



**POLYTECHNIQUE
MONTRÉAL**

**UNIVERSITÉ
D'INGÉNIERIE**

INF6804 - Visions par ordinateur

Hiver 2024

Travail Pratique 02

Groupe 01

Alexandre Gélinas - 2083465

Elizabeth Michaud - 2073093

Soumis à : Khalil Sabri

Guillaume-Alexandre Bilodeau

Date :

18 mars 2024

Table des matières

1. Présentation des deux méthodes à comparer.....	2
2. Hypothèses de performance pour des cas spécifiques.....	2
3. Description des expériences, séquence de la base de données et critères d'évaluation.....	3
4. Description des deux implémentations utilisées.....	4
5. Présentation des résultats de tests.....	5
5.1. Résultats pour le dataset Highway.....	5
5.2. Résultats pour le dataset Office.....	7
5.3. Résultats pour le dataset Pedestrians.....	8
5.4. Résultats pour le dataset PETS2006.....	9
6. Discussion des résultats et retour sur les hypothèses.....	10
Références.....	12

1. Présentation des deux méthodes à comparer

Ce travail pratique s'intéresse à deux algorithmes de segmentation des régions d'intérêt dans les vidéos, soit la segmentation des régions d'intérêt par soustraction d'arrière-plan et la segmentation des régions d'intérêt par segmentation d'instances.

La segmentation des régions d'intérêt par soustraction d'arrière-plan capture d'abord un modèle d'arrière-plan à l'aide d'une ou plusieurs trames de la vidéo. Cela a pour but de déterminer les éléments statiques de la vidéo qui ne bougeront pas ou qui bougeront très peu ou très lentement. Ensuite, ce modèle d'arrière-plan est comparé avec chacune des trames de la vidéo dans le but de trouver les éléments qui sont différents entre les deux images. Il faut faire une comparaison pixel par pixel des deux images en soustrayant la valeur du pixel de la première image de la valeur du pixel correspondant de la seconde image. Si le résultat en valeur absolue de cette soustraction est plus élevé qu'un certain seuil établi, alors le pixel sur la trame de la vidéo est considéré comme faisant partie d'un élément d'avant-plan.

La segmentation des régions d'intérêt par segmentation par instances essaie d'identifier les objets présents dans l'image afin de leur appliquer un masque pour les différencier. De ce fait, lorsque le modèle retrouve plusieurs types d'objets dans la même image, nous pouvons facilement différencier ces objets. De plus, la région identifiée contient plusieurs variants de pixel afin de déterminer le niveau possible d'appartenance de l'objet à la région. Il y a ainsi des pixels moins forts sur les bords de l'objet détecté et beaucoup plus forts dans le centre de ce dernier. Le but est donc de sélectionner uniquement les objets mobiles de l'image afin d'y retrouver l'arrière-plan en appliquant un filtre sur les objets trouvés. Enfin, pour que le modèle soit fonctionnel, celui-ci doit être entraîné afin de pouvoir détecter les objets voulus.

Ainsi, le but de ce travail pratique est de comparer ces deux méthodes de segmentation des régions d'intérêt dans les vidéos pour trouver quelle méthode est la meilleure dans quelles circonstances. Nous ferons cela en appliquant chaque méthode sur quatre vidéos différentes et en analysant les résultats obtenus avec une certaine technique de comparaison.

2. Hypothèses de performance pour des cas spécifiques

Selon notre compréhension des deux méthodes, nous posons l'hypothèse que la méthode de segmentation par instances sera meilleure que la méthode de segmentation par soustraction d'arrière-plan dans le cas où les régions d'intérêt ne sont pas en mouvement (**hypothèse 1**). En effet, puisque la segmentation par soustraction d'arrière-plan se base sur plusieurs images pour former son modèle d'arrière-plan, si un objet est en mouvement au début de la vidéo, mais s'immobilise peu de temps après pour rester statique pendant un long moment, il pourrait être considéré comme un élément de l'arrière-plan. Ainsi, cet objet ne serait plus détecté par la soustraction d'arrière-plan.

Ensuite, nous pensons que la méthode de segmentation par instances sera également meilleure que la méthode de segmentation par soustraction d'arrière-plan dans les cas où la région d'intérêt principale devient

partiellement occultée par un autre objet (**hypothèse 2**). Effectivement, la segmentation par instances est capable de détecter et de catégoriser les différents objets et régions d'intérêt dans une vidéo, alors que la segmentation par soustraction d'arrière-plan ne différencie pas les objets entre eux. Donc, la segmentation par instances sera en mesure de trouver plusieurs régions d'intérêt, mais la segmentation par soustraction d'arrière-plan n'en trouvera qu'une seule.

Par la suite, nous pensons que la méthode de segmentation par instances sera meilleure que la méthode de segmentation par soustraction d'arrière-plan dans les cas où l'arrière-plan n'est pas complètement statique (**hypothèse 3**). Puisque le modèle d'arrière-plan de la méthode par soustraction d'arrière-plan se base sur une idée générale de l'arrière-plan, les petites variations de ce dernier peuvent créer du bruit et affecter négativement les résultats en identifiant des régions d'intérêt là où il n'y en a pas.

Notre dernière hypothèse est que la méthode de segmentation par soustraction d'arrière-plan sera meilleure que la méthode de segmentation par instances dans les cas où la région d'intérêt est un objet qui se déplace très vite créant ainsi du flou dans la vidéo (**hypothèse 4**). Nous pensons cela, car la segmentation par instances essaie de déterminer les contours précis d'une région d'intérêt, ce qui est difficile si la région est floue.

Finalement, de manière générale, nos hypothèses convergent vers l'idée que la méthode par segmentation d'instances sera beaucoup plus précise dans les résultats que celle par soustraction d'arrière-plan. Elle devrait, selon nous, détecter mieux les principales régions d'intérêts de la vidéo.

3. Description des expériences, séquence de la base de données et critères d'évaluation

Pour commencer nos expériences, nous avons tout simplement implémenté la logique nécessaire à l'exécution de chacun des modèles. Lorsque l'implémentation était fonctionnelle, nous avons ajusté les paramètres de nos modèles en fonction de petit échantillon sur notre base de données. Les détails d'ajustement des paramètres se trouvent dans la section suivante. Ensuite, nous avons importé les images de *ground truth* de notre base de données pour les comparer avec les résultats de nos méthodes. Nous avons utilisé la normalisation pour l'évaluation des modèles, soit la sommation absolue de la différence des pixels capturés. Nous avons terminé par la normalisation de nos résultats en fonction de la taille de l'image d'entrée afin d'avoir un pourcentage de différence entre notre résultat et celui voulu.

Pour tester notre hypothèse 1, nous utiliserons une trame parmi les trames 1100 à 1275 de la vidéo Office. Durant ces trames, la personne, qui était entrée dans la pièce dans les trames précédentes, ne se déplace presque pas, ce qui permet bien de tester notre première hypothèse.

Pour tester notre hypothèse 2, nous utiliserons la trame 602 de la vidéo PETS2006. Dans cette trame, il y a une personne qui se déplace derrière une autre personne qui ne bouge pas. Nous pourrions ainsi voir si la segmentation par instances différencie ces deux personnes.

Pour tester notre hypothèse 3, nous utiliserons la trame 1071 de la vidéo Highway. Dans cette vidéo, les feuilles des arbres sur le côté gauche bougent beaucoup, ce qui représente un défi pour la segmentation par soustraction d'arrière-plan.

Pour tester notre hypothèse 4, nous utiliserons la trame 480 de la vidéo Pedestrians, puisqu'elle contient un cycliste flou se déplaçant rapidement. Nous pourrions ainsi comparer le déplacement et l'identification de l'objet.

En ce qui concerne la base de données, elle nous a été fournie en même temps que l'énoncé de ce travail pratique. De ce fait, nous avons utilisé la catégorie *Baseline* de la base de données CDNET 2012. Celle-ci inclut d'ailleurs des images de *ground truth* nous permettant de mieux comparer les résultats de nos expériences. Nous avons dû retirer le début de chaque vidéo fournie avec l'énoncé, puisque les images *ground truth* n'étaient visiblement pas bonnes. Elles étaient entièrement grises, alors qu'il y avait clairement des régions d'intérêt dans la vidéo.

4. Description des deux implémentations utilisées

Notre implémentation de la méthode par soustraction d'arrière-plan est fortement inspirée de l'implémentation non paramétrique ViBe présentée en classe et disponible sur ce site GitHub : [INF6804/NonParamBGS.ipynb at master · gabilodeau/INF6804 · GitHub](https://github.com/INF6804/NonParamBGS.ipynb). Tout d'abord, notre modèle d'arrière-plan est composé de plusieurs trames. Nous sélectionnons 1 trame à toutes les 50 trames, donc le nombre de trames formant le modèle d'arrière-plan dépend de la longueur de la vidéo. Nous avons décidé de prendre 1 trame sur 50, car plus de trames que cela (par exemple, 1 trame sur 20) n'apporte pas d'amélioration significative et ralentit l'exécution de notre implémentation et moins de trames que cela (par exemple, 1 trame sur 80) ne donnent pas assez d'information pour obtenir un modèle d'arrière-plan représentatif de toutes les parties des vidéos à analyser. Ensuite, pour obtenir les régions d'intérêt, nous comparons chaque trame de la vidéo avec chaque trame faisant partie du modèle d'arrière-plan (soit 1 trame sur 50). Pour chacune de ces comparaisons, nous effectuons une comparaison pixel par pixel en soustrayant la valeur des pixels de l'image faisant partie du modèle d'arrière-plan à la valeur des pixels de la trame à analyser. Les résultats de ces soustractions sont mis en valeur absolue pour évaluer si la différence des pixels correspondants est plus grande qu'un certain seuil que nous avons établi à 25. Nous avons testé plusieurs seuils et 25 est celui qui donne les meilleurs résultats, car un seuil plus bas est trop sensible aux bruits et détecte beaucoup de régions d'intérêt non pertinentes alors qu'un seuil plus haut ne détecte pas suffisamment bien les régions d'intérêt. Lorsque la différence est supérieure ou égale à 25, le pixel est considéré comme faisant partie de l'avant-plan pour cette image particulière du modèle de l'arrière-plan. Pour chaque pixel, nous cumulons combien de fois le pixel est considéré comme faisant partie de l'avant-plan. Tous les pixels qui font partie de l'avant-plan dans 70% des images ou formant le modèle d'arrière-plan ou plus sont identifiés comme faisant partie d'une région d'intérêt. Nous avons choisi 70% comme seuil en testant des seuils différents encore une fois. Nous avons suivi le même raisonnement que pour le seuil de 25 expliqué précédemment. Pour la lecture des trames (images), nous avons utilisé la librairie *cv2* et pour mettre les résultats de soustraction en valeur absolue, nous avons utilisé la librairie *numpy*.

Pour ce qui est de la méthode par segmentation d'instances, nous avons directement utilisé *maskrcnn_resnet50_fpn* de la librairie PyTorch avec les données d'entraînement par défaut, soit ceux de COCO_V1. Comme le résultat du modèle sépare chaque objet individuellement dans un masque, nous avons simplement additionné l'ensemble de ses masques afin de créer un seul masque de région d'intérêt. En ce qui concerne la lecture, pour ce qui est du seuil de confiance du modèle, nous avons utilisé un seuil de 30% afin d'enlever les principaux objets qui ne sont pas bien détectés. De plus, nous n'avons pas appliqué de filtre sur les masques afin de bien comparer les résultats des deux modèles pour n'importe quel type de vidéo et non seulement ceux qui nous ont été donnés. De ce fait, il pourra y avoir des éléments statiques qui sont détectés, mais qui ne forment pas de régions d'intérêts dans nos résultats pour ce modèle, mais cela est voulu.

5. Présentation des résultats de tests

Nous avons mis en forme nos résultats en graphique pour chaque base de données utilisée afin de pouvoir facilement comparer ces dernières et pour pouvoir étudier leurs différences.

5.1. Résultats pour le dataset Highway

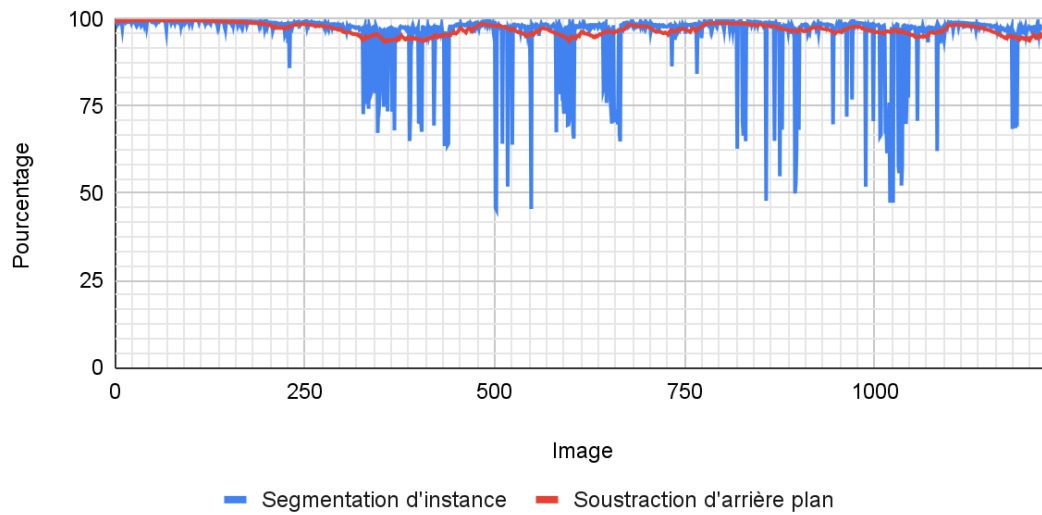
Le tableau I ainsi que le graphique I montrent les résultats obtenus pour les deux méthodes pour la base de données Highway. Les valeurs possibles se situent entre 0% pour une complète différence de correspondance avec le ground truth et 100% pour une correspondance parfaite. Nous avons résumé l'ensemble des résultats par une moyenne de pointage qu'on retrouve dans le tableau.

Tableau I : Résultat global de précisions pour la base de données Highway

Highway		
Résultat	Segmentation d'instances	Soustraction d'arrière-plan
Moyenne	95,61%	97,09%


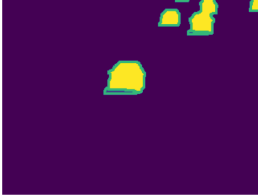
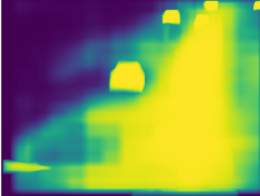

Selon le tableau I, nous pouvons voir que la méthode par segmentation d'instances correspond à environ 96% du ground truth et que celle par soustraction d'arrière-plan correspond à 97%. Ainsi, cette dernière méthode a donc été plus précise.

Graphique I : Pourcentage de comparaison de la détection de région d'intérêt pour le dataset Highway



Nous pouvons remarquer qu'il y a beaucoup de variations dans la méthode de segmentation d'instances alors que celle par soustraction d'arrière-plan est plutôt stable. Cependant, on remarque que la méthode par segmentation d'instances semble souvent être plus précise lorsqu'il n'y a pas de bruit.

Tableau II : Exemple de résultats pour la base de données Highway de l'image 1071

Image réelle	Ground Truth	Segmentation d'instances	Soustraction d'arrière-plan
			

Les images se trouvant dans le tableau II sont celles qui seront utilisées pour évaluer notre hypothèse 3.

5.2. Résultats pour le dataset Office

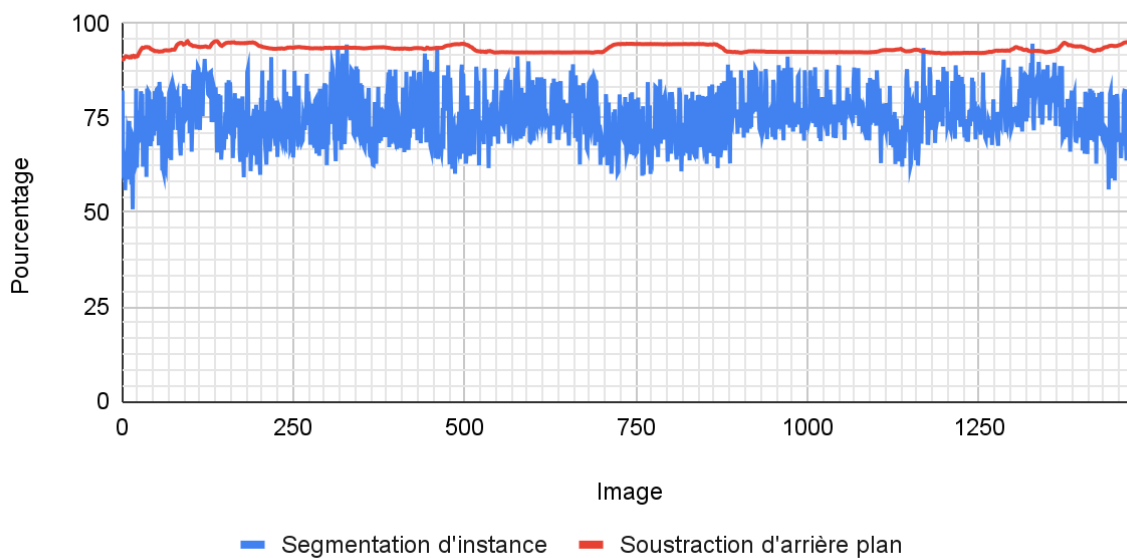
Du côté de notre base de données de Pedestrians, on peut remarquer une bien plus grande différence de résultat.

Tableau III : Résultat global de précisions pour la base de données Office

Office		
Résultat	Segmentation d'instances	Soustraction d'arrière-plan
Moyenne	75,25%	93,26%

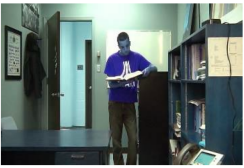

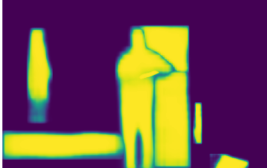

Comme le tableau III nous l'indique, la méthode par segmentation d'instances obtient une précision d'environ 75% alors que la méthode par soustraction d'arrière-plan obtient un score d'environ 93%.

Graphique II : Pourcentage de comparaison de la détection de région d'intérêt pour le dataset Office



À partir des résultats du graphique et de l'exemple de résultat ci-dessous, nous pouvons bien remarquer que le nombre d'objets que la méthode par segmentation d'instances identifie varie grandement et fait ainsi énormément varier nos résultats.

Tableau IV : Exemple de résultats pour la base de données Office de l'image 1220

Image réelle	Ground Truth	Segmentation d'instances	Soustraction d'arrière-plan
			

Les images se trouvant dans le tableau IV sont celles qui seront utilisées pour évaluer notre hypothèse 1.

5.3. Résultats pour le dataset Pedestrians

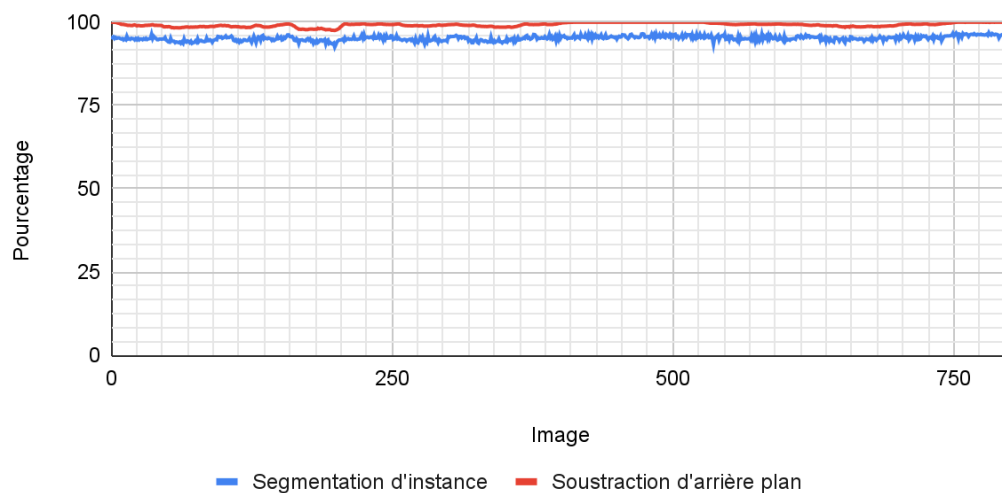
En ce qui concerne la base de données de Pedestrians, on peut y retrouver une bien meilleure précision pour nos deux algorithmes.

Tableau V : Résultat global de précisions pour la base de donnée Pedestrians

Pedestrians		
Résultat	Segmentation d'instances	Soustraction d'arrière-plan
Moyenne	95,18%	99,22%


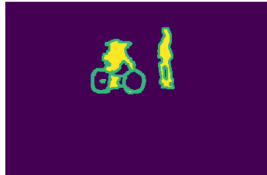
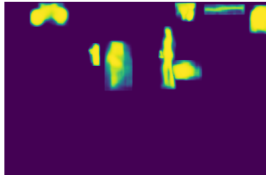
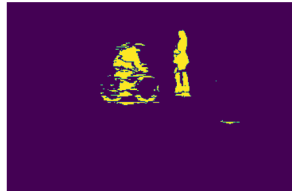
Selon le tableau V, nous pouvons voir que le résultat de la méthode par segmentation d'instances est d'environ 95% alors que celle par soustraction d'arrière-plan est d'environ 99%.

Graphique III : Pourcentage de comparaison de la détection de région d'intérêt pour le dataset Pedestrians



On remarque d'ailleurs avec le graphique III une stabilisation des résultats pour la méthode par segmentation d'instances. Pour ce qui est de la soustraction d'arrière-plan, on se retrouve majoritairement très près d'une note parfaite.

Tableau VI : Exemple de résultats pour la base de donnée Pedestrians de l'image 480

Image réelle	Ground Truth	Segmentation d'instances	Soustraction d'arrière-plan
			

Les images se trouvant dans le tableau VI sont celles qui seront utilisées pour évaluer notre hypothèse 4.

5.4. Résultats pour le dataset PETS2006

Pour ce qui est des résultats de la base de données de PETS2006, on retrouve une faible précision du modèle par segmentation d'instances et une très forte précision pour la méthode par soustraction d'arrière-plan.

Tableau VII : Résultat global de précisions pour la base de données PETS2006

PETS2006		
Résultat	Segmentation d'instances	Soustraction d'arrière-plan
Moyenne	77,83%	99,16%

Comme le tableau VII l'indique, on peut constater une précision d'environ 78% pour la méthode par segmentation d'instances et d'environ 99% pour la méthode par soustraction d'arrière-plan.

Graphique IV : Pourcentage de comparaison de la détection de région d'intérêt pour le dataset PETS2006

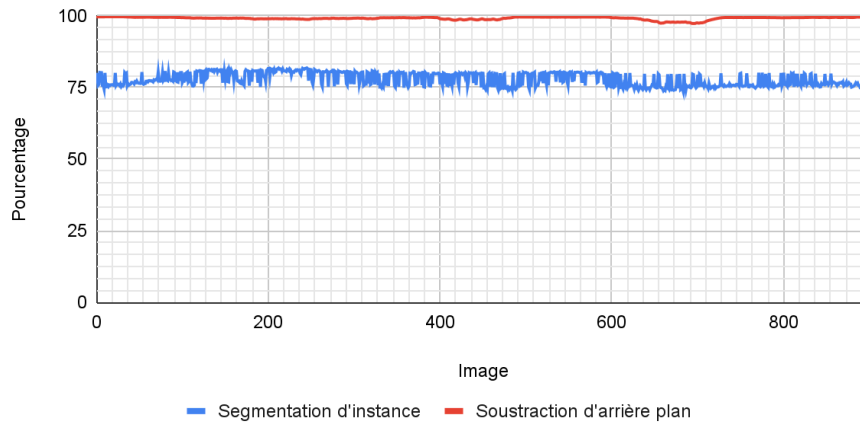


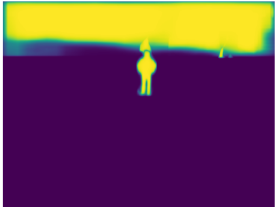
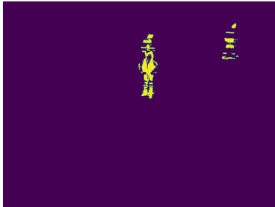


Tableau VIII : Exemple de résultats pour la base de données PETS2006 de l'image 602

Image réelle	Ground Truth	Segmentation d'instances	Soustraction d'arrière-plan
			

Les images se trouvant dans le tableau VIII sont celles qui seront utilisées pour évaluer notre hypothèse 2.

6. Discussion des résultats et retour sur les hypothèses

De manière générale, nous pouvons remarquer que la méthode par segmentation d'instances identifie souvent des objets comme étant des régions d'intérêt, alors que le ground truth n'en tenait pas compte. C'est par exemple le cas dans la trame 1220 de la vidéo Office où la segmentation d'instances reconnaît le manteau accroché, la table, le téléphone, le tableau et le meuble comme étant des régions d'intérêt en plus de la personne comme le montre le tableau IV. Cependant, le ground truth montre qu'il ne fallait que détecter la personne. La segmentation par soustraction d'arrière-plan ne détecte pas ces objets en plus, ce qui explique pourquoi elle est la méthode qui donne de meilleurs résultats pour toutes les vidéos en termes de pourcentage de ressemblance avec le ground truth.

Pour l'hypothèse 1, nous pouvons remarquer que dans la trame 1220 de la vidéo Office (tableau IV), la segmentation par soustraction d'arrière-plan ne détecte que les contours de la personne, mais l'intérieur de la forme détectée est principalement vide (donc non catégorisée comme région d'intérêt). Cela s'explique par le

fait que la personne reste longtemps presque immobile à lire un livre et puisque le modèle d'arrière-plan se base sur plusieurs trames de la vidéo à intervalle régulier, plusieurs trames du modèle d'arrière-plan contiennent la personne qui ne bouge presque pas. Cela fait en sorte que la personne est considérée comme faisant partie de l'arrière-plan et n'est pas aussi bien détectée que dans la segmentation d'instances. Cette hypothèse est donc confirmée.

Pour l'hypothèse 2, nous pouvons voir dans le tableau VIII que la segmentation d'instances arrive à différencier les deux individus même si un de ceux-ci occulte l'autre. En effet, nous pouvons apercevoir un trait vert entre les deux personnes indiquant qu'il s'agit de deux régions d'intérêt différentes. Par contre, la segmentation par soustraction d'arrière-plan n'arrive pas à faire cette différence. Donc, notre hypothèse 2 est confirmée, la segmentation d'instances est meilleure dans les cas où on souhaite différencier des régions d'intérêt qui se chevauchent.

Pour l'hypothèse 3, nous pouvons identifier les multiples détections de feuilles dans la méthode de soustraction d'arrière-plan pour l'image 1071 de la vidéo de Highway. Comme les feuilles ne sont pas un élément d'intérêt, celles-ci ne sont pas considérées dans le ground truth. De ce fait, le déplacement de l'arrière-plan, quoique minime, affecte cette méthode. Si nous n'utilisons que les objets voulus pour la méthode par segmentation d'instances, nous aurions une bien plus grande précision que celle par soustraction d'arrière-plan. Cependant, ceci doit absolument être analysé manuellement ou automatiquement par de l'intelligence artificielle afin de détecter les objets voulus. On confirme donc notre hypothèse, car cette validation peut être faite surtout si nous savons à l'avance la majorité des objets que nous voulons enlever de notre méthode par segmentation d'instances. Bien qu'on semble obtenir de moins bon résultat, on peut tout de même le confirmer dans le tableau II où l'on montre très précisément que la méthode détecte les voitures, mais détecte aussi d'autres objets.

Pour l'hypothèse 4, nous observons un résultat très convaincant présent dans notre tableau VI. Le modèle par segmentation d'instances détecte le vélo immobile assez similaire à celui du cycliste, mais ne détecte pas celui du cycliste en raison de sa vitesse. Du côté de la méthode par soustraction d'arrière-plan, on peut identifier très bien le vélo même s'il va à une vitesse assez élevée pour la vidéo. On peut donc facilement confirmer notre hypothèse que la méthode par soustraction d'arrière-plan capte beaucoup mieux le déplacement rapide que par l'identification des objets avec la méthode par segmentation d'instances.

Pour conclure, bien que cela ne réside pas dans nos mesures de performance, nous avons remarqué une très faible performance de vitesse d'exécution de la méthode par segmentation d'instances comparativement à celle par soustraction d'arrière-plan. De ce fait, bien que la méthode d'identification d'objet semble plus précise, celle-ci n'est pas nécessairement adéquate pour l'analyse de vidéos en temps réel.

Références

- [1] OpenCV: <https://docs.opencv.org/4.x/index.html>
- [2] Scikit-image: https://scikit-image.org/docs/stable/auto_examples/index.html
- [3] NumPy: <https://docs.scipy.org/doc/numpy/user/quickstart.html>
- [4] MathWorks: <https://www.mathworks.com/help/vision/ref/extractlbpfeatures.html>
- [5] CDNET 2012 dataset: <http://jacarini.dinf.usherbrooke.ca/dataset2012>
- [6] GitHub gabilodeau/INF6804 <https://github.com/gabilodeau/INF6804/blob/master/NonParamBGS.ipynb>