Internship proposition (Ms. Student)

Multimodal visual saliency: salient plane detection for driver assistance systems.

Keywords: Artificial visual attention, visual salience maps, deep learning, multimodal vision, event cameras.

1. Context

The driver's risk assessment depends on his perception of the scene and his relevance to the objects perceived to his driving task. These objects correspond to parts of the scene on which he has focused, either because they are prominent and catch the eye or because they are deemed relevant by the driver to accomplish his goal of driving safely. In a scene perceived by a driver, the salient objects are those which, compared to the other elements present in the visual field, have the most capacity to attract his attention and direct the gaze (and therefore to divert it from other areas. potentially at risk). Indeed, the attentional brain activity is organized and optimized to take into account the visual characteristics of the scene and the personal driver motivation/interest. The meeting of the human vision/artificial vision communities has given rise to a standard formalism that proposes to model low-level attentional systems using salience maps. These saliency maps, resulting from the fusion of primitives based on orientation, contrast, and movement, encode and capture the salient characteristics of objects. They result from exogenous mechanisms guided only by the stimuli present in the observer's visual field. In addition, specific works highlight a model which distinguishes search salience (ease of finding a sought-after object in a scene) from attentional salience (the ability of objects to "jump out").

We propose in this internship to take advantage of the work carried out at the IBISC laboratories in the field of the detection of salient planes from a mobile camera [4]. In this work, we propose a saliency map adapted to the dynamic case (of a vehicle in motion), which would thus take into account the vehicle's movement, structure, and environment movement. The project will show the contribution and the level at which dynamic vision and geometry are integrated into the construction of salience maps.

The vision system considered is said to be "multimodal." It has the particularity of including so-called event cameras. Neuromorphic cameras create a real craze in the scientific community, particularly among roboticists. The best known as DVS (Dynamic Vision Sensor) aim to overcome the known drawbacks of conventional cameras, namely: temporal redundancy due to an arbitrary acquisition frequency which does not depend on changes/movements in the scene; their low dynamics (in the order of 60 to 70db versus 140db for a natural scene); and a signal-to-noise ratio that is highly dependent on the stage lighting.

Event cameras are bio-inspired sensors that operate at the lowest level, radically different from traditional cameras. Instead of acquiring images with a fixed rate, each pixel's brightness changes are calculated entirely asynchronously. The result is a flow of events, with the possibility, for each event, of finding its timestamp, x and y position, and the sign of the change in brightness called polarity. Compared to traditional cameras, these cameras have exceptional properties: a very high dynamic range (140 dB against 60 dB), high temporal resolution (of the order of µs), and shallow

power consumption. In addition, the movement of objects or the camera does not generate blur. These advantages make them cameras with great potential for robotics and computer vision in scenarios that challenge traditional cameras, especially in the case of high speeds and/or high dynamic ranges. However, their use requires rethinking and readjusting classic algorithms.

2. Objectives

The objectives of the work are as follows:

- Understand the current work on models describing human attention in a driving situation.

- Understand the progress of an "attentional" processing chain based on a dual bottom-up (starting from low-level visual primitives) and top-down (depending on the task in progress or targeted by the user). The bottom-up approach is based on the construction of low-level saliency maps based on position, spatial frequencies, color, orientations, texture, and movement. For the top-down approach, neural network learning techniques could be considered to detect objects present in road scenes and label them semantically (markings, signs, pedestrians, vegetation, road, etc.).

- Lead a data-set composition with visual and event-based camera information from real-world driving scenarios.

- Implement an algorithm for detecting salient planes from an onboard event camera.

3. Required profile

You are curious, creative, independent, and rigorous. Your interest in innovation, research in computer vision and image processing characterizes you. Your interest is in research issues open at the crossroads between several disciplines (neurosciences, physiology, computer science).

4. Academic training sought

- Master / Engineer with skills in computer vision and image processing.
- Languages: C, C++
- English: good level.

5. Application procedures: Applications (CV + cover letter) must be sent directly by email to the contact below. The transmission of transcripts will be appreciated.

Training period: 6 months

Start date of internship: February or March 2022

Supervision and contact:

- Fabien Bonardi (assistant professor, IBISC), fabien.bonardi@ibisc.univ-evry.fr
- Samia Bouchafa-Bruneau (full Professor, IBISC), samia.bouchafa@ibisc.univ-evry.fr

References:

[1] Laurent Itti, Christof Koch and Ernst Niebur: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence, 20(11):1254–1259, 1998.

[2] Ali Borji, Laurent Itti: Sate-of-the-art in visual attention modeling. IEEE Transactions on pattern analysis and machine intelligence, 35(1):185–207, 2013.

[3] Richard Veale, Ziad M. Hafed and Masatoshi Yoshida: How is visual salience computed in the brain? Insights from behavior, neurobiology, and modelling. Philos Trans R Soc Lond B Biol Sci. 2017 Feb 19;372(1714).

[4] Viachaslau Kachurka, Kurosh Madani, Christophe Sabourin, Vladimir Golovko. Visual saliency based approach to object detection in computer vision systems: Real life applications. IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015.

[5] Tan Khoa Mai. Vers un système de vision artificielle opportuniste pour l'analyse de scènes complexes à partir de caméras embarquées. Thèse de L'université Paris-Saclay, le 13 décembre 2018.

[6] Benois-Pineau, J., & Mitrea, M. (2017, November). Extraction of saliency in images and video: Problems, methods and applications. A survey. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6).

[7] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco and Tobi Delbruck "Retinomorphic Event-Based Vision Sensors : Bioinspired Cameras With Spiking Output".

[8] Ryad Benosman et al. "Event-based visual flow". In :IEEE transactions on neural net-works and learning systems25.2 (2013), pp. 407-417

[9] Guillermo Gallego et al. "Event-based vision : A survey". In :arXiv preprint arXiv :1904.08405(2019).

[10] Bodo Rueckauer and Tobi Delbruck. "Evaluation of event-based algo- rithms for opticalflow with ground-truth from inertial measurement sen- sor". In :Frontiers in neuroscience10 (2016), p. 176.

[11] Elias Mueggler et al. "Lifetime estimation of events from dynamic vision sensors". In :2015 IEEE international conference on Robotics and Automa- tion (ICRA). IEEE. 2015,pp. 4874-4881.

Proposition de stage niveau Master 2

Saillance visuelle multimodale.

Application à la détection de plans saillants dans un système d'aide à la conduite.

Mots clés : Attention visuelle artificielle, cartes de saillance visuelles, deep learning, vision multimodale, caméras évènementielles.

1. Contexte

L'évaluation du risque par le conducteur est dépendante de sa perception de la scène et de la pertinence qu'il attribue aux objets perçus vis-à-vis de sa tâche de conduite. Ces objets correspondent à des parties de la scène sur lesquelles s'est porté son regard, soit parce qu'elles sont saillantes et attirent le regard, soit parce qu'elles sont jugées pertinentes par le conducteur pour accomplir son objectif de conduire en sécurité. Dans une scène perçue par un conducteur, les objets saillants sont ceux qui, par rapport aux autres éléments présents dans le champ visuel, ont le plus de capacité à attirer son attention et à diriger le regard (et donc à le détourner d'autres zones potentiellement à risque). En effet, l'activité attentionnelle du cerveau est organisée et optimisée pour tenir compte des caractéristiques visuelles de la scène et de la motivation/intérêt subjectif du conducteur. La rencontre des communautés vision humaine/vision artificielle a donné lieu à un formalisme commun qui propose de modéliser les systèmes attentionnels bas niveau par des cartes de saillance. Ces cartes de saillance, issues de la fusion de primitives basées sur l'orientation, le contraste et le mouvement, codent et captent les caractéristiques saillantes des objets, elles sont le résultat de mécanismes exogènes guidés uniquement par les *stimuli* présents dans le champ visuel de l'observateur. Par ailleurs, certains travaux mettent en évidence un modèle qui distingue la saillance de recherche (facilité à trouver un objet recherché dans une scène) de la saillance attentionnelle (capacité des objets à « sauter aux yeux »).

Nous nous proposons dans ce stage de tirer profit des travaux menés au laboratoires IBISC dans le domaine de la détection des plans saillants à partir d'une caméra mobile [4] afin de proposer une carte de saillance adaptée au cas dynamique (d'un véhicule en mouvement), qui prendrait ainsi en compte le mouvement propre du véhicule, la structure et le mouvement de l'environnement. Le projet permettra de montrer l'apport et le niveau dans lesquels sont intégrés la vision dynamique et la géométrie dans la construction des cartes de saillance.

Le système de vision considéré est dit « multimodal ». Il a la particularité d'inclure des caméras dites évènementielles. Les cameras dites événementielles ou encore neuromorphiques créent actuellement un véritable engouement dans la communauté scientifique, chez les roboticiens notamment. Les plus connues comme DVS (Dynamic Vision Sensor) visent à pallier les inconvénients connus des caméras conventionnelles à savoir : la redondance temporelle due à une fréquence d'acquisition arbitraire qui ne dépend pas des changements/mouvements dans la scène

; leur faible dynamique (de l'ordre de 60 à 70db contre 140 db pour une scène naturelle) ; et un rapport signal sur bruit fortement dépendant de l'éclairage de la scène.

Les caméras évènementielles sont des capteurs bio-inspirés qui fonctionnent au plus bas niveau de manière radicalement différente des caméras traditionnelles. Au lieu d'acquérir des images avec une fréquence d'acquisition fixe, les changements de luminosité pour chaque pixel sont calculés de manière totalement asynchrone. Il en résulte un flux d'événements, avec pour chaque événement la possibilité de retrouver son horodatage (*timestamp*), sa position x et y, et le signe du changement de luminosité en ce point, appelé polarité. Ces caméras possèdent des propriétés exceptionnelles par rapport aux caméras traditionnelles : une plage de dynamique très élevée (140 dB contre 60 dB), une haute résolution temporelle (de l'ordre de µs) et une très faible consommation d'énergie. De plus, le mouvement des objets ou de la caméra ne génère pas de flou. Ces avantages en font des caméras à grand potentiel pour la robotique et la vision par ordinateur dans des scénarios qui mettent en difficulté les caméras traditionnelles, notamment dans les cas de vitesses élevées et/ou de plages dynamiques élevées. Cependant, leur utilisation nécessite de repenser et de réadapter les algorithmes classiques.

2. Objectifs

Les objectifs du travail sont les suivants :

- Comprendre les travaux existants sur les modèles décrivant l'attention humaine en situation de conduite.
- Comprendre le déroulement d'une chaîne de traitement « attentionnelle » basée sur une double approche *bottom-up* (partant des primitives visuelles de bas niveau) et *top-down* (dépendant de la tâche en cours ou visée par l'utilisateur). L'approche bottom-up est fondée sur la construction de cartes de saillance bas-niveau basées position, fréquences spatiales, couleur, orientations, texture et mouvement. Pour l'approche top-down, les techniques d'apprentissage par réseaux de neurones sont envisagés pour détecter les objets présents dans les scènes routières et les étiqueter sémantiquement (marquages, panneaux, piétons, végétation, route, etc.).
- Réaliser une campagne d'acquisition de données provenant de caméras visuelles et événementielles dans des scénarios de conduite réalistes.
- Mettre en œuvre un algorithme de détection de plans saillants à partir d'une caméra évènementielle embarquée.

3. Profil recherché

Vous êtes curieux, créatif autonome et rigoureux. Votre intérêt pour l'innovation, la recherche en vision par ordinateur en traitement d'images vous caractérise. Votre intérêt se porte sur les problématiques de recherche ouvertes au carrefour entre plusieurs disciplines (neurosciences, physiologie, informatique).

4. Formation académique recherchée

- Master/ingénieur avec des compétence en vision par ordinateur et en traitement d'images.
- Langages : C, C++

- Anglais : bon niveau.

5. Modalités de candidature : Les candidatures (CV + lettre de motivation) sont à envoyer directement par email au contact ci-dessous. La transmission de relevés de notes sera appréciée.

Durée du stage: 6 mois **Date de début de stage** : Février ou mars 2022

Encadrement et contact :

- Fabien Bonardi (Maître de conférences, IBISC), fabien.bonardi@ibisc.univ-evry.fr
- Samia Bouchafa-Bruneau (Professeure, IBISC), samia.bouchafa@ibisc.univ-evry.fr

Références :

[1] Laurent Itti, Christof Koch and Ernst Niebur : A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence, 20(11):1254–1259, 1998.

[2] Ali Borji, Laurent Itti: Sate-of-the-art in visual attention modeling. IEEE Transactions on pattern analysis and machine intelligence, 35(1):185–207, 2013.

[3] Richard Veale, Ziad M. Hafed and Masatoshi Yoshida : How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. Philos Trans R Soc Lond B Biol Sci. 2017 Feb 19;372(1714).

[4] Viachaslau Kachurka, Kurosh Madani, Christophe Sabourin, Vladimir Golovko. Visual saliency based approach to object detection in computer vision systems: Real life applications. IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015.

[5] Tan Khoa Mai. Vers un système de vision artificielle opportuniste pour l'analyse de scènes complexes à partir de caméras embarquées. Thèse de L'université Paris-Saclay, le 13 décembre 2018.

[6] Benois-Pineau, J., & Mitrea, M. (2017, November). Extraction of saliency in images and video: Problems, methods and applications. A survey. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6).

[7] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco and Tobi Delbruck "Retinomorphic Event-Based Vision Sensors : Bioinspired Cameras With Spiking Output".

[8] Ryad Benosman et al. "Event-based visual flow". In :IEEE transactions on neural net-works and learning systems25.2 (2013), pp. 407-417

[9] Guillermo Gallego et al. "Event-based vision : A survey". In :arXiv preprint arXiv :1904.08405(2019).

[10] Bodo Rueckauer and Tobi Delbruck. "Evaluation of event-based algo- rithms for opticalflow with ground-truth from inertial measurement sen- sor". In :Frontiers in neuroscience10 (2016), p. 176.

[11] Elias Mueggler et al. "Lifetime estimation of events from dynamic vision sensors". In :2015 IEEE international conference on Robotics and Automation (ICRA). IEEE. 2015,pp. 4874-4881.