Localisation de mouvements saillants dans des cartes de flot optique par l'interprétation d'un réseau de classification

Etienne Meunier, Patrick Bouthemy

Localisation de mouvements saillants dans des cartes de flot optique par l'interprétation d'un réseau de classification

Étienne Meunier[1] Patrick Bouthemy[1]

[1] Inria, Centre Rennes - Bretagne Atlantique, France

etienne.meunier@inria.fr

Résumé

Cet article s'intéresse à la localisation des mouvements saillants dans les images successives d'une séquence vidéo. Un mouvement saillant est un mouvement se démarquant de son contexte environnant. Notre méthode s'appuie sur l'interprétation, pour chaque instant de la séquence, d'un réseau convolutionnel de classification dont l'entrée est constituée du flot optique. Cette classification porte sur la présence ou non de mouvements saillants dans l'image. En combinant la carte d'interprétation déduite du réseau et une segmentation du flot optique, nous pouvons détourer précisément les mouvements saillants dans l'image et estimer leur degré de saillance. Un atout important de notre méthode est qu'elle ne nécessite pas de cartes de segmentation annotées pour l'entraînement du réseau. Les résultats expérimentaux et la comparaison avec les méthodes existantes démontrent la performance de notre méthode sur une variété de vidéos.

Mots Clef

Saillance de mouvement ; Apprentissage non supervisé ; CNN ; cartes d'attention.

Abstract

This paper is concerned with the segmentation of salient motions in a video sequence. We have defined a novel method based on the optical flow only. We leverage a frame-based classification network that predicts the presence of salient independent motions in a video frame. By combining the interpretation map inferred from the network and the segmentation of the optical flow, we can locate the salient motions in the frame and estimate their saliency degree. Experimental results on real videos and comparison with existing methods demonstrate the performance of our method.

Keywords

Motion saliency ; Unsupervised learning ; CNN ; Attention maps

1 Introduction et travaux connexes

Cet article présente une nouvelle approche pour segmenter les objets aux mouvements saillants dans les images successives d'une vidéo. L'estimation de la saillance du mouvement vise à mettre en évidence les mouvements locaux qui se distinguent du contexte qui les environne, et qui sont donc susceptibles de révéler un événement significatif. La saillance du mouvement a de nombreuses applications en vision par ordinateur. Elle peut être exploitée pour la navigation de véhicules autonomes afin d'anticiper les obstacles, ou pour la sécurité dans des espaces publics accueillant des foules afin de déclencher l'alerte à l'apparition de situations critiques ou dangereuses. Elle peut faciliter l'analyse ultérieure de vidéos où le mouvement joue un rôle primordial.

La saillance vidéo (SV) est un domaine de recherche actif, mais elle mélange généralement les questions d'apparence et de mouvement. Des approches antérieures, telles que [1], ont exploité les frontières de mouvement dans l'image pour faciliter la segmentation et la reconnaissance des objets mobiles à partir d'un modèle d'apparence. Dans [2], les auteurs supposent que les objets saillants ont des caractéristiques d'apparence et de mouvement (orientation et amplitude des vecteurs du flot optique) différentes de celles de l'arrière-plan, et utilisent une mesure inspirée de Tukey pour détecter les vecteurs anormaux et les étiqueter comme appartenant à des objets en mouvement. Des approches d'apprentissage profond ont également été étudiées. Dans [3], les auteurs ont développé une méthode s'appuyant sur un réseau convolutionnel (CNN) qui combine les dimensions spatiales et temporelles de l'image sans calcul explicite du flot optique. Dans [4], des caractéristiques sont extraites à l'aide d'un réseau et sont utilisées au sein d'une stratégie de segmentation multiéchelle pour prédire la saillance dynamique dans des vidéos.

La segmentation d'objets dans des vidéos (SOV) est un sujet connexe où l'apparence et le mouvement des objets sont aussi des facteurs importants. La disponibilité de grands ensembles de données annotées, [5], rend possible l'utilisation de techniques d'apprentissage profond supervisé pour la SOV. Dans [6, 7], un réseau convolutionnel et un réseau récurrent sont entraînés conjointement pour segmenter les

objets en mouvement. Une formulation proche, [8], utilise une double supervision combinant le module de saillance avec un module de flot optique.

Dans cet article, nous nous intéressons à la segmentation des mouvements saillants (SMS) dans les images successives d'une vidéo à partir du seul flot optique. Cet objectif est proche de la SV, mais présente aussi des spécificités sur lesquelles nous reviendrons plus loin. Il est différent de la SOV, car la notion d'objet n'est pas prépondérante pour la SMS, contrairement à la SOV. De plus, la SOV se concentre le plus souvent sur la segmentation d'un objet mobile unique d'intérêt, *a* priori en avant-plan de la scène observée. La SMS peut comprendre des configurations plus diverses que la SV qui implique également la notion d'apparence dans la spécification de la saillance. Par exemple, dans le cas de la détection d'anomalies dans une foule, la saillance peut provenir d'une personne qui se déplace à contre-courant de celles qui l'entourent [9], et donc seul le mouvement importe. Un autre exemple, relatif à la biologie, est celui de particules exhibant un mouvement particulier à l'intérieur d'une cellule, ou de cellules au mouvement se différenciant de ceux d'un ensemble de cellules. On pourrait également citer des exemples liés au trafic routier. Dans ces exemples, seul le mouvement opère dans la création de la saillance dynamique. La SMS a en fait à la fois un caractère plus spécifique que la SV en ne considérant que le mouvement, et une capacité à appréhender des configurations plus générales où l'apparence est sans utilité.

Dans cet article, nous allons en fait nous focaliser sur un cas particulier, mais important et représentatif, de la SMS : les mouvements saillants dans l'image correspondant aux mouvements indépendants dans une scène observée par une caméra mobile. Cela nous permettra également de mener une comparaison expérimentale avec des méthodes existantes de SV qui attaquent des problématiques similaires. Réaliser la SMS dans ce type de configuration à partir du seul flot optique pose un problème central et difficile que nous allons spécifier ci-après.

Puisque la vidéo est acquise par une caméra mobile, tout élément statique de la scène va posséder un mouvement apparent dans les images. D'autre part, le mouvement apparent 2D mesuré par le flot optique résulte à la fois du mouvement 3D relatif entre les éléments de la scène et la caméra, et de la profondeur des objets dans la scène. Les objets indépendants mobiles de la scène vont produire des segments distinctifs dans le flot optique. Cela sera également le cas des objets statiques en avant-plan qui présentent une rupture de profondeur nette avec le reste de la scène statique formant l'arrière-plan de la scène. Nous nommerons mouvements de parallaxe les mouvements apparents distinctifs associés aux objets statiques situés au premier plan de la scène La question qui se pose est donc de pouvoir éliminer les mouvements de parallaxe dans l'établissement de la SMS à partir du flot optique.

Ce problème rejoint pour partie la segmentation d'objets mobiles indépendants dans des vidéos, dont nous ne donnons ci-dessous que quelques éléments bibliographiques récents exploitant notamment des CNN. Une solution consiste à calculer le flot de la scène statique, c'est-à-dire le mouvement apparent dans l'image, induit par le mouvement de la caméra, de l'ensemble de la scène statique, avant-plan et arrière-plan compris [10, 11]. Ensuite, les objets se déplaçant de manière indépendante peuvent être assez directement identifiés par rapport à ce flot de scène statique. Ces techniques sont robustes aux mouvements de parallaxe puisque ces derniers font partie de ce flot de scène statique. Cependant, elles impliquent que les paramètres intrinsèques de la caméra soient disponibles et qu'une estimation précise de la profondeur de tous les objets de la scène et du mouvement 3D de la caméra soit fournie. Or, notre objectif est de segmenter les mouvements saillants dans une séquence vidéo uniquement à partir du flot optique calculé et de manière non supervisée.

Le reste de l'article est organisé comme suit. Dans la section 2, nous décrivons notre méthode. Puis, nous présentons les résultats expérimentaux dans la section 3. Enfin, la section 4 contient une discussion et des remarques finales.

2 Description de la méthode

Le principe général de notre méthode de segmentation des mouvements saillants est résumé à la figure 1. Elle se déroule en deux étapes principales, la première impliquant deux processus exécutés en parallèle.

2.1 Classification de présence de saillance de mouvement dans une image d'une vidéo

Nous utilisons le réseau de classification appliqué à chaque image d'une vidéo pour prédire la présence de mouvements saillants à partir de cartes de flot optique, introduit dans [13]. Ce réseau est entraîné à prédire la présence d'un mouvement indépendant saillant dans une image. Il est donc implicitement capable d'ignorer les mouvements de parallaxe. L'apprentissage est supervisé, mais le processus d'annotation est très léger. Il se résume à l'étiquetage au niveau de l'image, et en pratique, il peut être réalisé presque automatiquement, puisque le jeu de données d'entraînement était composé de vidéos qui étaient soit entièrement saillantes dynamiquement, soit entièrement non saillantes.
Le réseau [13] prend initialement en entrée le flot résiduel déduit du flot optique dont le mouvement dominant dans l'image a été éliminé par simple soustraction. Le mouvement dominant est celui produit par le mouvement de la caméra. Il est modélisé par un mouvement paramétrique 2D (affine ou quadratique) estimé par un algorithme robuste, multi résolution et incrémental [14]. Le réseau prédit si l'image est dynamiquement saillante, c'est-à-dire si elle contient un mouvement saillant, ou non. Bien que la structure du réseau soit simple et peu profonde, comprenant trois blocs convolutionnels et une couche entièrement connectée, il a donné une précision de 87.5% sur la base de vidéos de test utilisée dans [13].

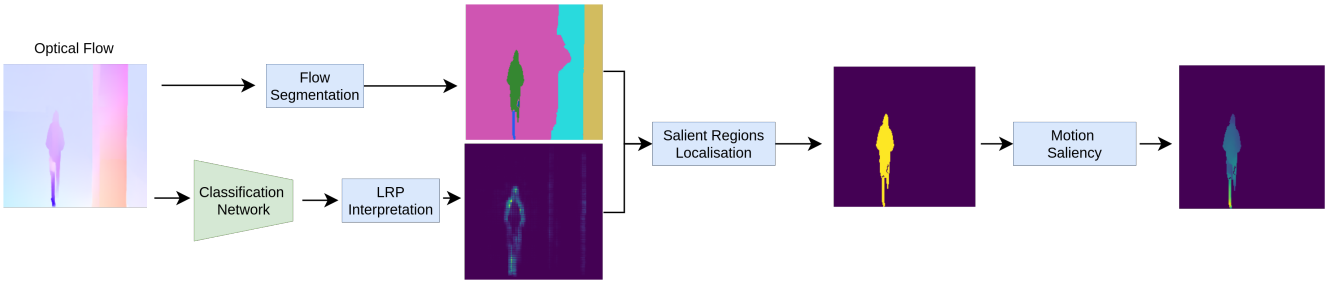

FIGURE 1 – Schéma général de notre méthode de SMS utilisant le flot optique en entrée. Le flot optique est ici affiché avec la représentation couleur HSV [12]. La branche supérieure segmente la carte de flot optique et la branche inférieure calcule la carte d'interprétation par la technique LRP. Les deux branches sont ensuite combinées pour extraire les régions du flot présentant un mouvement saillant. La procédure d'inpainting du flot est appliquée pour calculer un score de saillance de mouvement.

Nous avons modifié ce réseau de classification à plusieurs niveaux. Tout d'abord, comme nous ne voulions pas être dépendants de l'exactitude de la compensation du mouvement dominant par un seul modèle paramétrique, nous avons décidé de donner directement le flot optique en entrée du réseau au lieu du flot résiduel. Par conséquent, nous avons en partie ré-entraîné le réseau avec cette entrée modifiée en utilisant le même ensemble de vidéos d'entraînement que dans [13]. Ensuite, en suivant des recommandations sur l'application des techniques d'interprétation aux réseaux neuronaux profonds [15], nous avons apporté des modifications à l'ordre des couches internes du réseau au moment du test comme illustré dans la figure 2, puis fusionné les couches de normalisation et de convolution adjacentes. Cela a conduit à un réseau fonctionnellement identique à celui utilisé pour l'entraînement, mais capable de fournir des cartes d'interprétation moins bruitées. Le réseau modifié offre des performances de classification globalement similaires, tout en restant efficace lorsque la compensation de mouvement se serait avérée erronée, notamment dans le cas de scène statique à la structure complexe. Ce point est important pour les étapes suivantes.

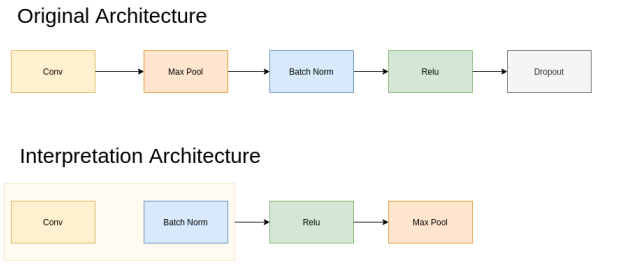

FIGURE 2 – Permutation des couches pour améliorer l'interprétabilité du réseau. En haut est représenté l'ordre des couches dans le réseau lors de l'entraînement et en bas l'ordre des couches dans le réseau utilisé pour l'interprétation.

2.2 Obtention des cartes d'interprétation

Dans [16], les auteurs établissent que le mouvement saillant ne peut pas être prédit par le mouvement environnant. Ils mettent en place une compétition entre un réseau "générateur" produisant un masque qui cache une partie du flot optique, et un réseau "impainter" qui essaye de reconstruire le flot à l'intérieur du masque à partir du contexte uniquement. Cependant, comme l'illustre plus loin la figure 4, cette méthode échoue en cas de mouvement de parallaxe, car le flot optique dépend aussi de la profondeur des objets qui est inconnue.

Nous adoptons une approche différente. Puisque le réseau de classification prédit correctement comme non saillantes les images qui ne comprennent que des mouvements de parallaxe, nous pouvons en déduire qu'il a appris à distinguer les mouvements indépendants dans la variété des mouvements apparents saillants qui peuvent être produits. Nous pouvons exploiter cette connaissance pour extraire et localiser les mouvements saillants dans l'image. Les techniques d'interprétation permettent de générer des cartes dites d'attribution à partir d'un réseau de neurones entraîné. Ces cartes indiquent quelles parties des données d'entrée contribuent réellement à la prédiction. En analyse d'image, elles ont été appliquées généralement à des réseaux dont l'entrée est l'image elle-même. Dans notre cas, l'entrée est une carte de vecteurs 2D de flot optique. Notre objectif est de déterminer quels vecteurs du flot optique donné en entrée du réseau induisent la classification de la scène comme dynamiquement saillante.

Plusieurs techniques existent pour l'interprétation des réseaux [17]. À notre connaissance, une telle interprétation de réseau n'a pas encore été étudiée pour l'analyse du flot optique. Une problématique technique importante est d'extraire des informations d'interprétation véritablement liées au fonctionnement interne du réseau entraîné, et non aux caractéristiques propres des données d'entrée ou à la structure du réseau [17]. Plusieurs travaux tels que [18] ont montré que dans certains cas, les cartes d'interprétation produites avec un modèle entraîné pouvaient être visuellement similaires à celles générées par un modèle aléatoire. Dans notre cas, nous n'avons pas observé un tel comportement.

La méthode des gradients intégrés [19] génère des cartes d'interprétation en cumulant les gradients le long d'une interpolation linéaire entre une image de référence choisie et une entrée donnée. Bien que cette technique ait donné

des résultats prometteurs sur plusieurs ensembles de données, les résultats varient considérablement en fonction de la référence choisie [20]. Dans le cas des images, l'image de référence par défaut est une image noire qui représente l'absence de caractéristiques iconiques. Cependant, en ce qui concerne le flot optique, il n'y a pas de choix évident pour représenter l'absence de mouvement saillant. Dans nos expériences, l'application de la méthode des gradients intégrés sur les champs de flot optique a principalement conduit à mettre en évidence les zones du flot avec une grande amplitude de mouvement.

Aussi, nous avons adopté la méthode "Layer-wise Relevance Propagation" (LRP) [21]. Elle fournit des cartes d'interprétation moins bruitées et supprime le besoin du choix d'une image de référence. Cette technique propage un score de pertinence de la couche de sortie à la couche d'entrée du réseau convolutionnel. Elle s'appuie sur un ensemble de règles de propagation particulières. Nous renvoyons le lecteur à [21] pour plus de détails. Nous avons suivi le schéma présenté dans [17, 22], en appliquant la règle z^+ à toutes les couches convolutionnelles et entièrement connectées, à l'exception de la première couche convolutionnelle, pour laquelle nous appliquons la règle z^β. La règle z^+ ne prend en compte que les poids positifs lors de la propagation du score d'importance. La règle z^β propage le score d'importance positif à travers les couches qui prennent en entrée des valeurs négatives, à l'aide d'un terme additif pour les valeurs extrêmes admissibles de l'espace d'entrée. Les couches max-pool sont traitées avec une stratégie de type "winner-takes-all". En ce qui nous concerne, nous n'avons pas besoin de traiter les couches de normalisation, car nous les avons fusionnées avec les couches convolutionnelles.

2.3 Segmentation du flot optique

Bien que les cartes d'attribution produites par le LRP mettent en évidence les zones de mouvement saillantes, le résultat reste fragmenté et parcellaire. De ce fait, elles ne suffisent pas à elles seules à produire des régions cohérentes et compactes de mouvement saillant. Aussi, nous effectuons en parallèle une segmentation en régions du flot optique basée sur un schéma de régression robuste et itératif.

Les travaux séminaux sur la segmentation du mouvement 2D apparent dans des images, en couches ou en régions [23, 24], prenaient deux images successives en entrée et estimaient un modèle affine de mouvement par couche ou par région. Ici, nous prenons des champs de flot optique en entrée, et nous décomposons ce flot optique en calculant itérativement un flot affine à travers une régression robuste impliquant la fonction de Huber $\mathcal{H}$. Plus précisément, nous minimisons la fonction suivante :

$$\sum_{p \in \Omega_k} \mathcal{H}(a_1 + a_2 x + a_3 y - u_p) + \mathcal{H}(a_4 + a_5 x + a_6 y - v_p),$$

où Ω_k est la partie de l'image intervenant à l'itération k, $p = (x, y)$ un pixel, $w_p = (u_p, v_p)$ son vecteur de flot optique, les a_i sont les six paramètres du flot affine.

À chaque itération, les points de l'image pour lesquels le flot optique est bien approché par le modèle de mouvement paramétrique, en pratique un modèle affine, sont sélectionnés pour former une couche. Ce sont les points de l'image où la norme L_1 de la différence entre les composantes du vecteur de flot affine et les celles du vecteur de flot optique, est inférieure à un seuil donné. Nous normalisons la norme L_1 par l'écart-type des différences afin de ne pas être affecté dans le choix du seuil par les variations d'amplitude du mouvement. Cela nous permet en fait de fixer par défaut le seuil à 1. Les points où le flot optique n'est pas correctement approché par le modèle affine estimé sont conservés pour les itérations suivantes. Nous itérons cette procédure jusqu'à ce qu'il ne reste qu'un petit nombre de points dans l'image, dont les vecteurs de flot optique ont été constamment incorrectement approchés par les modèles affines successivement introduits. Nous les regroupons alors pour former la dernière couche de la segmentation du flot optique.

Après cette décomposition itérative, chaque vecteur du flot optique appartient à une couche unique. Enfin, nous divisons chaque couche en composantes connectées pour former des segments de mouvement qui seront utilisés dans l'étape suivante de la SMS.

2.4 Localisation de régions de mouvement saillant

Dans cette étape, nous combinerons la carte d'attribution obtenue à l'aide de la méthode LRP, comme expliquée dans la sous-section 2.2, avec la carte de segmentation du flot optique décrite dans la sous-section 2.3. La seconde permettra de manipuler des régions bien segmentées, la première de désigner celles qui correspondent à des mouvements saillants. Selon [17], l'interprétation du réseau doit être principalement guidée par la force des valeurs d'attribution plutôt que par leur configuration spatiale. Par conséquent, nous prenons précisément en compte les valeurs d'attribution pour affecter ce qu'on appellera un score d'attribution à chaque région segmentée du flot optique. Ce score est donné par le rapport entre la somme des valeurs d'attribution dans la région, et la taille de la région. Ensuite, nous déterminons un seuil adaptatif sur le score d'attribution pour éliminer les régions aux scores d'attribution les plus bas. En pratique, nous prenons comme valeur de seuil le score de la plus grande région supposée correspondre à l'arrière-plan statique de la scène. Enfin, nous obtenons un ensemble de régions mobiles saillantes R_j.

2.5 Estimation des cartes de saillance du mouvement

Dans cette étape, nous souhaitons aller au-delà de la simple localisation des régions dynamiquement saillantes, en estimant une carte de saillance du mouvement, c'est-à-dire une carte de "chaleur" exprimant le degré de saillance dynamique de chaque pixel. Nous prenons tour à tour chaque

région mobile saillante R_j. Nous remplaçons le flot optique à l'intérieur de la région par un flot affine dont les paramètres ont été estimés grâce à une régression robuste sur les vecteurs du flot optique extérieur à la région. C'est une forme d'inpainting paramétrique du flot dans la région à partir du flot optique environnant. L'estimation robuste du modèle affine de flot est la même que celle décrite dans la sous-section 2.3. Les vecteurs de flot optique "inpainté" à l'intérieur de la région sont donnés par le modèle affine estimé à l'extérieur de la région, pour tout $p = (x, y) \in R_j$ par :

$$(u_p^{imp}, v_p^{imp}) = (\hat{a}_1 + \hat{a}_2 x + \hat{a}_3 y, \hat{a}_4 + \hat{a}_5 x + \hat{a}_6 y), \quad (1)$$

où les $\hat{a}_i$ désignent les paramètres du modèle de mouvement affine estimé. Ce flot paramétrique nous permet ensuite d'inférer la carte de saillance du mouvement ϕ à l'intérieur de R_j dont les valeurs sont comprises dans $[0, 1]$. Pour tout $p = (x, y) \in R_j$, nous avons :

$$\phi(p) = 1 - \exp(-\lambda ||w_p^{imp} - w_p||_2). \quad (2)$$

$\lambda > 0$ module la visualisation des cartes de saillance. $\phi(p) = 0$ pour les pixels n'appartenant pas aux régions mobiles saillantes.

3 Résultats expérimentaux

3.1 Détails de la mise en œuvre

Les champs de flot optique sont calculés à l'aide de la méthode RAFT [25] qui s'appuie sur des CNN. Nous avons utilisé une version modifiée de l'implémentation décrite dans [17] pour mettre en œuvre le LRP, Captum [26] pour tester la méthode des gradients intégrés, Scikit Learn [27] pour réaliser la régression de Huber. Nous avons transféré le modèle et les poids du réseau de [13] sur la librairie Pytorch [28], et mis à jour la dernière couche linéaire en utilisant l'ensemble de données d'entraînement original. Nous avons entraîné ce réseau pendant deux itérations, ce qui a nécessité environ 20 minutes sur un GPU GeForce MX150.

3.2 Évaluation quantitative

En raison du manque de benchmarks dédiés, l'évaluation expérimentale des méthodes de saillance du mouvement a été généralement liée aux benchmarks de segmentation d'objets vidéo (SOV). Cette situation n'est pas idéale, mais c'est la seule possibilité qui nous est offerte pour pouvoir comparer de manière quantitative notre méthode à des méthodes existantes proches. Le dataset DAVIS2016 est constitué de vidéos d'extérieur qui comportent presque toujours un seul objet mobile indépendant situé au premier plan et qui forme la vérité terrain. Cela induit un biais dans l'évaluation des méthodes de saillance de mouvement, puisque les objectifs respectifs entre SMS et SOV sont quelque peu différents comme souligné en introduction.

Comme le jeu de vidéos de DAVIS2016 ne comporte que très peu d'exemples de mouvement de parallaxe, cette évaluation quantitative se limitera à tester la précision et la régularité de la segmentation fournie par notre méthode. Soulignons que notre réseau de classification n'a pas été entraîné sur DAVIS2016, et qu'aucun exemple de ce jeu de vidéos n'a été utilisé pour concevoir l'étape de segmentation du flot optique. Pour chaque image, la carte des régions mobiles saillantes R_j extraite par notre méthode dans chaque image des vidéos de test de DAVIS2016, est comparée à la vérité terrain fournie dans le benchmark.

Nous comparons nos résultats uniquement à ceux des méthodes équivalentes du benchmark DAVIS2016, c'est-à-dire les méthodes "zero-shot" (au sens où elles n'utilisent pas la segmentation donnée par la vérité-terrain sur la première image de la vidéo) et non supervisées (au sens où elles ne sont pas entraînées sur les vidéos d'entraînement de DAVIS2016). Les résultats de comparaison sont rassemblés dans le tableau 1. Notre méthode se classe de manière satisfaisante, sachant que, en plus de ne pas utiliser d'étape de post-traitement, nous n'avons pas recours à l'apparence des objets, contrairement aux méthodes CIS et TIS_s.

TABLE 1 – Résultats sur le benchmark DAVIS2016 pour plusieurs méthodes non supervisées. J représente l'index de Jaccard (similarité des régions extraites à la vérité-terrain) et F rend compte de la précision des frontières des zones segmentées par rapport à la vérité-terrain.

	CIS [16]	TIS_s[2]	TIS_0[2]	BGM[29]	FTS [7]	Ours
J Mean ↑	71.5	67.6	58.6	62.5	55.8	62.3
F Mean ↑	70.5	63.9	47.5	59.3	51.1	60.9

3.3 Séparation des mouvements de parallaxe et des mouvements indépendants

Plus important encore, nous voulons évaluer la capacité de notre méthode à bien distinguer les mouvements de parallaxe et les mouvements indépendants dans les cartes de flot optique pouvant comprendre simultanément ces deux types de mouvement, puisque c'est en l'occurrence le véritable objectif de notre méthode de saillance du mouvement. C'est une tâche difficile pour laquelle la plupart des méthodes de segmentation même supervisées échouent du fait du manque d'exemples de ce type dans la base de vidéos d'entraînement.

L'évaluation ne pourra cependant qu'être visuelle et donc qualitative, car il n'existe pas de benchmarks conséquents disponibles avec vérité-terrain correspondante. Aussi à ce stade, nous avons constitué un petit ensemble de 5 vidéos acquises en milieu urbain avec une caméra tenue à la main, le porteur étant lui-même en mouvement. Chaque vidéo comprend à la fois des objets statiques sur le devant de la scène, comme des arbres, des pancartes ou des poteaux, induisant des mouvements de parallaxe, et des objets en mouvement dans la scène comme des piétons, des cyclistes ou des véhicules.

Au préalable, nous évaluons les performances du réseau de classification en lui-même en effectuant un test d'ablation. Nous dessinons un masque autour d'un objet au mouvement apparent distinctif au sein du flot optique, et nous remplaçons le flot optique à l'intérieur du masque par in-

FIGURE 3 – Impact des régions sur la prédiction de classification de saillance dynamique. En partant du flot optique (à gauche), nous remplaçons le flot dans les régions sélectionnées en utilisant la technique d'inpainting décrite à la section 2.5. Lorsque la région du tronc d'arbre au premier plan est ainsi inpaintée, l'image est toujours classée comme dynamiquement saillante (au milieu). En revanche, l'inpainting des régions correspondant aux deux personnes en mouvement fait alors passer l'image dans la classe des images dynamiquement non saillante (à droite). Voir le score de prédiction au-dessus de chaque image.

painting paramétrique à partir du flot environnant, comme expliqué en section 2.5. Cela permet de masquer le mouvement apparent de l'objet en question. Cette nouvelle carte de flot est ensuite donnée au réseau de classification et nous observons la modification éventuellement induite dans la prédiction.

La procédure est illustrée à la figure 3. Nous pouvons observer que seul le masquage du flot optique des objets mobiles indépendants (les deux piétons) fait significativement baisser le score de saillance dynamique, ce qui conduit le réseau à classer l'image ainsi modifiée comme "non dynamiquement saillante". Inversement, lorsque nous masquons dans le flot optique la zone correspondant à l'objet statique en avant-plan (le tronc d'arbre) causant le mouvement de parallaxe, le réseau continue à prédire l'image comme "dynamiquement saillante" en raison de la présence d'objets mobiles indépendants dans la scène observée. Cela montre que le réseau de classification de saillance dynamique fonde réellement sa prédiction sur la présence d'un mouvement indépendant.

Nous évaluons maintenant visuellement la localisation des régions mobiles saillantes. La figure 4 contient une comparaison entre la segmentation des mouvements saillants effectuée par notre méthode et celle fournie par la méthode CIS [16] qui donnait les meilleurs résultats globaux sur le dataset DAVIS2016 (voir tableau 1). Dans cette figure, nous représentons, par une surimpression en vert sur l'image d'origine de la vidéo, le masque binaire obtenu après l'étape de localisation des régions de mouvement saillant, décrite dans le paragraphe 2.4. Nous pouvons noter que notre méthode est capable d'extraire correctement les personnes qui marchent comme des régions mobiles saillantes, tout en n'étant pas sensible aux objets statiques du premier plan. Au contraire, la méthode CIS [16] segmente également ces objets statiques, qui produisent des mouvements de parallaxe proéminents dans le flot optique, comme des régions mobiles saillantes.

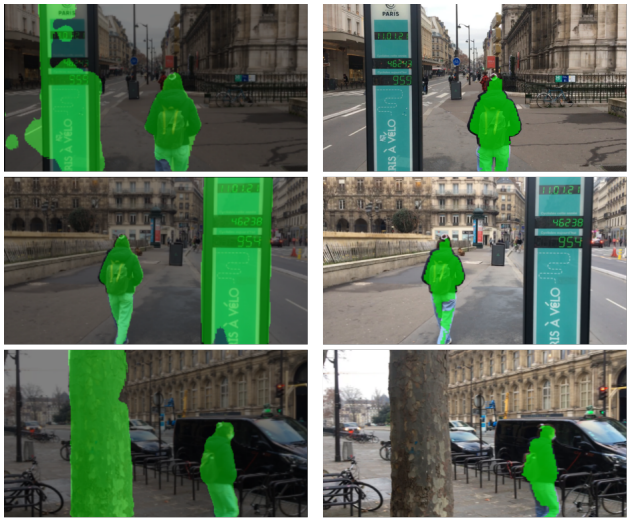

FIGURE 4 – Trois exemples de comparaison entre la méthode CIS [16] qui détecte, en plus de l'élément mobile de la scène, l'objet statique au premier plan (surimpression en vert, colonne de gauche), et notre méthode qui ne détecte que l'objet au mouvement indépendant (surimpression en vert, colonne de droite).

Enfin, nous affichons des échantillons de cartes de saillance de mouvement à la figure 5. Ces cartes sont le résultat obtenu à la fin de notre pipeline de segmentation des mouvements saillants. Rappelons que les valeurs de saillance sont données par la fonction ϕ définie à l'équation 2. Sur ces cartes, plus la couleur tend vers le jaune, plus la zone correspondante est reconnue comme dynamiquement saillante. Nous pouvons relever que les objets mobiles indépendants sont bien mis en évidence. Soulignons que sur ces exemples notre méthode réussit à distinguer les mouvements indépendants des mouvements de parallaxe. En effet les objets statiques au premier plan, le tronc d'arbre et le poteau dans les deux exemples les plus à gauche de la première rangée ainsi que la barrière dans celui de la séquence "Parkour" sont bien trouvés comme non saillants, ce qui n'est pas trivial, car leur flot optique correspondant à un mouvement de parallaxe est également proéminent. De plus, comme le montre l'exemple central de la première rangée, notre méthode est capable de détecter simultanément deux objets dynamiquement saillants, bien qu'ils aient un mouvement différent. Relevons au passage que le reflet mobile du cygne sur l'eau ainsi que l'eau écumant sous le kitesurfer sont correctement trouvés comme régions mobiles saillantes. Par contre, notre méthode s'en était trouvée pénalisée dans le benchmark DAVIS2016 (voir tableau 1, car ce dernier se fixe sur l'objet lui-même, objectif d'une SOV, et la vérité-terrain de ce benchmark se limite ainsi au cygne et au kitesurfer uniquement.

4 Conclusion

Nous avons conçu une méthode originale et efficace pour la segmentation de mouvements saillants. Elle ne nécessite que le flot optique en entrée, et implique une annota-

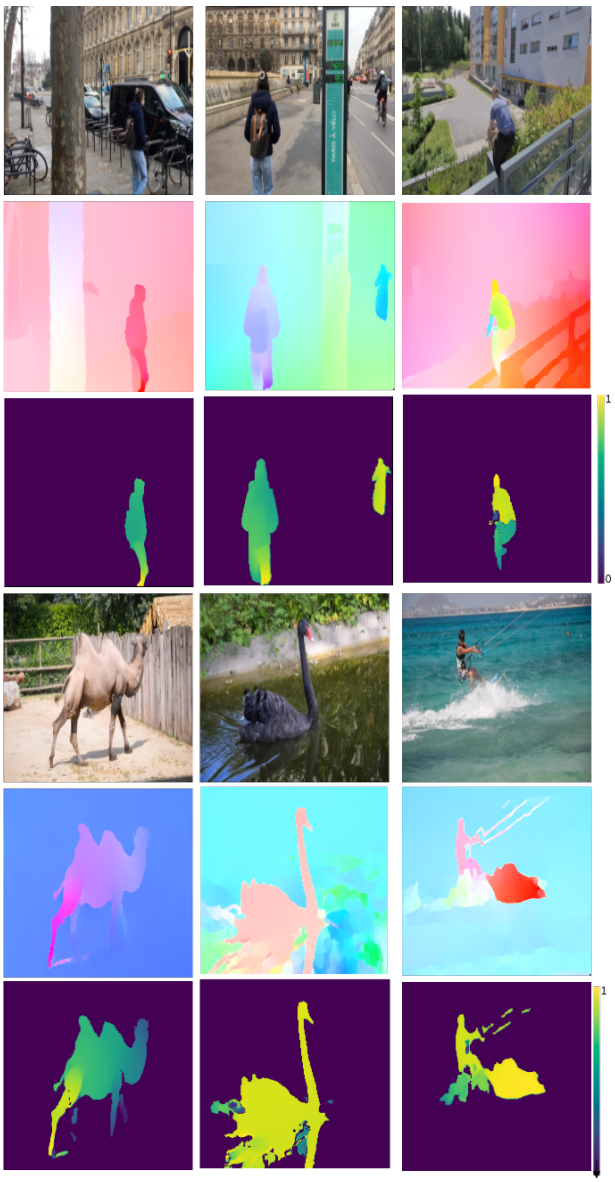

FIGURE 5 – Six exemples de cartes de saillance de mouvement. Sur la première rangée : de gauche à droite, deux vidéos en extérieur que nous avons acquises, et l'exemple "Parkour" du dataset DAVIS2016. Sur la deuxième rangée, trois exemples de DAVIS16 : "Camel", "Blackswan" et "Kite-surf". Dans chaque rangée : image de la vidéo (haut) ; flot optique (milieu) ; cartes de saillance de mouvement calculées par notre méthode (bas). Le violet correspond à des valeurs de saillance nulles, le jaune aux valeurs de saillance les plus élevées.

tion initiale très légère (en pratique au niveau de la vidéo) pour la classification basée sur les cartes de flot optique. Par contre, la segmentation elle-même est non supervisée. Notre méthode est capable de distinguer les mouvements indépendants d'une scène filmée par une caméra mobile, des mouvements de parallaxe dus aux objets statiques au premier plan de la scène, et ce sans aucune information 3D. Elle combine l'interprétation LRP du réseau de classification et la segmentation du flot optique. Au-delà de la segmentation des mouvements saillants, elle produit l'estimation de cartes de saillance de mouvement en exploitant une technique d'inpaintaing paramétrique du flot optique. Elle est ainsi d'intérêt pour différentes applications liées à l'analyse de scènes dynamiques par vidéo. Des résultats expérimentaux ont permis d'évaluer favorablement sa précision et sa fiabilité.

Remerciements

Ce travail a été financé par Bpifrance à travers le contrat LiChIE dans le cadre du PSPC.

Références

[1] Anestis Papazoglou and Vittorio Ferrari, "Fast Object Segmentation in Unconstrained Video," in *2013 IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 2013, pp. 1777–1784, IEEE.

[2] Brent Griffin and Jason Corso, "Tukey-Inspired Video Object Segmentation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 1723–1733, IEEE.

[3] Wenguan Wang, Jianbing Shen, and Ling Shao, "Video Salient Object Detection via Fully Convolutional Networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[4] T. Le and A. Sugimoto, "Video Salient Object Detection Using Spatiotemporal Deep Features," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5002–5015, Oct. 2018.

[5] J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, June 2016, pp. 724–732, IEEE.

[6] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid, "Learning Video Object Segmentation with Visual Memory," *arXiv :1704.05737 [cs]*, July 2017.

[7] Hongmei Song et al., "Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection," in *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds., vol. 11215, pp. 744–760. Springer International Publishing, Cham, 2018.

[8] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman, "FusionSeg : Learning to combine motion and appearance for fully automatic segmention of generic objects in videos," *arXiv :1701.05384 [cs]*, Apr. 2017.

[9] Juan-Manuel Pérez-Rúa, Antoine Basset, and Patrick Bouthemy, "Detection and Localization of Anomalous Motion

in Video Sequences from Local Histograms of Labeled Affine Flows," *Frontiers in information and communication technologies*, May 2017.

[10] Anurag Ranjan et al., "Competitive Collaboration : Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12240–12249.

[11] Pia Bideau, Rakesh R. Menon, and Erik Learned-Miller, "MoA-Net : Self-Supervised Motion Segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[12] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski, "A database and evaluation methodology for optical flow," *International journal of computer vision*, vol. 92, no. 1, pp. 1–31, 2011.

[13] Léo Maczyta, Patrick Bouthemy, and Olivier Le Meur, "CNN-based temporal detection of motion saliency in videos," *Pattern Recognition Letters*, vol. 128, pp. 298, Dec. 2019.

[14] J. M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, Dec. 1995.

[15] Mathilde Guillemot et al., "Breaking Batch Normalization for better explainability of Deep Neural Networks through Layer-wise Relevance Propagation," *arXiv :2002.11018 [cs, stat]*, Feb. 2020.

[16] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto, "Unsupervised Moving Object Detection via Contextual Information Separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 879–888.

[17] Jindong Gu, Yinchong Yang, and Volker Tresp, "Understanding individual decisions of cnns via contrastive backpropagation," in *Asian Conference on Computer Vision*. 2018, pp. 119–134, Springer.

[18] Julius Adebayo et al., "Sanity Checks for Saliency Maps," *arXiv :1810.03292 [cs, stat]*, Nov. 2020.

[19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic Attribution for Deep Networks," *arXiv :1703.01365 [cs]*, June 2017.

[20] Pascal Sturmfels, Scott Lundberg, and Su-In Lee, "Visualizing the Impact of Feature Attribution Baselines," *Distill*, vol. 5, no. 1, pp. e22, Jan. 2020.

[21] Grégoire Montavon et al., "Layer-Wise Relevance Propagation : An Overview," in *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, Eds., vol. 11700, pp. 193–209. Springer International Publishing, Cham, 2019.

[22] Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller, "Explaining NonLinear Classification Decisions with Deep Taylor Decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, May 2017.

[23] J.Y.A. Wang and E.H. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, Sept./1994.

[24] J.-M. Odobez and P. Bouthemy, "MRF-based motion segmentation exploiting a 2D motion model robust estimation," in *Proceedings., International Conference on Image Processing*, Washington, DC, USA, 1995, vol. 3, pp. 628–631, IEEE Comput. Soc. Press.

[25] Zachary Teed and Jia Deng, "RAFT : Recurrent All-Pairs Field Transforms for Optical Flow," *arXiv :2003.12039 [cs]*, Aug. 2020.

[26] Narine Kokhlikyan et al., "PyTorch captum," *GitHub repository*, 2019.

[27] Fabian Pedregosa et al., "Scikit-learn : Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[28] Adam Paszke et al., "PyTorch : An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[29] Scott Wehrwein and Richard Szeliski, "Video Segmentation with Background Motion Models," in *Procedings of the British Machine Vision Conference 2017*, London, UK, 2017, p. 96, British Machine Vision Association.