

2 Travaux connexes

L'utilisation de données synthétiques pour entraîner des modèles d'apprentissage profond a montré son efficacité dans plusieurs domaines, par exemple pour le comptage de piétons [5] et celui de la classification et la détection de défauts dans l'acier [2]. Ces études démontrent que les données synthétiques peuvent combler les lacunes des ensembles de données réelles, notamment en cas de déséquilibre ou de manque de diversité. Une étude récente explore l'utilisation des données synthétiques pour l'entraînement des modèles de détection d'objets en milieu industriel [8]. L'étude évalue plusieurs proportions de données réelles et synthétiques pour l'entraînement des modèles de détection YOLOv8, et démontre qu'un jeu de données mixte de 890 échantillons, dont seulement 3% sont réels, permet d'atteindre des performances satisfaisantes. Enfin, l'étude de l'utilisation de jeux de données synthétiques pour l'estimation de la pose 3D d'objets industriels a été menée [6]. Elle confirme que les données synthétiques représentent une solution efficace face à la rareté des données réelles, tout en offrant une meilleure adaptation aux besoins spécifiques des cas d'usage. Ces travaux nous ont motivés à intégrer les données synthétiques à l'entraînement de notre modèle de détection, et à évaluer l'impact de cet ajout en fonction du ratio de données réelles utilisées. L'objectif est de déterminer le ratio minimal de données réelles qui permet d'obtenir des performances acceptables en termes de précision des détections. Cette approche vise à réduire la dépendance aux données réelles annotées manuellement, souvent coûteuses et limitées en quantité.

3 Méthode

Dans ce travail, nous nous intéressons à la caractérisation de l'affordance des outils utilisés par un opérateur humain sur un poste de travail, dans le cadre d'une tâche d'assemblage de composants. Cette caractérisation repose sur la détection des outils à l'aide de nos caméras, permettant ainsi le suivi en temps réel de leur utilisation. Notre approche repose sur l'apprentissage par transfert (*transfer learning*), une technique qui consiste à initialiser le modèle avec des poids pré-entraînés sur un grand ensemble de données (COCO pour YOLOv9) avant de l'adapter à notre tâche spécifique. L'entraînement a été réalisé sur 100 epoch avec une taille de lot (*batch size*) de 16 en utilisant les poids de YOLOv9 (Gelan-C).

3.1 Indicateurs de performance

Pour évaluer les performances du modèle, nous avons choisi d'utiliser les indicateurs mAP@0.5 (mAP : mean Average Precision) et mAP@0.5-95 et le f1-score, employés pour la détection d'objets :

- **mAP@0.5** : seuil de chevauchement (Intersection over Union, IoU) de 50% qui évalue la capacité du modèle à détecter et localiser les objets dans l'image.
- **mAP@0.5-0.95** : moyenne des mAP pour un intervalle de seuils d'IoU de [50%,$\cdots$,95%]. Ceci fournit une mesure plus précise de la détection et de la localisation des objets.
- **f1-score** : évalue la performance du modèle à distinguer les vrais positifs des faux positifs et des faux négatifs.

3.2 Collecte de données réelles

Des données réelles ont été collectées lors d'expérimentations sur un poste de travail manuel réel, où un opérateur réalise diverses actions avec les outils. Un ensemble de 200 images réelles a été enregistré. 80% servent à composer les bases de données d'entraînement et 20% sont dédiées à la base de données de test. L'annotation de ces images a été réalisée manuellement à l'aide de l'outil Roboflow, garantissant des annotations précises adaptées au format YOLO. Le temps d'annotation d'une image réelle est de l'ordre de 5 minutes sachant que celui-ci peut considérablement s'allonger en fonction de la complexité de la scène traitée et du nombre d'éléments à annoter.

3.3 Génération des données synthétiques

Les données synthétiques ont été générées avec Unity 3D et son module *Perception Package*, permettant la création automatisée de jeux de données annotés pour la vision par ordinateur. Afin de simuler un poste de travail manuel dans

des conditions réalistes, le Jumeau Numérique de l'Atelier Flexible de Production (JN-UFP), développé par l'équipe de recherche du campus CESI Rouen, a été exploité (cf. Figure 1-a). Ce jumeau numérique a été grandement amélioré dans le cadre du projet JENII [1], notamment son niveau de détail et la qualité de son rendu 3D hautement réaliste. Les outils modélisés au sein du JN-UFP correspondent aux cinq classes d'outils principales présentes sur le poste de travail réel, illustrées dans la Figure 1 : tournevis cruciforme, tournevis plat, clé Allen, clé plate et un jeu de clés Allen. La génération des données synthétiques a mobilisé des ressources de calcul modérées : environ 2 heures ont suffi pour produire 320 images sur une station dotée d'un processeur Intel Xeon W-2245 (3,9 GHz, 8 cœurs, 16 threads, 16,5 Mo de cache L3) et de 64 Go de RAM ainsi qu'une carte graphique NVIDIA Quadro RTX 6000.

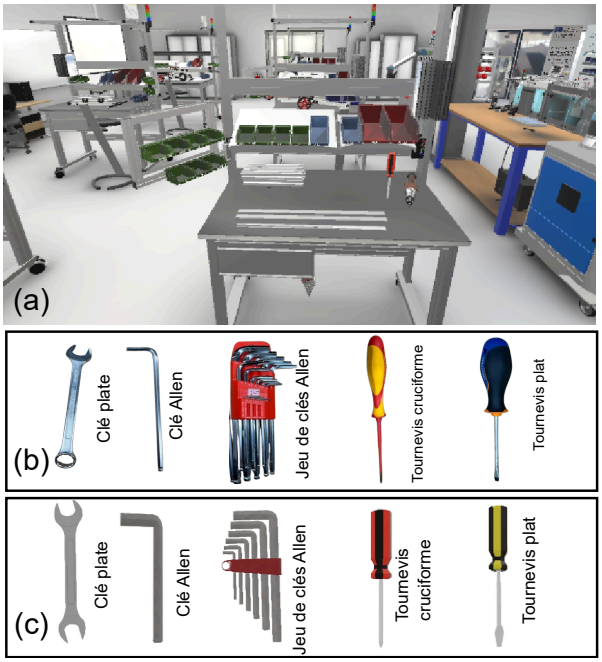

FIGURE 1 – (a) Vue d'un poste de travail du Jumeau numérique UFP (b) et (c) Vues des cinq classes d'outils mobilisés respectivement au sein du poste de travail réel et du poste de travail virtuel du JN-UFP.

4 Résultats et discussion

L'objectif de cette analyse est d'évaluer l'influence de la proportion des données réelles dans un jeu de données d'entraînement mixtes. Notre démarche vise à caractériser l'effort d'annotation de données réelles requis pour atteindre des performances acceptables pour l'application visée.

4.1 Performances obtenues avec un entraînement sur des données réelles

Pour rappel 200 images constituent le jeu de données réelles qui a été scindé en 160 images d'entrainement et 40 images pour le test. Ainsi le modèle entraîné sur ces données réelles a obtenu des performances prometteuses avec un mAP@0.5 de 0.943, un mAP@0.5-0.95 de 0.66 et un f1-score de 0.92. Ces résultats démontrent une bonne capacité de détection des objets avec un seuil d'IoU de 50%, bien que la précision diminue avec des seuils plus stricts. Néanmoins, un effort d'annotation important a été nécessaire pour labelliser les images. Dès lors, il est intéressant de questionner la proportion minimale de données réelles nécessaire dans le jeu de données d'entraînement en y ajoutant des données synthétiques issues du jumeau numérique.

4.2 Performances obtenues à partir de données mixtes

Nous avons créé onze bases de données d'entraînement contenant un nombre variable d'images synthétiques et réelles. La taille totale des différents jeux de données a été maintenue constante à 320 images. Le nombre d'images réelles dans les différents jeux de données d'entraînement a été progressivement augmenté jusqu'à atteindre un ratio de 50% d'images réelles, à savoir 160 images (totalité du nombre d'images réelles disponibles pour entraîner un modèle). La base de test est identique à celle utilisée pour l'entraînement sur le jeu de données réelles et commune à tous les jeux de données d'apprentissage mixtes : 40 images réelles (cf.4.1). Les résultats présentés Figure 2 montrent d'une part l'évolution des indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et f1-score(mixtes) pour les modèles entraînés sur les bases de données mixtes, en fonction du pourcentage de données réelles (courbes pleines) et d'autre part l'évolution des indicateurs mAP@0.5(réelles), mAP@0.5-0.95(réelles) et f1-score(réelles) pour les modèles entraînés uniquement à partir des données réelles considérées pour chacune des bases de données mixtes (courbes en pointillés).

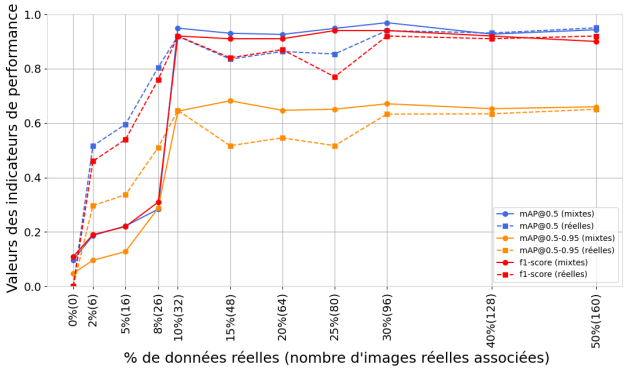

FIGURE 2 – Évolution des indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et f1-score(mixtes) (courbes pleines) / Évolution des indicateurs mAP@0.5(réelles), mAP@0.5-0.95(réelles) et f1-score(réelles) (courbes en pointillés).

On observe que lorsque le jeu de données mixtes comporte moins de 10% de données réelles, les performances via les indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et en f1-score(mixtes) se révèlent faibles. En revanche, dès que la proportion de données réelles est au delà de 10%,

les scores mAP@0.5(mixtes), mAP@0.5–0.95(mixtes) et f1-score(mixtes) augmentent drastiquement puis se stabilisent à des niveaux de performances beaucoup plus élevés, compris respectivement dans les intervalles [0.926, 0.969], [0.644, 0.682] et [0.90, 0.94]. On note également qu'entre 10% et 50%, la dispersion des valeurs autour de la valeur médiane, calculée pour chacun des indicateurs f1-score(mixtes), mAP@0.5(mixtes) et mAP@0.5-0.95(mixtes), varie peu. La meilleure performance est obtenue avec un ratio de 30% de données réelles : mAP@0.5(mixtes) de 0.969, mAP@0.5-0.95(mixtes) de 0.671 et un f1-score(mixtes) de 0.94. Si, désormais on s'intéresse aux évolutions des indicateurs mAP@0.5(réelles), mAP@0.5-0.95(réelles) et f1-score(réelles) comparativement à celles obtenues pour les indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et f1-score(mixtes) on observe, sur la figure 2 trois comportements distinctifs. Pour l'intervalle [0%(0) à 10%(32)[, les performances des indicateurs (mixtes) sont très inférieures à celles des indicateurs (réelles). Ceci souligne que les données synthétiques associées aux données réelles dégradent ici globalement la performance des modèles. Pour l'intervalle [10%(32) à 30%(96)], les performances des indicateurs (mixtes) sont supérieures à celles des indicateurs (réelles). Les données synthétiques associées aux données réelles contribuent ici de manière significative aux performances des modèles sauf pour le jeu de données à 10% où seul la valeur de l'indicateur mAP@0.5(mixtes) se détache légèrement de celle de mAP@0.5(réelles). Pour l'intervalle]30% (96), 50% (160)], l'ajout de données synthétiques n'apporte plus d'amélioration significative sur les performances des modèles. En conclusion, les résultats obtenus pour l'intervalle de pourcentage de données réelles [10% 30%] mettent en évidence l'intérêt d'intégrer à un noyau minimal de données réelles des données synthétiques pour minimiser l'effort d'annotation tout en atteignant une performance acceptable pour le cas d'étude considéré.

5 Conclusion

Ce travail s'inscrit dans la perspective de contribuer à la caractérisation des affordances des éléments constitutifs d'un environnement de travail en développant un modèle de détection et de suivi de trajectoires d'outils basé sur l'apprentissage profond. Pour pallier le manque de données annotées en milieu industriel, nous avons combiné des données réelles issues d'expérimentations et des données synthétiques générées à partir d'un jumeau numérique. Nos expériences ont montré que l'ajout de données synthétiques à un noyau minimal de données réelles permet de franchir un palier en termes de performances. Dans le cadre de notre cas d'étude, cette proportion est comprise entre 10% et 30% pour la base d'entraînement. Les pistes d'amélioration envisagées portent sur l'optimisation de la contribution des données synthétiques aux performances du modèle, tout en visant à réduire la quantité de données réelles pour minimiser l'effort d'annotation. Selon nos analyses en cours, cela implique notamment l'amélioration du processus de génération des données virtuelles, par l'augmentation entre autres du nombre de points de vue et la réduction de l'écart de réalité entre les environnements réel et virtuel.

Remerciements

Ces travaux sont financés dans le cadre du projet Ecole de la Batterie (EDLB), opération soutenue par l'État dans le cadre de l'AMI « Compétences et Métiers d'Avenir » du Programme France 2030, opéré par la Caisse des Dépôts » (La Banque des Territoires).

Références

[1] ANR. Projet JENII. https ://anr.fr/ProjetIA-21-DMES-0006, 2025.

[2] Aleksei Boikov, Vladimir Payor, Roman Savelev, and Alexandr Kolesnikov. Synthetic data generation for steel defect detection and classification using deep learning. *Symmetry*, 13 :1176, 2021.

[3] James Jerome Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[4] Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and Jose Santos-Victor. Affordances in psychology, neuroscience, and robotics : A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10 :4–25, 2018.

[5] Hadi Keivan Ekbatani, Oriol Pujol, and Santi Segui. Synthetic data generation for deep learning in counting pedestrians. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, page 318–323. SCITEPRESS - Science and Technology Publications, 2017.

[6] Aristide Laignel, Nicolas Ragot, Fabrice Duval, and Sarah Ouarab. *Synthetic Datasets for 6D Pose Estimation of Industrial Objects : Framework, Benchmark and Guidelines*, page 227–241. Springer Nature Switzerland, 2024.

[7] Donald A Norman. *The design of everyday things*. Bantam Doubleday Dell Publishing Group, 1990.

[8] Sarah Ouarab, Rémi Boutteau, Katerine Romeo, Christele Lecomte, Aristid Laignel, Nicolas Ragot, and Fabrice Duval. *Industrial Object Detection : Leveraging Synthetic Data for Training Deep Learning Models*, page 200–212. Springer Nature Switzerland, 2024.

[9] Stéphane Simonian, Rawad Chaker, and Jonathan Kaplan. Affordance en e-formation et régulation de l'apprentissage : une exploration dans un contexte d'études universitaires. *TransFormations : Recherche en éducation et formation des adultes*, 2020.

[10] T Zgheib, H Borges, V Feldman, T Guntz, C Di Loreto, O Desmaison, and F Corduant. Détection d'objets en temps réel : Entraînement de réseaux de neurones convolutifs sur images réelles et synthétiques. In *Conférence Nationale en Intelligence Artificielle*, 2022.