Détection d'objets en mouvement dans un milieu urbain par fusion de données RVB et événementielles

Zhuyun ZHOU[1] Zongwei WU[1, 2] Rémi BOUTTEAU[3] Fan YANG[1] Cédric DEMONCEAUX[1, 4] Dominique GINHAC[1]

[1]ICB, UMR CNRS 6303, Université de Bourgogne, Dijon, France

[2]CVL, ETH Zurich, Zürich, Suisse

[3]LITIS UR 4108, INSA Rouen, Univ Rouen Normandie, Rouen, France

[4]LORIA, Université de Lorraine, CNRS, INRIA, Nancy, France

Résumé – La détection d'objets en mouvement (DOM) est une tâche essentielle pour parvenir à une conduite autonome sûre. En vision par ordinateur, les méthodes d'apprentissage profond donnent des résultats intéressants mais elles sont généralement basées sur des images classiques et souffrent par conséquent des limites de ce type de capteur dans des scènes hautement dynamiques. Les avancées technologiques récentes dans le développement de capteurs bio-inspirés, en particulier les caméras événementielles, peuvent naturellement compléter les données issues des caméras conventionnelles afin de mieux modéliser les objets en mouvement. Nous proposons dans cet article, un nouveau réseau de fusion de données multimodales RVB/événements, appelé RENet (*RGB-Event fusion Network*), qui exploite conjointement ces deux modalités complémentaires pour obtenir une DOM plus robuste dans des scénarios difficiles de conduite autonome.

Abstract – Moving Object Detection (MOD) is a critical vision task for successfully achieving safe autonomous driving. Despite plausible results of deep learning methods, most existing approaches are only frame-based and may fail to reach reasonable performance when dealing with dynamic traffic participants. Recent advances in sensor technologies, especially the Event camera, can naturally complement the conventional camera approach to better model moving objects. we propose RENet, a novel RGB-Event fusion Network, which jointly exploits the two complementary modalities to achieve more robust MOD under challenging scenarios for autonomous driving.

1 Introduction

Malgré les avancées significatives des réseaux de neurones profonds en vision par ordinateur, ceux-ci ne donnent pas encore des résultats satisfaisants pour la détection d'objets en mouvement (DOM) dans les scènes à haute dynamique en utilisant uniquement une caméra RVB. Ceci est dû aux limites intrinsèques de ce type de capteur. Les récentes caméras bio-inspirées dites événementielles suscitent beaucoup d'intérêt, en particulier pour la conduite des véhicules autonomes. En effet, leur mode de fonctionnement asynchrone avec une fiable latence, permet de capter de riches informations temporelles, même dans des environnements complexes comme la nuit ou la sortie d'un tunnel.

Un événement est un signal à 4 entrées (x, y, p, t), avec (x, y) les coordonnées d'un pixel dans l'image et p sa polarité de mouvement à l'instant t. Dans la littérature, on trouve 3 représentations populaires des événements correspondant respectivement aux formats Image (x, y, p) [1, 2], Temps-Surface (x, y, t) [3] et aussi Voxel (x, y, p, t) [4, 5]. Seule la dernière permet d'exploiter pleinement toutes les informations spatio-temporelles, mais sa structure 4D réclame un coût de calcul trop élevé pour envisager une application en temps réel comme la conduite autonome.

En ce qui concerne les travaux existants sur la fusion de données RVB et événements, la plupart des méthodes prédéfinit d'abord un intervalle fixe de temps pour représenter les événements avant de combiner les caractéristiques de deux modalités d'une manière unidirectionnelle (événement vers RVB). Ces approches souffrent de 3 inconvénients majeurs : 1) une fenêtre temporelle fixe ne peut pas s'adapter aux différents scénarios sans ajustement manuel, 2) un petit intervalle temporel comme le temps d'exposition de la caméra RVB ne permet pas de profiter pleinement des riches informations temporelles fournies par la caméra événementielle, tandis qu'une plus grande valeur ajoute du bruit d'alignement et 3) une fusion unidirectionnelle exclut la possibilité d'exploiter les informations croisées inter-modales.

Dans ce papier, nous proposons d'abord un nouveau format de représentation d'événement basé sur un module d'agrégation multi-échelle permettant d'utiliser pleinement les riches informations temporelles des événements qui sont cruciales pour la DOM. Ensuite, nous introduisons un nouveau schéma de fusion RVB/événements, modélisant simultanément les propriétés spécifiques de chaque modalité, ainsi que leur relation inter-modale. Pour tester et valider notre réseau RENet, nous proposons également un jeu de données DSEC-MOD, qui est construit à partir de DSEC [6], sur lequel différents modules de fusion de l'état de l'art ont été évalués. L'étude comparative démontre la supériorité de notre approche.

2 Travaux connexes

Avec le développement des caméras événementielles, la fusion multimodale RVB/événements a attiré de plus en plus l'attention des chercheurs. Certains travaux adoptent un réseau pré-entraîné pour réaliser la conversion événement-image [7]. Ensuite, plusieurs travaux traitent directement la vidéo générée

de la même manière que la vidéo RVB pour atteindre l'application cible. [8] adopte une conception de fusion intermédiaire par le biais d'une simple concaténation de convolutions. [9] exploite en outre l'attention du transformateur pour mieux guider la fusion des caractéristiques. Partageant une idée similaire, nous suivons également la conception de fusion intermédiaire pour fusionner RVB et événements. Contrairement à [8], [9] qui utilisent une conception de fusion unidirectionnelle, nous modélisons explicitement les caractéristiques spécifiques à la modalité et à la représentation partagée. En outre, nous introduisons un module d'étalonnage intermodal pour corriger attentivement les réponses bruitées de chaque modalité avant la fusion des caractéristiques. Contrairement à [9], qui se concentre sur l'axe de canal, notre module d'attention prend en compte la dimension spatiale qui joue un rôle important dans la localisation des objets au sein d'une image.

3 Méthodologie

La Figure 1 illustre notre réseau RENet composé d'un module E-TMA (*Event-based Temporal Multi-scale Aggregation*), d'un double encodeur pour extraire les caractéristiques de chaque modalité, de modules bi-directionnels de fusion BDC (*Bi-Directional Calibration*) et d'un décodeur pour détecter et localiser les objets en mouvement. Nous décrivons par la suite le fonctionnement de chaque composant.

3.1 E-TMA : Agrégation multi-échelle temporelle des événements

Les objets en mouvement peuvent légèrement bouger dans un petit intervalle temporel : plus la plage temporelle augmente, plus il y a d'événements, c'est-à-dire que les contours des objets deviennent plus larges (voir la Figure 2). A partir de ce constat, et pour capter les riches informations temporelles fournies par la caméra événementielle, nous proposons un module d'agrégation multi-échelle temporelle des événements. Pour cela, nous prenons d'abord 3 échelles temporelles à l'entrée du module E-TMA, respectivement $\varepsilon_1 = 15ms$, $\varepsilon_2 = 30ms$ et $\varepsilon_3 = 50ms$. Ensuite, les événements sont projetés dans un espace 2D de caractéristiques en appliquant η. Cette étape peut se décrire mathématiquement comme suit :

$$E'_i = \eta(\varepsilon_i), i \in 1, 2, 3. \tag{1}$$

En fonction de l'échelle temporelle, les opérations de regroupement (*pooling*) ont été réalisées en utilisant différentes tailles de fenêtre dans l'objectif de mieux modéliser les caractéristiques locales. Les résultats de *pooling* ont subi des sur-échantillonnages avant d'être concaténés.

Contrairement aux travaux existants, nous effectuons la fusion des événements de différentes échelles dans l'espace de caractéristiques à l'aide des opérations de *pooling* en exploitant leur propriété d'invariance spatiale et en rotation. En particulier, la fonction de Max *pooling* a été choisie pour extraire les informations les plus significatives en vue de mieux détecter les objets en mouvement.

3.2 BDC : Calibrage Bi-Directionnel

A la sortie du module E-TMA, les caractéristiques agrégées des événements sont envoyées à un encodeur ainsi que l'image RVB (voir la Figure 1). Le rôle des encodeurs consiste à réaliser la modélisation sémantique de chaque modalité. Du fait que les images RVB possèdent beaucoup plus de textures par rapport aux images des événements, nous avons appliqué 2 architectures différentes d'encodeur (ResNet-101 pour RVB et ResNet-18 pour les événements), ceci nous permet d'optimiser le coût de calcul. A la sortie de différents étages d'encodeur, avant de réaliser la fusion des 2 modalités, les projections ramènent les cartes sémantiques des événements à la même taille que celles du RVB.

La fusion RVB/événements est un domaine de recherche relativement récent, donc très peu de travaux existent dans la littérature, en particulier pour la fusion au niveau des encodeurs. Par conséquent, nous avons conçu un nouveau schéma de fusion d'attention pour prendre en compte simultanément les indices les plus significatifs spatialement et aussi en fonction de l'axe des canaux des encodeurs.

Prenons la fusion au niveau sémantique le plus élevé comme exemple, pour l'élément RVB $f_r \in \mathbb{R}^{C\times h\times w}$ et l'élément d'événement projeté $f_e \in \mathbb{R}^{C\times h\times w}$ (C, h, et w correspondent respectivement au nombre de canaux, la hauteur et la largeur de chaque carte), nous effectuons des fusions attentives allant du grain grossier au grain fin. Plus précisément, nous utilisons tout d'abord une multiplication des pixels et une addition pour améliorer grossièrement les caractéristiques les plus informatives dans chaque modalité. Formellement, les caractéristiques améliorées f'_r et f'_e sont calculées par :

$$f'_r = f_r \otimes f_e + f_r; \quad f'_e = f_r \otimes f_e + f_e. \tag{2}$$

Pour l'axe du canal, nous procédons à un calibrage bi-directionnel, c'est-à-dire que nous apprenons les caractéristiques d'une modalité et les appliquons à l'autre. Ensuite, nous fusionnons les caractéristiques multimodales calibrées tout en conservant les caractéristiques les plus informatives. Soit CA le module d'attention au canal de [11], le calibrage croisé le long de l'axe du canal peut être formulé comme suit :

$$f_r^{CA} = CA(f'_e) \otimes f'_r + f'_r; \quad f_e^{CA} = CA(f'_r) \otimes f'_e + f'_e. \tag{3}$$

En utilisant le même protocole, nous affinons les caractéristiques de manière spatiale. Soit SA l'attention spatiale de [11], nous obtenons les caractéristiques améliorées finales f_r^{enh} et f_e^{enh} comme suit :

$$\begin{aligned} f_r^{enh} &= SA(f_e^{CA}) \otimes f_r^{CA} + f_r^{CA}; \\ f_e^{enh} &= SA(f_r^{CA}) \otimes f_e^{CA} + f_e^{CA}. \end{aligned} \tag{4}$$

Ces modèles visent à trouver les canaux/pixels les plus pertinents et les plus sûrs au sein de chaque caractéristique spécifique à une modalité. En outre, le calibrage croisé permet de modéliser sélectivement les caractéristiques bruitées tout en conservant les informations pertinentes, ce qui constitue un moyen simple mais efficace de renforcer la modélisation des caractéristiques à l'aide d'indices complémentaires. Enfin, les caractéristiques extraites sont soigneusement fusionnées par le biais d'une convolution qui apprend les poids de contribution des composants les plus informatifs afin de former la sortie partagée :

$$f = Conv_{3\times3}(Concat(f_r^{enh} \otimes f_e^{enh}; max(f_r^{enh}, f_e^{enh}))). \tag{5}$$

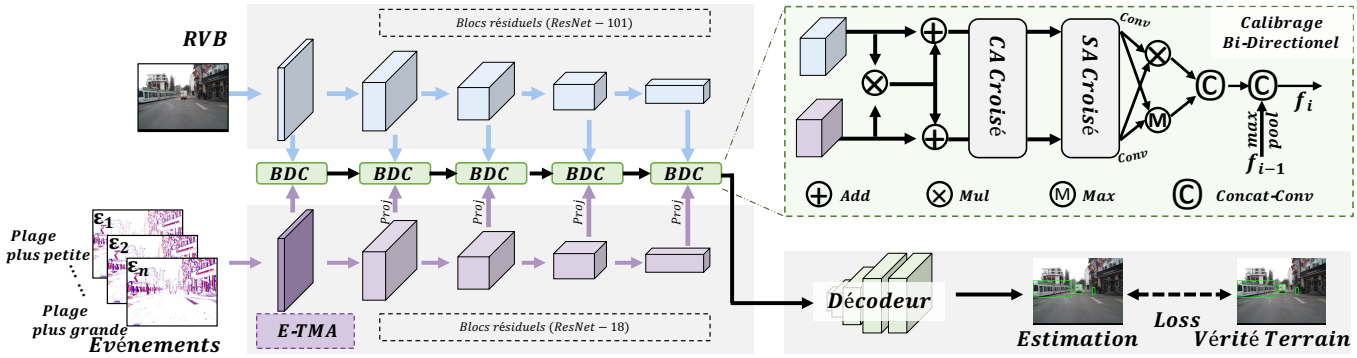

FIGURE 1 : Architecture de RENet : l'originalité de notre réseau RENet consiste en un module E-TMA pour agréger les événements temporels et des modules BDC de fusion bi-directionnelle RVB/événements au niveau intermédiaire. Le décodeur et la fonction "*Loss*" sont inspirés de [10].

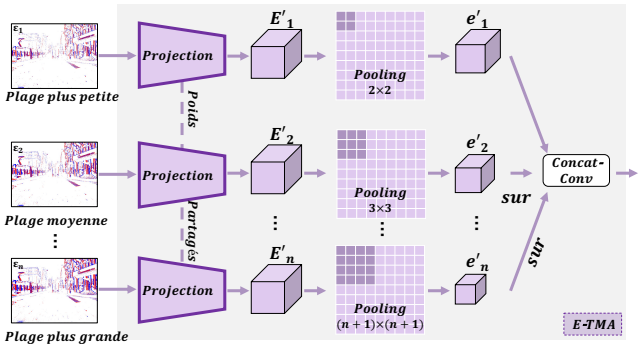

FIGURE 2 : Illustration du module E-TMA : agrégation multi-échelle temporelle des événements avec un réseau de neurones à poids partagés.

4 Jeu de données

A notre connaissance, DSEC est le plus grand jeu de données RVB/événements disponible pour la conduite autonome. Cependant, DSEC ne fournit pas les annotations des objets en mouvement. Pour expérimenter notre proposition, la première étape consiste à construire un jeu de données adéquat. Pour ce faire, nous avons d'abord calibré les 2 entrées : les images RVB sont rectifiées en suivant la méthode proposée par [6] pour que les 2 cartes possèdent le même champ de vue et la même résolution. Ensuite, en visualisant les séquences vidéo, nous avons manuellement annoté les objets en mouvement pour chaque image en les encadrant par des boîtes englobantes. Dans ce jeu de données DSEC-MOD (*Moving Object Detection*), nous considérons 8 types d'objets mobiles : voiture, camion, bus, train, piéton, cycliste, motocycliste et autres. Chaque séquence contient au moins 3 différents types d'objets en mouvement. Au total, nous avons labélisé 16 séquences (13314 images), avec 11 séquences (10495 images) pour l'apprentissage et 5 autres séquences (2819 images) pour le test. DSEC-MOD est le plus grand jeu de données de ce type.

A partir de DSEC-MOD, beaucoup de travaux peuvent être réalisés, notamment la classification et la segmentation. La classification correspond à catégoriser les objets par classe ; la Figure 3 montre deux exemples de DSEC-MOD avec l'étiquetage des objets. La segmentation peut être considérée comme une détection plus fine, au niveau des pixels, des objets en mouvement. Ce travail a été déjà réalisé par DSEC-MOS [12] (voir Figure 4 comme exemples).

FIGURE 3 : Deux exemples de la classification.

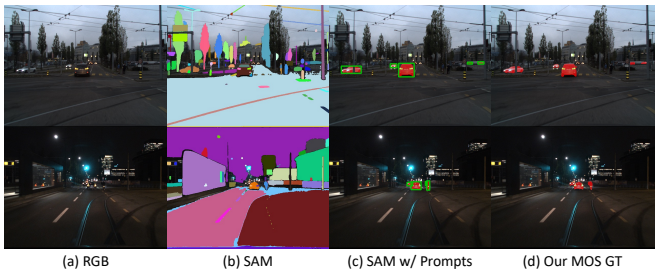

FIGURE 4 : Deux exemples DSEC-MOS [12] : (a) RVB calibrés sur événements ; (b) SAM [13] direct ; (c) SAM avec des boîtes englobantes de DSEC-MOD ; (d) Résultats de segmentation DSEC-MOS, visualisées en rouge sur les images RVB.

5 Résultats expérimentaux

Protocole : Puisque nous avons utilisé ResNet-101 et ResNet-18 comme encodeurs, les images d'entrée sont ramenées à 288×288 pixels. Les techniques classiques d'augmentation de données ont été appliquées à notre RENet. Le coefficient d'apprentissage initial est de 5e-4, diminuant avec un facteur de 10 respectivement aux époques 10 et 15 en utilisant l'optimiseur Adam. L'apprentissage a été accompli au bout de 30 époques. Nous évaluons les performances du RENet avec F.mAP (*Frame mean Average Precision*) qui mesure la qualité de détection pour chaque image.

Résultats et discussions : Pour connaître la contribution de chaque composant de RENet, des expériences d'ablation ont aussi été conduites dont les résultats sont illustrés dans la Table 1. A la configuration de base RVB, nous avons graduellement ajouté les différents modules, les performances de détection augmentent au fur et à mesure. Nous avons aussi remplacé le module d'agrégation E-TMA par la récente méthode "multi-échelle accumulation" qui fait baisser le taux de réussite : une

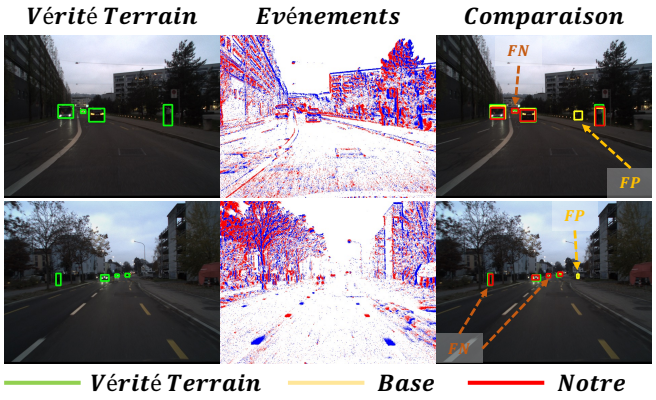

FIGURE 5 : Deux exemples de détection d'objets en mouvement : la fusion RVB/événements obtient les meilleurs résultats. FP et FN sont les abréviations de *False Positive* et *False Negative*, qui signifient les fausses détections et les détections manquantes.

TABLE 1 : Etudes d'ablation : R-E correspond à une simple fusion RVB/événements à l'aide de "concat-conv" sans aucune forme d'attention. CA et SA sont les abréviations de *Channel Attention* et *Spatial Attention*. F. mAP est la moyenne de la précision moyenne des frames, 0.5 est le seuil.

#	R-E	CA	SA	E-TMA	Acc. [9]	F. mAP@0.5
1	✓					34.25
2	✓			✓		35.56
3	✓	✓		✓		36.53
4	✓		✓	✓		37.75
5	✓	✓	✓		✓	36.34
6	✓	✓	✓	✓		**38.38**

simple concaténation de multiple images d'événements à l'entrée ne suffit pas pour mieux exploiter leurs riches informations temporelles.

Pour mieux comprendre l'apport de la caméra événementielle, nous avons aussi analysé les performances obtenues sous différentes conditions d'éclairage. La figure 5 montrent 2 exemples avec la vérité terrain, la fusion RVB et événements permet de réduire les fausses détections.

Pour valider notre proposition, une étude comparative avec l'état de l'art a été faite, les détails peuvent être trouvés dans [14].

6 Conclusions et perspectives

Dans ce papier, nous avons proposé une nouvelle architecture de fusion RVB/événements pour la détection d'objets en mouvement. Nous introduisons une agrégation multi-échelle des événements pour exploiter les riches informations temporelles asynchrones. De plus, les caractéristiques hétérogènes RVB/événements ont été fusionnées d'une manière grossière à fine, ce qui donne un mécanisme de calibrage simple et efficace à l'aide de différentes formes d'attention.

Nous avons aussi construit un nouveau jeu de données DSEC-MOD, à partir de scènes de DSEC, pour encourager et promouvoir les recherches sur la fusion RVB/événements pour la détection d'objets en mouvement. Les nombreux résultats expérimentaux valident notre approche par rapport à l'état de l'art. Dans le futur, nous allons optimiser le réseau RENet pour réduire sa complexité de calcul et les ressources nécessaires des mémoires et ceci pour pouvoir l'implémenter dans des systèmes embarqués.

Remerciements : Ces travaux sont financés par l'Agence Nationale de la Recherche (CERBERE, ANR-21-CE22-0006).

Références

[1] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet : Self-supervised optical flow estimation for event-based cameras," in *RSS*, 2018.

[2] M. Liu and T. Delbruck, "Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors," in *BMVC*, 2018.

[3] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, "Semi-dense 3d reconstruction with a stereo event camera," in *ECCV*, 2018.

[4] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens : Event-based video frame interpolation," in *CVPR*, 2021.

[5] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video : Bringing modern computer vision to event cameras," in *CVPR*, 2019.

[6] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec : A stereo event camera dataset for driving scenarios," *RAL*, vol. 6, no. 3, pp. 4947–4954, 2021.

[7] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *TPAMI*, 2019.

[8] A. Tomy, A. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, "Fusing event-based and rgb camera for robust object detection in adverse conditions," in *ICRA*, 2022.

[9] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. Van Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *ECCV*, 2022.

[10] Y. Li, Z. Wang, L. Wang, and G. Wu, "Actions as moving points," in *ECCV*, 2020.

[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam : Convolutional block attention module," in *ECCV*, 2018.

[12] Z. Zhou, Z. Wu, R. Boutteau, F. Yang, and D. Ginhac, "Dsec-mos : Segment any moving object with moving ego vehicle," *arXiv preprint arXiv :2305.00126*, 2023.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv :2304.02643*, 2023.

[14] Z. Zhou, Z. Wu, R. Boutteau, F. Yang, C. Demonceaux, and D. Ginhac, "Rgb-event fusion for moving object detection in autonomous driving," in *ICRA*, 2023.