POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D'INGÉNIERIE

INF6804 – Vision par ordinateur

Hiver 2020

Rapport TP No. 1 :

Segmentation Vidéo

2035719 – Matthias RAMOS

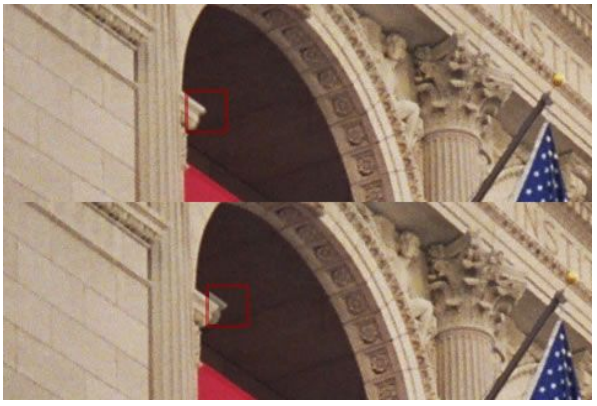
2035967 – Yoann Heitz

Dans le cadre de ce premier travail pratique nous avons décidé de comparer la méthode du flot optique et Mask R-CNN pour la séparation de zones d'avant plan/ arrière plan dans une vidéo.

I) Présentation des deux méthodes :

1) Flot optique :

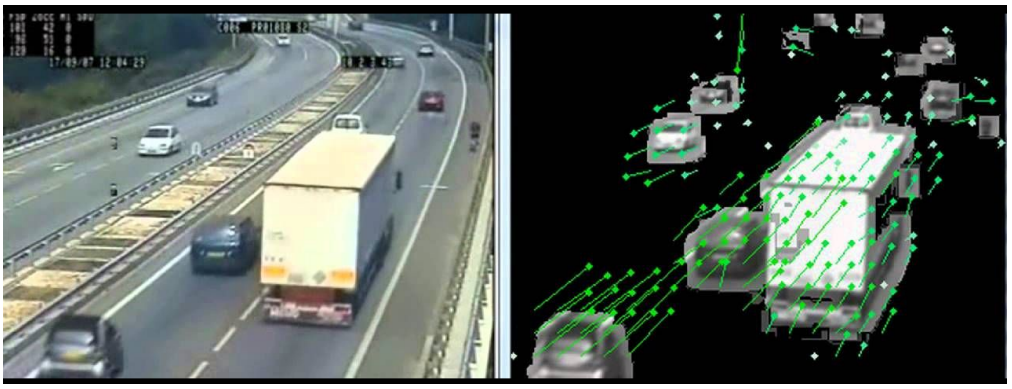
La méthode du flot optique est une méthode de détection de mouvements. La région d'intérêt ou avant plan doit alors être une partie de la vidéo qui est en mouvement par rapport à l'arrière plan qui serait fixe idéalement. Cette méthode a pour objectif la construction d'une image représentant le déplacement de certains pixel. Pour le flot optique dense, tous les pixels sont pris en compte et un vecteur de déplacement est donné pour chaque pixel sur l'image finale. Le flot optique creux quant à lui ne calcule que le déplacement de certains pixels pouvant être renseignés à l'avance. Dans le cadre de ce travail nous avons choisi d'utiliser la méthode du flot optique creux. Les pixels pour lesquels le déplacement est calculé sont alors des pixels situés sur des coins ou contours dans l'image. La détection de ces points d'intérêts utilise la méthode du minimum des valeurs propres (méthode de Harris) : une fenêtre glissante parcours l'image. Une matrice représentant les changements d'intensité selon les axes au niveau de cette matrice est construite. Cette matrice possède 2 valeurs propres : si elles sont toutes les 2 grandes alors il y a un coin au niveau de la matrice (le changement d'intensité est fort dans 2 directions au niveau d'un coin).



Détection de points d'intérêts avec la méthode de Harris

Une fois les coins/contours localisés sur la première image, l'algorithme du flot optique va déterminer leur déplacement lors du passage à la seconde image. Cela passe par la résolution d'une équation liant les dérivées de l'intensité des pixels d'intérêts en X, en Y et dans le temps (qui peuvent être déterminés lisant les intensités des pixels sur les images) et les vitesses de ces pixels selon X et Y (qui sont nos inconnues). On doit alors résoudre une équation à 2 inconnues. La méthode de Lucas-Kanade est alors utilisée. Cette méthode suppose que des pixels voisins ont le même déplacement. On obtient alors une système de plusieurs équations à 2 inconnues qui peut alors être résolu. Les vitesses en X et Y pour nos pixels sont alors déterminées : un champ vectoriel représentant les déplacements de chaque pixel d'intérêt peut donc être construit.

Le champ construit, il reste encore le problème de la segmentation à résoudre. pour cela nous avons décidé de former des enveloppes convexes pour chaque zone de l'image où l'on retrouve des champs de déplacement similaires et proches. On obtient alors des ensembles connexes de pixels qui sont donc nos segments finaux. Ces segments sont nos régions d'intérêts qui sont en mouvement d'une image à l'autre.



Resultats d'un flot optique appliqué à une capture vidéo

2) Mask R-CNN :

Mask R-CNN est une extension de Fast R-CNN qui permet non seulement une détection et classification d'objets dans une image, mais aussi la segmentation des objets détectés dans l'image (composante supplémentaire par rapport à Fast R-CNN).

Fonctionnement de Fast R-CNN:
Fast R-CNN fonctionne en plusieurs étapes. Tout d'abord, l'image à analyser passe dans un réseau neuronal convolutif (CNN). Chaque neurone du réseau applique un filtre de convolution à une partie de l'image. Sur une couche donnée du réseau, tous les neurones appliquent le même filtre aux zones qui leur sont attribuées. Cette phase permet de faire ressortir des caractéristiques précises de l'image selon les filtres qui sont utilisés. Un filtre de convolution est un filtre qui à partir d'un ensemble de pixels créé un nouveau pixel en fonctions de poids attribués à chacun des pixels en entrée. A la sortie d'une convolution sur une image, la taille de l'image n'est pas forcément conservée. Dans un réseau neuronal convolutif la taille de l'image sera très souvent réduite puis restaurée par la suite. A la sortie de ce réseau neuronal convolutif, des cartes de caractéristiques sont obtenues.
Ces cartes de caractéristiques sont ensuite traitées (parfois elles passent par une phase de pooling pour réduire leurs dimensions ou nombre) par un réseau neuronal de proposition de région (RPN). Le but de ce réseau est d'obtenir des régions d'intérêt (bounding boxes) dans l'image de départ qui sont susceptibles de contenir des objets d'intérêt. Cependant à ce stade on sait que des objets sont probablement présents dans nos régions d'intérêts mais nous ne savons pas encore à quelle position ni de quels objets il s'agit. Beaucoup de régions d'intérêt peuvent être générées mais seules quelques une seront sélectionnées quand ces régions se superposent. Pour les sélectionner le rapport entre l'intersection de certaines de ces régions et de leur union sera étudié et seules les régions avec un grand rapport IoU (intersection on union) seront gardées. Un autre réseau de neurone va étudier ces régions d'intérêt pour diminuer les tailles des bounding boxes afin qu'elles encadrent au mieux l'objet détecté et proposer une classe pour l'objet présent.

A ce stade du traitement nous avons alors des bounding boxes et une classe ainsi que sa probabilité pour chaque objet détecté sur l'image. Mask R-CNN effectue alors un dernier traitement sur chacune des bounding boxes afin de produire un masque sur l'objet détecté. Ce traitement est une segmentation de haut niveau (semantic segmentation) : pour chaque bounding box une seule instance d'un objet est présente et sa classe est connue, il faut donc faire une classification binaire pour déterminer quels pixels appartiennent à cet objet et quels pixels font partie de l'arrière plan. Pour cette dernière étape une autre réseau neuronal convolutif peut être utilisé.

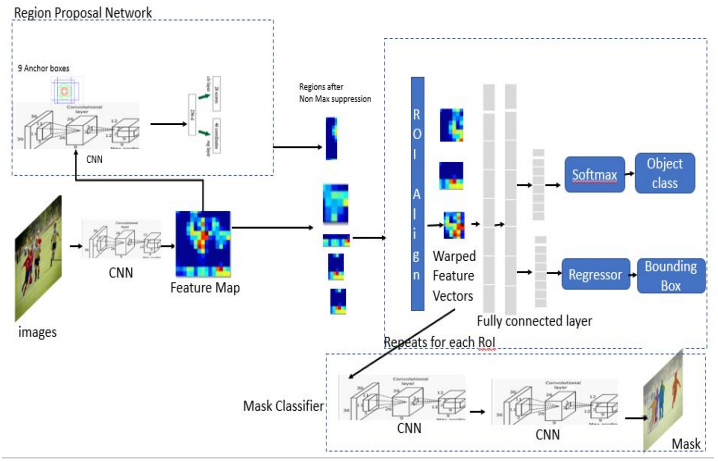


Schéma d'un algorithme Mask R-CNN utilisant 2 réseaux convolutifs pour la génération des masques

Résultats possibles après traitement par Mask R-CNN

II) Hypothèse de performances

1) Flot optique

Grâce à une pré-sélection des zones d'intérêt (par les coins), on s'attend à ce que la méthode du flot optique soit plus robuste au bruit, et non perturbé par les images avec beaucoup de couleurs. Cette méthode est plus générique que les méthodes pré entraînées, peut détecter des objets inconnus, et il n'est pas nécessaire d'avoir une base de données au préalable pour l'utiliser, ce qui représente un gain en ressources : temps de traitement, mémoire.

En revanche, plusieurs inconvénients apparaissent avec une méthode aussi naïve. Déjà le flot optique se révèle inutilisable avec une caméra trop mobile. Ensuite des problèmes apparaissent lorsqu'il y a beaucoup d'objets en mouvements et qu'ils se couvrent les uns par rapport aux autres (camion qui roule à la même vitesse qu'une voiture juste devant lui par exemple), ou des objets avec un mouvement trop rapide (cette méthode suppose que d'une image à l'autre un pixel reste dans son voisinage) par rapport à la fréquence d'acquisition (fps), ou encore si les pixels changent de couleur au cours de leur mouvement (ce qui peut arriver lorsqu'il y a des changements de luminosité mais pas de mouvement). Aussi les images uniformes avec trop peu de coins détectables peuvent poser problème pour cibler l'avant plan (beaucoup de pixels auront la même intensité et seront confondus entre eux lors de l'étude de leur mouvement). Enfin, contrairement aux autres méthodes, le flot optique ne peut détecter les objets en avant plan que s'ils sont en mouvement (cette méthode ne peut être utilisée que pour des vidéos et non pas pour des images seules et sera inefficace dans beaucoup de cas : voiture qui s'arrête dans le cadre d'une surveillance du trafic routier par exemple).

2) Mask R-CNN

Mask R-CNN fait intervenir plusieurs réseaux de neurones et nécessite un entraînement sur une base de données. Cet algorithme devrait donc être capable de détecter efficacement et de produire des masques avec une grande précision sur les objets voulus dans des situations voulues lorsqu'il a été entraîné pour ça (par exemple si la base de données contient des vidéos par mauvais temps l'algorithme sera capable de détecter les objets voulus sur des vidéos dont la qualité peut être amoindrie par la météo). Par ailleurs Mask R-CNN effectue plusieurs convolutions successives. Cet algorithme a donc tendance à transformer les données en métadonnées. Des caractéristiques autres que la couleur seront donc extraites de l'image et l'algorithme sera donc plus robuste aux changements de couleurs par exemple.

Cependant entraîner l'algorithme sur une base de données nécessite d'avoir accès à ces données et demande du temps de calcule ainsi que de la mémoire. Par ailleurs la génération d'un masque avec une réseau de convolution reste un procédé gourmand et donc souvent lent, cet algorithme ne sera donc sûrement pas utilisable pour traiter des vidéos en temps réel. Aussi l'algorithme ne saura pas s'adapter à des cas inconnus alors que le flot optique n'a pas à connaître les caractéristiques d'un objet en mouvement pour le détecter.

3) Cas d'utilisation

- Vidéo avec une faible fréquence d'acquisition : On s'attend alors comme on l'a dit à avoir un flot optique moins efficace car incapable de faire la relation entre les pixels ayant eu un mouvement trop ample entre deux images. Tandis que le Mask R-CNN fonctionnera sans aucune différence puisque son analyse se fait image par image.
  Mask R-CNN sera donc plus efficace.

- Mauvais temps : Avec un mauvais temps, les perturbation des images risquent d'affecter les deux méthodes. Pour une vidéo avec des averses de neige par exemple, une texture et/ou couleur "uniforme" peut s'établir sur l'image, limitant la détection de coins. Dans ce cas la technique du flot optique sera beaucoup moins efficaces voir inutile, alors que Mask R-CNN peut être plus robuste si il a été entraîné sur une base de données adaptée.
  Mask R-CNN sera donc plus efficace

- Forte présence d'ombres : le flot risque de détecter les ombres si leur contraste avec l'arrière plan éclairé est assez fort (puisque les ombres se déplaceront avec l'objet cible), alors que Mask R-CNN sera très sûrement entraîné à les ignorer. Cependant la présence de fortes ombres implique un fort contraste et donc une bonne détection de contours/coins. Le flot optique détectera donc bien les objets en mouvements bien qu'il détectera aussi leur ombre. Les 2 méthodes vont détecter les mouvements. Cependant Mask R-CNN filtrera mieux les ombres.

- Vidéos de nuit : dans ce type de vidéos, les forts changements d'intensité lumineuse (entre un lampadaire ou un phare de voiture et le ciel noir par exemple) auront tendance à étaler les sources de lumière. Si on prend l'exemple d'une voiture qui fait face à la caméra, son phare va recouvrir une partie de la voiture par une sphère lumineuse. Cette sphère (de grande intensité par rapport au fond de l'image) se déplacera à la même vitesse que la voiture et la voiture sera donc détectée par le flot optique. Mask R-CNN quant à lui risque de ne pas détecter la voiture recouverte de lumière. Le flot optique sera plus efficace

- Caméra mobile : L'efficacité de Mask R-CNN ne sera pas du tout affectée puisque les images sont traitées les unes indépendamment des autres. Le flot optique détectera cependant un mouvement global sur les images. Si la caméra n'effectue que des translations et pas de rotations la différence d'amplitude de mouvement entre le fond et les régions d'intérêts resteront les mêmes (et donc les régions d'intérêt pourront être détectées). Cependant si la caméra effectue des rotations alors les extrémités de l'image se déplaceront plus rapidement que le centre et des différences de vitesses seront détectées. Le traitement sera alors affecté. Mask R-CNN sera plus efficace

III) Expériences

1) Evaluation des performances

Nous utiliserons la base de données CDNET qui rassemble 11 catégories de vidéos qui rassemblent chacune 4 à 6 videos. Parmis ces catégories se trouvent celles que nous avons choisi d'utiliser pour comparer les deux méthodes. Pour chaque vidéo nous trouvons aussi une version "groundtruth" qui représente la segmentation de chaque image avec l'arrière plan en noir et l'avant plan en blanc. Ces vidéos nous permettront d'évaluer nos résultats.

Image de la séquence bad weather : blizzard et segmentation correspondante dans le dossier groundtruth

Une première vérification "à l'oeil" sera alors réalisée pour d'abord tester la bonne configuration des implémentations, et avoir une première idée des résultats de performance. Ensuite nous procéderons à une étude plus poussée grâce à l'utilisation de plusieures métriques de similitude avec les images du "groundtruth".

Pour la précision on compare le nombre de pixels en communs de notre méthode avec les pixels de groundtruth. Pour chaque pixel on mesure la différence d'intensité avec son intensité attendue :

$$\Delta_i = | I_i - I_{i(gt)} |/255 \quad \text{(0 si le pixel est idem au "groundtruth", 1 sinon)}$$

$$\Delta_{tot} = 1/N \sum_{i=0}^{N} \Delta_i$$

De même nous évaluons le recall pour chaque vidéo, en effet, le nombre de pixels contenu dans la région d'intérêt est faible par rapport aux tailles des images (moins de 10% en général). Produire des masques ne contenant aucun pixels mènerait alors une précision de plus de 90% dans ces cas avec la métrique précédente. Le rappel permet alors de mesurer l'exhaustivité de nos résultats. Cette mesure correspond au nombre de pixels classifiés comme étant dans le masque et y étant vraiment sur le nombre total de pixels réellement dans le masque

En classant tous les pixels dans le masque on aurait un recall de 1, on a donc décidé de rajouter une dernière métrique discriminante pour la comparaison des masques : l'indice de Jaccard ou IoU (Intersection on Union). Cette mesure est le rapport entre le nombre de pixels appartenant à l'intersection du vrai masque et du masque prédit et le nombre de pixel dans la réunion des 2 masques.

Le temps de traitement moyen pour chaque image est aussi pris en compte dans l'évaluation des performances de chaque méthode.

Aussi étant donné l'implémentation des algorithmes, il est à prévoir que Mask R-CNN devrait obtenir des meilleurs résultats (au prix d'un plus grand temps de calcul). Pour évaluer les variations de performances nous étudierons l'écart entre Mask R-CNN et le flot optique divisé par la moyenne pour chacune des métriques.

2) Implémentations

Pour le flot optique nous nous sommes inspirés des notebooks laissés sur le moodle du cours. Ainsi nous avons implémenté d'abord la détection de pixel d'intérêt grâce aux mouvements des coins detecté à l'aide de l'API OpenCV. Ensuite, pour chaque image, afin de construire un masque depuis les coordonnées des pixels retenus (nous ne gardons que les pixels qui bougent suffisamment d'une image à l'autre) nous utilisons la librairie scipy.spatial et plus particulièrement la classe Delaunay qui nous permet de construire l'enveloppe convexe des points retenus. Avant de construire cette enveloppe convexe nous appliquons d'abord un algorithme DBScan implémenté par nous même sur les points retenus : cela permet en effet de les regrouper en clusters avant de récupérer l'enveloppe convexe de chacun de ces clusters (sans ce traitement on n'aurait qu'une seule grande enveloppe convexe si 2 régions d'intérêt isolées sont présentes d'une part et d'autre de l'image. Une fois les enveloppes convexes de chaque région d'intérêt récupérées nous récupérons les coordonnées de tous les points contenus dans ces enveloppes convexes.

Nous comparons ensuite les points obtenus avec les points des masques du dossier groundtruth (qui ont été prétraités afin d'enlever les pixels des ombres) selon les métriques définies. Les paramètres principaux qui varient pour cette algorithme sont les nombres de coins détectés par l'algorithme de détection de OpenCV (ainsi que les seuils de détections de ces coins) et le seuil d'appartenance à un cluster dans l'algorithme DBScan. Pour la détection de coins nous avons gardé les paramètres de base trouvés sur le moodle. Nous les avons testé en les faisant varier mais les coins détectés étaient de moins bonne qualité qu'auparavant. Nous les avons donc laissé inchangés. Pour le seuil d'appartenance à un cluster nous l'avons fait "à l'oeil". Si ce seuil est trop petit, une personne/voiture dans une vidéo sera coupée en plusieurs morceaux. Si il est trop grand, 2 personnes/voitures dans une vidéo seront regroupé en une enveloppe convexe. Nous avons donc testé des cas de bases en faisant varier ce seuil pour obtenir des résultats convenables (les objets ne sont pas découpés mais 2 objets très proches peuvent être regroupés en une enveloppe convexe).

Le modèle Mask RCNN utilisé provient de l'API Coco, nous avons réutilisé et modifier le sample du repo https://github.com/matterport/Mask_RCNN/blob/master/samples/demo.ipynb qui permet d'instancier un modèle déjà préentrainé et testé dont l'efficacité est indéniable. D'abord les paramètres ont été laissés tels qu'on les reçoit par défaut, puis nous avons remarqué que pour certaines vidéo, le modèle détectait en premier plan des objets que le groundtruth souhaite laisser en arrière plan. Nous avons alors retouché le modèle pour chaque vidéo afin de gagner en performance.

3) Séquences vidéos sélectionnées

Nous avons sélectionné une vidéo pour chacun des cas d'utilisation spécifiés précédemment ainsi qu'une vidéo dans des conditions normales ("highway") pour obtenir des mesures des base. Pour la vidéo par mauvais temps nous avons sélectionné la vidéo "blizzard". Cette vidéo montre une route à double sens sur laquelle se déplacent des voitures pendant une tempête de neige. A cause des fortes chutes de neiges, le contraste entre les voitures et le fond (blanc de neige) est diminué. Cela devrait amoindrir les performances de l'algorithme de détection de coins et rajouter de la difficulté au suivi de pixels dans l'algorithme de flot optique (les chutes de neiges uniformisent la couleur et la texture sur les images)

Pour la vidéo avec faible fréquence d'acquisition nous avons choisi la séquence "turnpike_0_5fps" qui zoom de profil sur une autoroute. Etant donné la vitesse des voitures, le zoom sur la route et la faible fréquence d'images (0.5 par seconde) les voitures se déplacent d'une distance considérable d'une image à une autre, tout en étant très visibles. Cette vidéo permettra de tester le suivi de points dans le flot optique.

Pour la vidéo avec ombres nous avons sélectionné la séquence "busStation" qui montre des personnes à une station de bus très éclairée par le soleil. Leurs ombres sont fortement projetées et ont peut y voir la forme des personnes très nettement. Ainsi on pourra regarder à quel point le flot optique mesure le déplacement des ombres et voir si Mask R-CNN détecte des personnes au niveau des ombres (qui ont des formes humaines).

Pour la vidéo de nuit nous avons choisi la vidéo "fluidHighway" qui montre une autoroute de face la nuit. Les phares des voitures éblouissent la caméra. On pourra donc vérifier si Mask R CNN arrive toujours à détecter des voitures et si le flot optique reste robuste.

Pour la vidéo avec caméra mobile nous avons choisi la vidéo "traffic". Cette vidéo montre des voitures sur une route avec un caméra qui bouge. La vidéo est nette. Sans le mouvement de caméra cette vidéo ne poserait aucune difficulté.

Nous n'avons sélectionné des vidéos qu'avec des voitures/camions et humains puisque Mask R-CNN a été réduit à détecter ce type d'objet, sans cela il détecte aussi tout type d'objet qui se trouvent en arrière plan, ce qui peut biaiser significativement ses résultats de performance.

IV) Résultats

1) Présentation

Pour chaque vidéo, nous générons pour chaque image un masque qui sera comparé avec le groundtruth selon chacune des métriques décrites précédemment. Les résultats alors présentés sont les moyennes de ces métriques par catégorie de vidéo. Les tests effectués sur les vidéos de baseline nous permettent d'évaluer nos modèles sur des images ne présentant pas de difficulté particulière afin de pouvoir comparer les performances avec les autres catégories.

Baseline:

	Accuracy (%)	Recal (%)	IoU	Temps de traitement par image (µs)
Flot optique	89.36	35.53	0.30	60.4
Mask RCNN brut	94,13	71.17	0.63	327
Ecart/moyenne	0,052	0.67	0.71	

Low framerate:

	Accuracy (%)	Recal (%)	IoU	Temps de traitement par image (µs)
Flot optique	96.41	35,57	0.123	27.9
Mask RCNN brut	96.53	41.71	0.3	330
Ecart/moyenne	0.0012	0.16	0.84	

Bad weather:

	Accuracy (%)	Recal (%)	IoU	Temps de traitement par image (µs)
Flot optique	96.86	7.3	0.278	135
Mask RCNN brut	99.32	82.24	0.789	324
Ecart/moyenne	0.025	1.67	0.96	

Shadow:

	Accuracy (%)	Recal (%)	IoU	Temps de traitement par image (µs)
Flot optique	96,7	62.2	0.445	24.8
Mask RCNN brut	98,9	83,2	0.785	416
Ecart/moyenne	0,041	0,289	0,55	

Night:

	Accuracy (%)	Recal (%)	IoU	Temps de traitement par image (µs)
Flot optique	99,34	34,18	0,28	45.5
Mask RCNN brut	97,69	39.17	0.35	327
Ecart/moyenne	0.017	0.136	0.22	

CameraJitter:

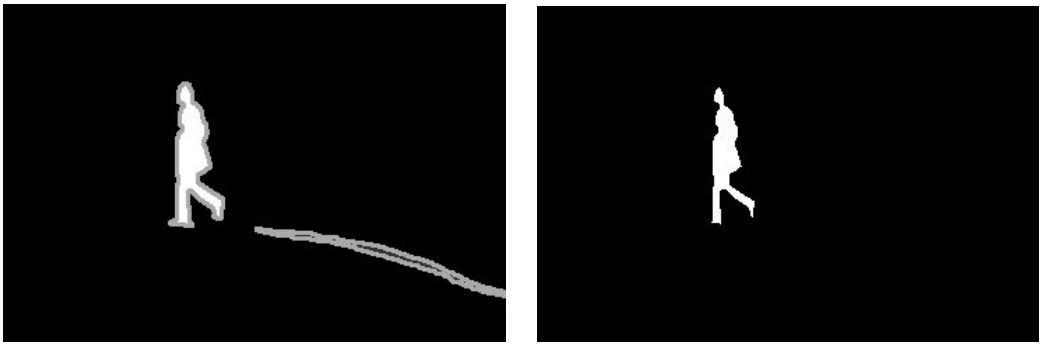
	Accuracy (%)	Recal (%)	IoU	Temps de traitement (µs)
Flot optique	77,65	8,59	0,07	108
Mask RCNN brut	93,29	45,52	0.44	327
Ecart/moyenne	0,183	1,36	1,45	

2) Analyse

D'après ces résultats, on note très vite l'efficacité de détection souvent bien supérieure des réseaux RCNN par rapport à une méthode naïve telle que le flot optique. Ce défaut du flot optique creux provient aussi très probablement de l'imprécision de la reconstitutions des masques depuis les coins mobiles bien que le DBScan soit assez satisfaisant. Cette imprécision provient du peu de coins générés. Néanmoins cette méthode reste très rapide (malgré l'implémentation non optimale de DBScan) mais génère des résultats "grossiers" dans la majorité des cas.

Détection d'avant plan (origine, vecteurs, masques) par flot optique

Plus en détail, on remarque que les performances de ces méthodes sur les différents cas d'utilisation diffèrent légèrement de ce qui a été prédit dans nos hypothèses précédentes. D'abord sur les présences d'ombres, le flot optique est moins affecté que ce que l'on aurait pu penser et conserve tout de même une bonne précision. Une explication, en analysant de plus près les masques généré, serait que les coins sont moins souvent détectés sur les ombres que sur les objets d'avant plans et donc ne viennent pas interférer dans les résultats.



Transformation des images groundtruth en images binaires pour supprimer les ombres

Le flot optique perd beaucoup moins en performances que Mask R-CNN sur la vidéo de nuit. C'est par ailleurs sur cette vidéo que Mask R-CNN a les moins bonnes performances (scores de recall et IoU qui pointent l'imprécision des masques) alors que le flot optique reste stable par rapport à la vidéo baseline. Ceci vient donc confirmer nos hypothèses.

Sur la vidéo avec basse fréquence d'acquisition, le flot optique reste constant en performances sur le recall alors que Mask R-CNN décroît dans toutes les métriques. Cela peut provenir d'un mauvais choix de vidéo de test pour la vidéo de base ou celle avec un faible taux d'acquisition. Cependant il est a noté que l'IoU du flot optique est très bas. Cela provient du fait que le flot optique a produit des masques beaucoup plus grands que ceux ciblés (en comparaison avec le recall). Une meilleure paramétrisation du DBScan pourrait régler ce problème.

Pour la vidéo en mauvaises conditions météos et la vidéo avec caméra mobile, les performances du flot optique sont fortement amoindries comme ce qui avait été prévu. Mask R-CNN perd en performances sur la vidéo avec caméra mobile mais reste très robuste sur la vidéo en mauvaises conditions météos (scores de recall et IoU élevés). Pour la caméra mobile cependant,

avec une meilleure segmentation les capacités du flot optique pourraient être améliorées si on ne prend en compte que les pixels dont le déplacement a dépassé un certain seuil (à déterminer).

Détection d'avant plan via Mask RCNN

Enfin les temps d'exécutions sont assez constants pour les 2 méthodes, le flot optique peut parfois voir son temps d'exécution augmenter sur des vidéos avec une forte mobilité à analyser comme Bad weather ou Camera Jitter. Le flot optique reste plus rapide (d'un facteur 10 à 20) que Mask RCNN ce qui paraît tout à fait logique au vue de la détection plus poussée faite par les réseaux de convolution. Le temps d'analyse de cette dernière méthode est, cela dit, tout à fait acceptable au vue de sa précision.

3) Conclusions

Après cette première étude, nous pouvons conclure que le flot optique bien que plus adapté pour effectuer des analyse rapide sur certains types de vidéos, reste bien moins efficace que des méthodes plus sophistiquées comme celle du Mask RCNN. Et en prenant en compte le rapport efficacité/temps d'exécution les réseaux de convolution reste bien plus attrayant pour la détection d'avant plan que le flot optique. Cependant seul le flot optique creux a été étudié dans ce travail pratique. Ses performances sont assez limitées par la qualité de la détection des coins. Avec un flot optique dense on devrait pouvoir augmenter les performances de prédiction de masques (puisque tous les points seraient étudiés) mais il faudrait aussi implémenter un algorithme de segmentation efficace pour traiter le champs de vecteurs obtenus. Cependant cette méthode serait plus gourmande en calculs et en temps consommé (par ailleurs la vitesse de calcul de DBScan pourrait aussi être améliorée en utilisant des librairies adaptées). Aussi, les mesures pourraient être affinées en cherchant les meilleurs paramètres de DBScan et de détection de coins pour chaque vidéo pour le flot optique. Cela nécessiterait néanmoins de faire plus de tests sur le dataset (qui pourrait aussi être élargi) ce qui peut prendre un temps considérable pour de longues vidéos. Aussi, les performances des modèles ont été évaluées sur la génération de masque (qui est coûteuse en temps). Les modèles pourraient aussi être évalués sur la génération de bounding boxes où le flot optique creux pourrait gagner en performance relativement à Mask R-CNN (qui ne serait plus très adapté puisque la génération d'un masque propre à Mask R-CNN serait une étape supplémentaire non nécessaire).