L'imagerie neuromorphique révolutionne la vision artificielle

Solution de rupture inspirée par le fonctionnement du cerveau humain et dédiée aux systèmes de vision artificielle et de cybervision, la technologie neuromorphique de vision pilotée par les événements qu'a développée la start-up française Chronocam, créée en 2014, arrive aujourd'hui à maturité. La jeune pousse qui cible plusieurs applications allant du domaine biomédical et scientifique à l'industrie et l'automobile explique pour L'Embarqué les principes de sa technologie.

AUTEURS

Christoph Posch, cofondateur et CTO, et Thomas Finateu, responsable du développement des capteurs, Chronocam.

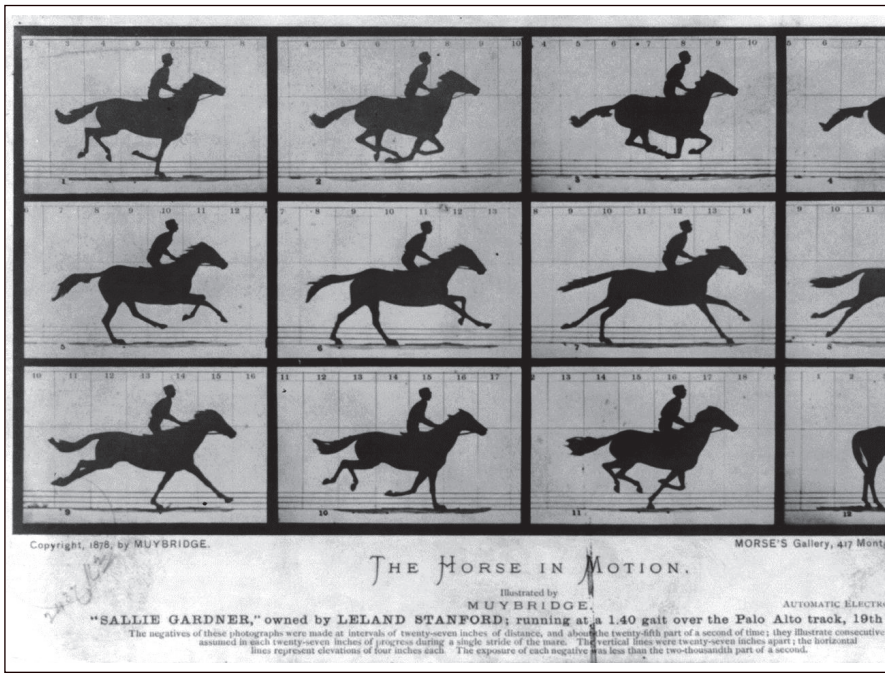
Le mode d'exploitation utilisé par les capteurs d'images actuels n'a d'intérêt que dans une seule perspective : réaliser des photographies, par exemple, pour immortaliser une scène fixe telle qu'un coucher de soleil resplendissant ou un paysage de montagne enchanteur. L'exposition d'une matrice de pixels au flux lumineux provenant de la scène à restituer, et ce pendant une période donnée, est la procédure type pour capturer un contenu visuel. Ce genre d'image consiste en un instantané, pris à un moment précis, et ne contenant aucune information dynamique. Pourtant cette méthode d'acquisition des informations visuelles est présente dans presque tous les systèmes de vision artificielle conçus pour la capture et la compréhension de scènes dynamiques. Malgré la multiplication des applications de vision intégrées sur des plates-formes diverses et variées, depuis les appareils mobiles intelligents jusqu'aux objets en mouvement comme les voitures, les avions ou les drones, les systèmes actuels de vision artificielle sont tous construits autour de ces instantanés qui apparaissent de moins en moins performants au regard des tâches qui leur sont demandées et qui sont, elles, de plus en plus nombreuses et de plus en plus complexes.

Un paradigme universel… erroné

Au 19e siècle déjà, Eadweard Muybridge avait eu l'idée d'installer une rangée d'appareils photo pour capturer les mouvements d'un cheval (photo A). Depuis ce temps, nombre de personnes sont persuadées que la capture rapide d'une succession d'images est un moyen efficace pour capturer le mouvement visuel. Cette conviction semble être confortée par le mode d'enregistrement et le rendu des films, du point de vue d'un spectateur humain. L'observation d'un mouvement en apparence fluide et continu à partir d'une certaine cadence d'images est toutefois plus liée aux perceptions de l'œil et du cerveau humain qu'à la qualité d'acquisition et de codage des informations visuelles sous forme de séries d'images fixes.

Dès qu'un changement de mouvement se produit, ce qui est le cas pour toutes les applications de vision artificielle dynamique, le paradigme universel reposant sur l'acquisition visuelle d'images est toutefois erroné. Lorsqu'une caméra observe une scène dynamique, quelle que soit la fréquence d'acquisition des images, elle n'est pas représentative de la réalité. En effet, les différents éléments d'une scène comportent généralement des contenus dynamiques différents. Par conséquent, un taux d'échantillonnage unique gouvernant l'exposition de l'ensemble des pixels d'une matrice d'imagerie ne pourra pas assurer l'acquisition adéquate des différentes dynamiques à l'œuvre en même temps.

Prenons par exemple une scène naturelle avec un sujet en déplacement rapide devant un fond statique comme un joueur de base-ball. Lors de l'acquisition de ce type de scène avec une caméra vidéo conventionnelle, le flou dû au mouvement de l'objet et à son déplacement d'une image à l'autre entraîne un sous-échantillonnage du mouvement rapide du bras et de la balle. D'autre part, l'acquisition répétée d'un fond statique génère des quantités énormes de données redondantes



● B.- Lors de l'acquisition de ce type de scène avec une caméra vidéo conventionnelle, le flou dû au mouvement de l'objet et à son déplacement d'une image à l'autre entraîne un sous-échantillonnage du mouvement rapide du bras et de la balle. D'autre part, l'acquisition répétée d'un fond statique génère des quantités énormes de données redondantes n'apportant aucune information nouvelle. Par conséquent, la scène est sous et sur-échantillonnée en même temps !

n'apportant aucune information nouvelle. Par conséquent, la scène est sous et sur-échantillonnée en même temps (photo B) ! Aucune autre solution n'est possible tant que l'ensemble des pixels du capteur d'image partage la même source de synchronisation et que cette dernière contrôle leur exposition.

La conclusion évidente de cette observation est que la vision artificielle, dans la plupart des cas, doit se satisfaire d'un mélange de données inutiles et de données de mauvaise qualité ! (Ces données seront soit inutiles soit mauvaises selon la cadence choisie et l'information de la scène.) Pour assurer des résultats exploitables, les fabricants de systèmes de vision artificielle doivent donc investir en permanence dans des solutions de traitement complexes, gourmandes en puissance et en ressources pour compenser une acquisition insuffisante. Cette approche, basée sur la puissance brute, n'est toutefois plus adaptée aux nouvelles tâches exigeantes de vision. Celles-ci requièrent en effet une compréhension des scènes en temps réel. Elles demandent par ailleurs un traitement visuel dans des environnements disposant de peu de puissance électrique, d'une bande passante réduite et de ressources informatiques limitées tels que des plates-formes mobiles avec batterie, des drones ou des robots. Toutefois, compte tenu de l'absence d'alternative convaincante, la stratégie d'acquisition d'informations visuelles dynamiques, à l'efficacité limitée, a fini par être adoptée par le marché de la vision artificielle. Elle est intégrée dans presque toutes les applications de vision depuis des décennies. Mais tout cela va changer !

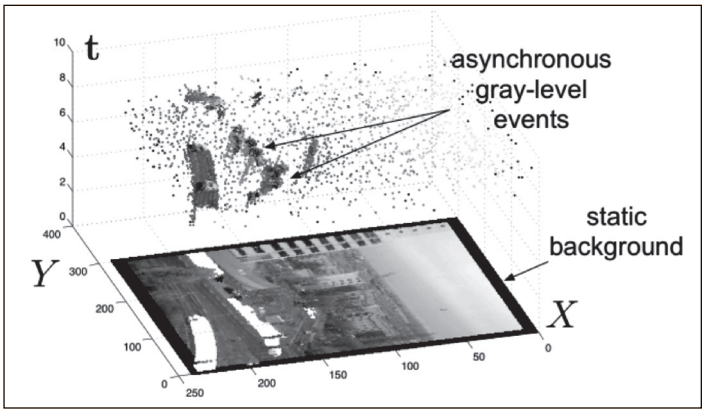
● A.- Depuis le XIXe siècle, nombre de personnes sont persuadées que la capture rapide d'une succession d'images est un moyen efficace pour capturer le mouvement visuel.

L'échantillonnage par franchissement de niveaux à la rescousse

Dans un scénario idéal, un capteur d'image doit échantillonner les différentes parties de la scène englobant les mouvements et changements rapides avec des taux élevés, tandis que les zones à changements lents sont capturées plus lentement, toujours en même temps. Les zones sans changement, quant à elles, ont un taux d'échantillonnage nul. Pour procéder ainsi, il n'est plus possible d'utiliser un taux d'échantillonnage unique (= cadence) pour l'ensemble des pixels. Par ailleurs, dans la mesure où les informations sur les emplacements et la vitesse de ces changements ne sont généralement pas connues d'avance, il faudrait autant de taux d'échantillonnage que de pixels sur le capteur, et adapter la fréquence d'images de chaque pixel à la scène observée.

Idéalement il faudrait abandonner la cadence commune et laisser chaque pixel adapter son propre taux d'échantillonnage en fonction de l'entrée visuelle reçue. Ainsi, chaque pixel définit le timing de ses propres points d'échantillonnage en réponse à son entrée visuelle, en réagissant aux changements quantitatifs de lumière incidente. Par conséquent, l'ensemble du processus d'échantillonnage n'est plus gouverné par une source fixe (artificielle) de timing, mais par le signal à échantillonner lui-même, ou plus précisément par les variations du signal en amplitude et dans le temps. La sortie générée par cette caméra n'est plus une séquence d'images, mais un flux constant dans le temps de données de pixels individuels, généré et transmis de façon conditionnelle et pertinente, en fonction des modifications de la scène (photo C).

L'échantillonnage d'amplitude, également appelé échantillonnage par franchissement de niveaux, a fait par le passé l'objet de recherches, notamment concernant les signaux unidimensionnels tels que les signaux audio. De nos jours, ce paradigme d'échantillonnage intervient dans l'acquisition en temps réel de données d'image en deux dimensions. En puisant dans deux décennies de recherche dans le domaine de l'ingénierie neuromorphique, nous avons développé un capteur d'images contenant une matrice de pixels fonctionnant de manière autonome et combinant un détecteur asynchrone par franchissement de



● C.- La sortie générée par une caméra à technologie neuromorphique n'est plus une séquence d'images, mais un flux constant dans le temps de données de pixels individuels, généré et transmis de façon conditionnelle et pertinente, en fonction des modifications de la scène filmée.

niveaux et un circuit distinct de mesure de l'exposition. Toute mesure d'exposition à un pixel est déclenchée par un événement de franchissement de niveaux (photo D).

Par analogie au processus biologique, chaque pixel de ces capteurs optimise son propre échantillonnage en fonction des informations visuelles perçues ! Si les choses changent rapidement, le pixel échantillonne à un taux élevé, et si rien ne change, il cesse l'acquisition de données (celles-ci étant alors perçues comme redondantes). Il se met en veille jusqu'à ce qu'un nouvel évènement se produise dans son champ de vision. De plus, chaque pixel échantillonne de façon indépendante son éclairement lors de la détection d'un changement d'une certaine magnitude dans cette même luminance. Ceci lui permet d'établir instantanément son niveau de gris, dès la modification. Le résultat de la mesure d'exposition (par exemple le nouveau niveau de gris) est sorti de façon asynchrone du capteur en même temps que les coordonnées du pixel dans la matrice du capteur. Par conséquent, les informations visuelles ne sont pas acquises et transmises image par image, mais en continu et de façon conditionnelle. Seuls les éléments de la scène qui comportent de nouvelles informations visuelles sont transmis. Ou, en d'autres mots, seules les informations pertinentes, car inconnues, sont acquises, transmises, stockées et enfin traitées par des algorithmes de vision artificielle.

Dans la mesure où le fonctionnement des pixels est désormais asynchrone et sachant que les circuits de pixels sont conçus pour réagir très rapidement, l'acquisition de données fortement redondantes et inutiles par sur-échantillonnage est éliminée. Il en est de même pour le sous-échantillonnage de la dynamique des scènes rapides, du fait que le pixel asynchrone n'est plus lié à la limitation de la résolution temporelle par une cadence fixe et lente. L'acquisition de pixels et des temps de lecture de l'ordre de la milliseconde, voire de la microseconde, sont ainsi possibles. Cette nouvelle approche permet de générer des résolutions temporelles équivalentes à celles de capteurs conventionnels, s'exécutant à des cadences rapides de plusieurs dizaines ou centaines de milliers d'images par seconde. Elle ouvre des possibilités très intéressantes pour la vision artificielle. Désormais, et pour la première fois, le compromis strict entre la résolution temporelle et le débit binaire, qui limite l'acquisition sur les systèmes reposant sur des images, peut être surmonté. Dans cette nouvelle configuration, la résolution temporelle du processus d'échantillonnage des données d'image n'est plus gouvernée par une horloge fixe pilotant tous les pixels. Par conséquent, le volume de données de la sortie du capteur, indépendamment de la résolution temporelle disponible pour l'acquisition au niveau des pixels, dépend uniquement du contenu dynamique de la scène visuelle. L'acquisition de données visuelles devient ainsi à la fois rapide et concise, ouvrant la voie à une acquisition ultra-rapide et à une consommation réduite en matière de ressources, de bande passante nécessaire pour la transmission et de mémoire.

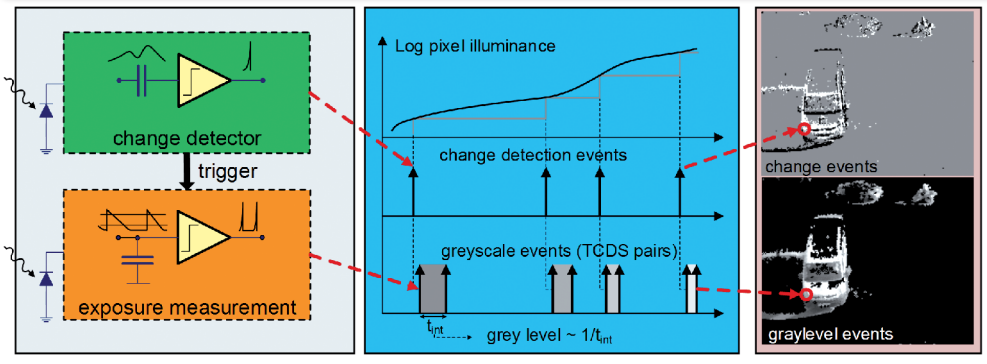
Des paradigmes du traitement de la vision repensés

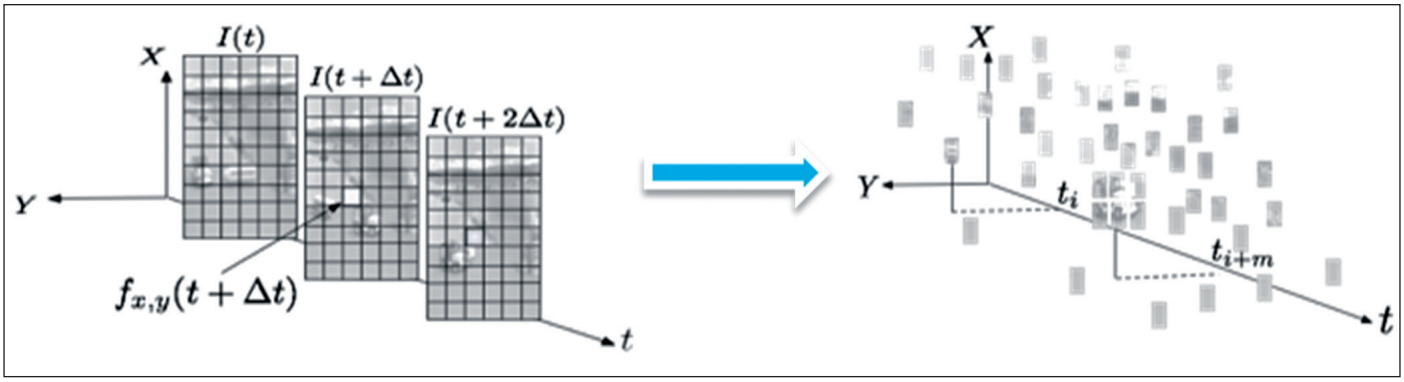
L'avantage de ce type de traitement des informations visuelles dynamiques ne se limite pas au capteur. Afin de libérer totalement le potentiel de ces capteurs de vision sensibles aux événements, les paradigmes du traitement de la vision doivent être repensés de fond en comble.

Tout d'abord, la notion d'image comme base du traitement de la vision est complètement abandonnée. Dans la mesure où les capteurs codent la dynamique visuelle sous forme de schémas spatio-temporels à forte résolution d'événements (représentant les aspects voulus de la dynamique de la scène tels que des contours d'objets mobiles, des trajectoires, la vélocité, etc.), les algorithmes de traitement s'appliquent désormais à des événements et à des caractéristiques temporels, et non plus à des images statiques distinctes. Les formules mathématiques décrivant ces caractéristiques dans l'espace et le temps sont simples et élégantes. Elles permettent de produire des algorithmes ultra-efficaces et des règles de calcul permettant d'exécuter en temps réel des systèmes de traitement sensoriels, tout en limitant au strict minimum le recours aux ressources de calcul. Ceci a pour effet de réduire la consommation électrique. Voyons pourquoi.

La résolution temporelle très fine du processus d'acquisition et de codage, combinée à la concision des données du capteur, permet pour la pre-

● D.- Chronocam a développé un capteur d'images contenant une matrice de pixels fonctionnant de manière autonome et combinant un détecteur asynchrone par franchissement de niveaux et un circuit distinct de mesure de l'exposition. Toute mesure d'exposition à un pixel est déclenchée par un événement de franchissement de niveaux.





● E.- La résolution temporelle très fine du processus d'acquisition et de codage, combinée à la concision des données du capteur, permet pour la première fois de prendre véritablement en compte les volumes espace-temps continus pour les calculs visuels

mière fois de prendre véritablement en compte les volumes espace-temps continus pour les calculs visuels (photo E). Dans ces espaces multidimensionnels, les schémas spatio-temporels des données visuelles présentent un niveau élevé d'orthogonalité entre des caractéristiques d'apparence similaire (lorsqu'elles sont représentées par des trames d'images statiques). Ces dernières peuvent être exploitées notamment au moyen du traitement simultané de nuages d'événements dans les domaines fréquentiel et temporel. Les objets mobiles décrivent des plans continus dans le volume espace-temps, ce qui donne un accès direct à la vitesse du mouvement de l'objet et autorise des prédictions fiables des positions futures, ainsi que la possibilité de séparation en cas d'occlusions.

L'utilisation d'informations temporelles avec une résolution fine n'est pas une nouveauté dans le domaine de la vision artificielle. La nature code des informations par ordonnancement de « pics » d'amplitude uniforme et base ses calculs uniquement sur leurs propriétés temporelles et les temps d'arrivée. Récemment l'étude du calcul des systèmes biologiques a révélé une approche totalement nouvelle pour les problèmes de vision artificielle. L'une des découvertes les plus importantes a été que plusieurs problèmes mathématiques mal posés peuvent être réécrits de façon intelligente dans le cadre du calcul de temps. En appliquant ces algorithmes mathématiques à des sorties basées sur les événements en termes de pixels perçus par nos capteurs de vision à échantillonnage automatique, nous avons pu démontrer que les restitutions de scène visuelles en temps réel peuvent atteindre des résolutions temporelles de dizaines, voire de centaines de kHz, là où l'imagerie conventionnelle atteint difficilement 30 ou 60 Hz. Ces résultats démontrent la force d'un calcul de vision basé sur les événements et inspiré par la biologie. Celui-ci replace les problèmes de vision dans un cadre mathématique incrémental, de sorte que chaque événement visuel provenant d'un pixel du capteur produit un petit calcul qui consommera très peu de ressources.

Le fait de privilégier la dimension temps (événements) aux informations spatiales des images statiques a pour avantage que tout calcul reposant sur des événements temporels précis peut être exprimé sous forme de coïncidences temporelles, par analogie avec le mode de calcul qu'utilisent les neurones biologiques. Nos capteurs codent l'évolution temporelle et la luminance absolue de façon individuelle sur chaque pixel lors de la synchronisation des événements. Ceci permet d'exprimer des calculs complexes de corrélation sous forme de simples ensembles de coïncidences entre les pixels. Il en va de même pour la reconstitution de la profondeur 3D des systèmes de stéréovision à plusieurs caméras. Ces opérations complexes de concordance sont transposables en de simples détections de coïncidences des activations de pixels sur différentes caméras. En effet, étant donné la forte résolution temporelle des capteurs, si deux pixels de deux caméras émettent des événements en même temps, il est probable qu'ils observent le même point de la scène. De plus, des résultats récents indiquent que l'utilisation d'informations temporelles pour l'évolution temporelle de l'éclairement sur un pixel unique permet de repenser plusieurs problèmes réputés difficiles de vision en temps réel, tels que les algorithmes de localisation et de cartographie simultanée (SLAM), la reconnaissance d'objets ou l'évitement des obstacles.

L'avènement de la vision basée sur des événements

L'acquisition artificielle de données, pertinentes ou non, selon une synchronisation choisie au hasard (mais bien souvent lente, 30 ou 60 images/s) par rapport au signal à acquérir, limite de façon considérable les systèmes actuels de vision artificielle. L'avènement de la vision reposant sur des événements – qui est inspirée du modèle biologique – amorce un changement de paradigme pour les systèmes de vision artificielle reposant sur l'acquisition et le traitement asynchrones des informations visuelles allant au-delà de Nyquist. Des résultats publiés récemment démontrent les avantages de l'approche neuromorphique par rapport à l'approche traditionnelle par image sous de nombreux aspects. Elle permet notamment de bénéficier d'une vitesse accrue, d'une plage dynamique ultra-élevée, d'une réduction des coûts de calcul et d'une plus grande robustesse. Par conséquent, les tâches de vision artificielle exigeantes, qui présentaient jusqu'à présent un coût prohibitif en raison de leur forte consommation en matière de ressources, comme la cartographie 3D en temps réel ou le suivi complexe de plusieurs objets, deviennent désormais possibles. D'autres types de tâches seront également désormais plus accessibles comme les boucles rapides de feedbacks visuels pour les fonctions sensorielles et motrices exécutées à des cadences en kilohertz sur du matériel de traitement bas de gamme avec batterie et entrées visuelles « toujours actives », ou comme les interactions avec l'utilisateur ainsi que la prise en compte du contexte environnemental sur des appareils mobiles intelligents. ■