Vidéo Protection Intelligente et IA

P. Bernas, G. Née, EVITECH SAS.

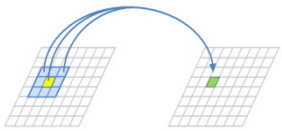
Considérations scientifiques, techniques, économiques, et juridiques

Dans cet article, nous examinons les impacts de l'IA sur les applications d'**analyse intelligente d'images pour la sécurité globale**, à partir d'images de caméras *CCTV* (vidéo-protection) couleur et thermiques dont nous sommes spécialistes à EVITECH depuis 2005[1]. Ces applications d'analyse d'images visent à détecter des situations dangereuses[2] pour prévenir ou limiter leurs conséquences : intrusion dans un site sensible, comportement individuel dangereux, détection d'une fuite d'hydrocarbure ou d'une fumée, maîtrise de la foule (comptage, flux, incidents), présence d'objet suspect ou dangereux, etc. Après une présentation générale de l'histoire de la discipline, nous identifions les apports les plus pertinents de l'IA pour l'analyse vidéo, en termes théoriques, mais aussi en termes pratiques (l'IA ayant facilité l'analyse dans l'espace public), les capacités des opérateurs de vidéosurveillance n'étant pas illimitées ni en termes d'énergie consommée, ni en prix. Nous ne pouvons conclure sans aborder la question légale : la Loi limite ces usages dans l'espace public.

1- L'impact de la science de l'IA sur le traitement d'images en général

1.1 – La discipline

Le traitement d'images est une discipline qui a émergé avec l'image numérique, dans les années 1980[3] (après que des précurseurs américains en aient posé les bases dès l'apparition des ordinateurs dans les années 60), et qui a progressé jusqu'en 2005 par ajout de techniques algorithmiques et mathématiques successives permettant d'analyser et de discerner de mieux en mieux les propriétés des images et donc celles de la vidéo : morphologie mathématique[4], filtrage, statistiques, calculs de gradients et recherche de points caractéristiques, histogrammes...



C'était une époque d'intuitions, de tâtonnements, de formulations mathématiques d'impressions que l'expert cherchait à automatiser dans un algorithme : « *le traitement d'images était dirigé par l'Homme* »[5]. Les réseaux neuronaux classiques, exploitant l'apprentissage, végétaient à l'époque autour des capacités de reconnaissance de caractères[6].

---

[1] « Intérêts et limites de la vidéo-surveillance intelligente pour la Sécurité Globale »,P. Bernas, Conf. AVIRS 2008.

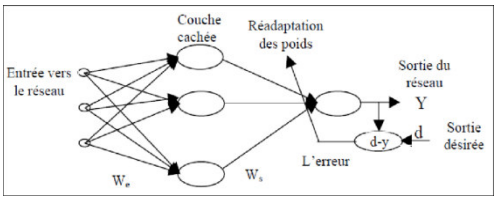
[2] Dans la photo ci-contre, une personne portant une arme (source Evitech Lynx).

[3] « Vision », David Marr: A Computational Investigation into the Human Representation and Processing of Visual Information, 1982, Editions Freeman, ISBN 978-0716712848.

[4] "*Image Analysis and Mathematical Morphology*", Jean Serra, G. Matheron, vol. 1, Academic Press, Londres, 1982, (ISBN 0-12-637242-X)

[5] Roger Mohr (1947-2017 †), ex-Directeur du laboratoire Inria GRAVIR et de l'Ensimag, et cofondateur d'Evitech.

[6] Le premier réseau convolutionnel Neocognitron (Kunihiko Fukushima) date de 1979, il est entraîné à reconnaître des motifs visuels. Il constitue encore la base des réseaux récents : K. Fukushima et S. Miyake, « *Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition », in Competition and Cooperation in Neural Nets*, Berlin, Heidelberg, 1980, p. 267-285.



La stratégie d'entraînement des réseaux a toutefois a bien évolué grâce à la « *back propagation* » (utilisation de l'erreur pour corriger l'apprentissage). Elle est utilisée pour l'entraînement des réseaux neuronaux à partir de 1985[7]. En 1989, Yann Lecun est le premier à mettre en application un réseau combinant la « *back propagation* »[8] et un réseau de neurones convolutionnel appliqué à la reconnaissance de caractères manuscrits[9].

En parallèle, les techniques d'apprentissage ont été développées comme une alternative automatique, avec notamment l'Analyse en Composantes Principales[10] et les *Support Vector Machines*[11]. S'appuyant sur des descripteurs issus du traitement classique (points caractéristiques, HOGs, …) elles ont montré des capacités de progrès intéressantes, mais sans parvenir à révolutionner le domaine, car les gains obtenus n'étaient pas spectaculaires.

C'est aussi vers 2005 que la puissance des processeurs a permis de décoder la vidéo numérique sans nécessairement utiliser un processeur parallèle, et de passer du traitement des photos à celui de la vidéo en temps réel. Aujourd'hui, un ordinateur d'un coût de l'ordre de 1000 €[12], consommant 150 W, peut décoder simultanément 5 à 10 flux vidéo, et leur appliquer un traitement algorithmique classique, ce qui place le coût matériel de ce traitement (hors coût du logiciel) dans une fourchette de 100 à 200 € par flux, avec une énergie consommée de 15 à 30 W par flux. De plus, ces traitements peuvent opérer de façon adaptée sur des parties de l'image, à différentes échelles de taille.

Autour de 2013, l'émergence des réseaux neuronaux profonds[13] et des réseaux neuronaux convolutionnels[14], en conjonction avec l'apparition de cartes graphiques puissantes, a permis de briser d'un seul coup le plafond de verre des performances en matière de reconnaissance d'objets dans des images, qui stagnait depuis plusieurs années.

En particulier, AlexNet[15] permet une progression très significative sur une compétition de reconnaissance d'objets dans des photographies qui faisait référence sur la base d'images couleur ImageNet[16] (14 millions d'images) depuis des années.

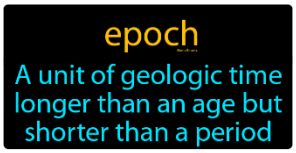
1.2 – Les réseaux neuronaux

Tout commence avec le *neurone*, sorte de micro-entité de calcul élémentaire possédant un poids et capable de communiquer avec ses voisins, lors des opérations d'apprentissage et de calcul. La

[7] « Learning Internal Representations by Error Propagation », 1985.
[8] Y. LeCun et al., « *Handwritten Digit Recognition with a Back-Propagation Network* », in Advances in Neural Information Processing Systems, 1989, vol. 2.
[9] Source du schéma : https://www.tellaw.org
[10] Jean-Paul Benzécri ; Analyse des données, Dunod, 1973.
[11] C. Cortes et V. Vapnik, « *Support-vector networks* », Mach Learn, vol. 20, n° 3, p. 273-297, sept. 1995.
[12] Processeur Intel core i7 12700, RAM 16 GB, en Oct 2022.
[13] Ronan Collobert et Jason Weston, « *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning* », Proceedings of the 25th International Conference on Machine Learning, New York, NY, USA, ACM, iCML '08, 2008, p. 160–167
[14] « Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation », DeepLearning 0.1, LISA Lab (Août 2013).
[15] Krizhevsky A, Sutskever I, Hinton G. *ImageNet classification with deep convolutional neural networks*. Communications of the ACM, 2017. 60(6): p. 84-90
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et L. Fei-Fei, « *ImageNet: A large-scale hierarchical image database* », in 2009 IEEE Conference on Computer Vision and Pattern Recognition, juin 2009, p. 248-255.

démarche générale consiste alors à structurer un réseau neuronal profond convolutionnel en entrées, sorties, et constitué d'un certain nombre de couches de neurones (couches de pooling, couches totalement connectées, couches de convolution), liées par des opérateurs (opérateurs mathématiques, et opérateurs de convolution, notamment, impliquant des quantités de neurones).
L'apprentissage fait circuler des données annotées (images ou autres) dans ce réseau, en appliquant ces opérateurs, et en rétro-propageant finement par de petites modifications successives, les poids des neurones, en fonction des résultats attendus à chaque donnée (les annotations). Ce procédé correctif se nomme une étape d'apprentissage. Un passage complet de la base d'exemples d'entraînement (un *dataset*) sur le réseau se nomme une *epoch*.

En parallèle de cet apprentissage, on utilise un second ensemble de données annotées, les données de tests, pour mesurer le score de reconnaissance du réseau sur ce second ensemble. Pour ne pas entacher le processus et risquer le sur-apprentissage, ces données doivent impérativement être réservées à la mesure de la performance du réseau entraîné.
Lorsqu'après un passage alterné de nombre d'*epochs* et de tests, le réseau devient optimal (il cesse de s'améliorer d'une *epoch* à la suivante), ou bien lorsqu'il commence à régresser, on s'arrête.



On s'aperçoit alors que ces réseaux élaborent automatiquement, dans leurs données intermédiaires, des descripteurs bien plus nombreux et riches que ceux que l'on cherchait à utiliser dans le traitement d'images classique.

Ce processus d'apprentissage modifie à chaque étape les poids des neurones du réseau de façon progressive, par de petites variations, image après image. On peut imaginer une notion de distance et de continuité entre deux mêmes réseaux, à partir de valeurs très proches de leurs poids. L'espace continu défini par tous ces réseaux possibles, avec toutes les valeurs possibles de poids, pourrait donc être vu comme une surface courbe dont les creux seraient les optima, pour la reconnaissance des objets visés. Dans ce paradigme, le processus d'apprentissage ferait « descendre » le réseau selon une convergence vers le creux le plus accessible en suivant la ligne de pente. Il est possible, lors d'un entraînement, de buter sur un minimum local, qui n'est pas un optimum absolu. Par des méthodes qui bousculent assez radicalement les poids des neurones, à l'initialisation, on peut aussi changer de point de départ, recommencer l'entraînement, et cette fois-ci essayer de converger vers un autre minimum, ailleurs, que l'on peut espérer être plus performant. Parfois, on tombe sur le réseau nul (poids à zéro partout).

Ce constat donne à la discipline un caractère expérimental qui peut sembler un peu choquant pour un scientifique, mais qui pour le moment est le lot commun des acteurs du domaine.

Un autre sujet qui possède également une forte dimension expérimentale est la constitution de la base d'images d'apprentissage : quelles images sélectionner et comment les annoter. Tout un ensemble de règles ont émergé de l'expérience progressive de l'apprentissage pour définir les bonnes conditions de constitution d'une base d'exemples pertinente pour l'apprentissage. Certaines ont été automatisées, comme l'inversion miroir des images, leur rotation, ou le redimensionnement, pour augmenter artificiellement le nombre d'exemples, au travers d'une discipline qui se nomme, à juste

titre, l'augmentation de données[17]. D'autres sont encore l'objet d'une expertise empirique : la variété des positions, des arrières plans, des éclairages est recherchée pour équilibrer un *dataset*. Les tentatives d'utiliser de fausses données[18] à l'entraînement (copier et insérer des imagettes « rares » d'objets dans une photo, pour essayer d'en multiplier les exemples) ont montré qu'elles ne conduisaient qu'à des détecteurs de fausses données[19].

Une fois la configuration du réseau obtenue, le fait de lui passer une image et d'en obtenir un résultat (par exemple des boîtes avec tous les objets reconnus) s'appelle un *forward*[20].

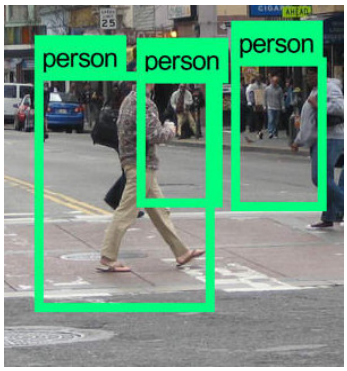
Le temps d'entraînement d'une *epoch* multiplie le temps de base du *forward* par le nombre de cas d'apprentissage (par exemple, cent mille, ou un million). Le nombre d'*epochs* nécessaire pour faire converger un réseau peut être important (des centaines). Le temps total d'entraînement est le produit, c'est à dire jusqu'à des dizaines ou centaines de millions de *forwards*[21]. Toutefois, ce temps n'est normalement engagé qu'une fois, pour élaborer le réseau.

A partir de cette invention, extraordinaire par les progrès atteints en termes de performances, la plupart des problèmes connus de traitement d'images ont été revisités avec ces techniques et ont connu des solutions dont les résultats étaient bien souvent meilleurs que par les approches précédentes.

2- Les opportunités offertes par l'IA sur l'image

2.1- Reconnaissance d'objets, personnes, visages, …

La première application de traitement d'images, la plus médiatisée, dont l'IA a nettement amélioré les performances, est la reconnaissance d'un objet en gros plan dans une image ou dans une photographie couleur. En entraînant un réseau neuronal profond convolutionnel à partir d'une grosse base d'images annotées (on parle de centaines de milliers, ou de millions), il est aujourd'hui possible de produire un algorithme de classification qui atteint des performances supérieures à 90% de reconnaissance de tels objets, dans un domaine d'usage réaliste (gros plan, éclairage, distance, netteté, …).



C'est le besoin des détecteurs embarqués dans les voitures, dont on s'attend à ce qu'ils aient une très grande fiabilité : reconnaître les objets immédiatement devant lui (ex : une personne qui traverse). Les progrès sont alors tels qu'ils permettent de dépasser le taux de reconnaissance d'un observateur humain[22].

On distingue ici la classification de photo (déterminer l'objet principal qui constitue la photo) et la détection (ou reconnaissance), qui consiste à identifier tous les objets reconnaissables[23] présents, même à l'arrière-plan de la photo.

---

[17] *A survey on Image Data Augmentation for Deep Learning,* Connor Shorten, Taghi M. Khoshgoftaar, in Journal of big data n°6, article n°60, 2019.

[18] Ici, un faux char.

[19] F. Jurie, sur la détection de chars depuis une vue satellite, par exemple.

[20] En français : une prédiction.

[21] En pratique, des jours, des semaines.

[22] Comparing Object Recognition in Humans and Deep Convolutional Neural Networks—An Eye Tracking Study, L van Dyck, R Kwitt, S Denzler, W. Gruber, Oct 2021, Frontiers in Neuroscience.

[23] Ici, YOLO : la personne en gris au premier plan est détectée, mais ni les voitures derrière à gauche, ni correctement chacune des autres personnes au fond. La détection de *buildings* dépend de l'existence de la classe *building* dans l'apprentissage.

Ainsi, dans le cas de la détection de visages, sur une caméra placée au-dessus d'un point de passage assez fréquenté, on atteint aujourd'hui des scores de plus en plus proches de 100%, la distance de reconnaissance augmentant au fil des années (donc pour des visages de plus en plus éloignés – petits– dans l'image).

Mais le taux de reconnaissance des objets à l'image est toujours relatif à une certaine base d'exemples annotés (le *dataset*), dont il faut examiner avec soin la résolution, l'éclairage, la qualité pour pouvoir en tirer des conclusions sur un usage ultérieur. Un détecteur performant et rapide aujourd'hui (YOLO v5 ici) atteint une performance de reconnaissance de **voitures** (*recall*) comprise entre 15 et 64% pour une précision (absence de fausses détections) de 25 à 57% sur un sous-ensemble de la base d'images couleur *PascalVOC* (Test Evitech 2022). Cette base comporte aussi des objets en second plan, des objets multiples, avec des recouvrements partiels, ce qui complique la reconnaissance et diminue les taux. Comme on le constate, il reste une large marge de progrès.

Sur cette même application, d'autres approches ont été et sont massivement investiguées pour permettre une reconnaissance, même lorsqu'on dispose de peu de données (par exemple seulement des dizaines ou des centaines d'images). Parmi celles-ci, on citera l'approche du *fine tuning* : on utilise au départ un réseau neuronal initialisé sur les millions d'images de classes générales annotées d'un grand *dataset* (on dispose d'un réseau qui reconnaît bien les objets de ce *dataset*), que l'on l'entraîne ensuite à apprendre une ou des classes très spécifiques que l'on recherche (une nouvelle classe, jusqu'alors inconnue[24]), sur les données spécifiques dont on dispose, en moindre nombre (dizaines, centaines). Il peut converger ou se spécialiser sur un système de reconnaissance assez performant : les scores de reconnaissance des nouvelles classes sont seulement légèrement inférieurs à ceux de l'entraînement complet.

En termes de temps d'exécution, la reconnaissance d'objets dans une image est en quelque sorte l'unité de base du traitement d'images par IA. Un forward sur une image 500x500 sur une carte GPU moyenne, d'un coût d'environ 1000 €[25], dotée de 4 à 6 GB de RAM, consommant autour de 290 W, mobilise la carte entre 10 et 30 millisecondes[26]. En dépit des aléas de disponibilité des cartes graphiques sur l'année 2021 et début 2022, la société NVIDIA est à cet égard celle qui fournit pour le moment les solutions les plus efficaces en rapport performances / prix. Ce délai de 10-30 ms est à comparer avec la fréquence des images de la vidéo, de 25 à 30 images par seconde (ips), donc une nouvelle image toutes les 33 à 40 ms. Sur une toute petite installation avec 4 caméras, une carte pourrait traiter au mieux toutes les images des 4 caméras à 25 ips (algorithme le plus rapide sur 500x500 pixels : 25 x 4 = 100, et 100 x 10 ms = 1 sec), et au pire elle ne les traite qu'à 8 ips, donc environ une image sur 3, ce qui diminue nettement les performances d'une application d'analyse qui serait entièrement basée sur l'IA pour détecter et suivre les objets en mouvement dans une vidéo.

Or, il y a plus de 4000 caméras couleur dans le Plan de Vidéo Protection de Paris[27], et plus de 500 caméras couleur dans chaque grande gare ferroviaire européenne (il n'y a pratiquement pas d'usage des caméras thermiques en milieu urbain/transports[28]).

[24] L'INRIA a réalisé un démonstrateur de détection de sièges de télésiège de cette façon.
[25] Une carte NVIDIA RTX 3070 Ti - 8Gb en oct 2022, prix courant sur internet.
[26] Pour des réseaux de type SSD, YOLO. Ce temps est celui du forward, il suppose le réseau DNN déjà chargé dans la carte.
[27] Le PVPP à Paris : Image AFP.
[28] Il n'y a pas non plus de bases d'images annotées faisant référence, ni de statistiques de détection connues sur cette imagerie, qui est très peu adressée par l'IA, même si des usages fortuits de classification par IA (entraînée sur images couleur), et appliquée sur des images thermiques peuvent parfois révéler des classifications pertinentes.

Dans le domaine de la biométrie faciale, derrière la reconnaissance d'une forme de visage, on va comparer deux à deux des paires de visages pour identifier un individu. On développe des réseaux dédiés à une telle comparaison, en les entraînant sur des paires de visages de la même personne pris avec différents éclairages, à des âges différents, avec des coiffures variées, etc. Les applications conçues comme des cascades de traitement (détection de visages puis comparaison) peuvent mobiliser plusieurs réseaux successifs pour produire un traitement.

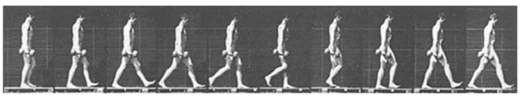
2.2- Postures

Un autre ensemble d'applications qui a été l'objet de recherches à succès porte sur la reconnaissance des parties du corps humain, ou encore la superposition dans l'image d'un squelette ou d'un modèle de corps humain.

Des réseaux comme OpenPose[29] ont permis de localiser toutes les parties des corps d'un groupe de danseurs dans une image. La technologie a été diffusée dans des vidéos devenues célèbres.

En termes de temps d'exécution, ces applications sont déjà 3 à 4 fois plus lentes que la reconnaissance d'objets. Un calcul de pose sur une image 500x500 sur la carte GPU moyenne citée plus haut, prend entre 70 et 120 millisecondes.

La reconnaissance d'une posture particulière d'une personne (chute au sol, pied en l'air) peut alors s'appuyer sur cette détection de posture pour reconnaître une attitude (le bras levé, par exemple, ou estimer l'angle entre les jambes) ou identifier des gestes particuliers (préhension d'un objet dans un rayon de magasin, puis mains regroupées au niveau du ventre pouvant suggérer un vol à l'étalage).



Dans le cas des gestes, on pourra aussi reconstituer une succession de positions et la comparer à des successions de positions qui auront été apprises comme étant à détecter : on en vient là à des solutions basées sur des doubles réseaux neuronaux, l'un pour élaborer les postures au fil des images, et constituer des blocs temporels (par exemple 50 postures successives de la même personne), et l'autre pour classifier les blocs temporels et reconnaître une gestuelle particulière, comme un coup de poing.

Le second réseau sera donc sollicité en continu sur une fenêtre glissante, ou ponctuellement par à-coups une fois toutes les secondes, ou toutes les deux secondes, par caméra (pour détecter sur un horizon de 50 images toutes les 25 à 50 images[30]), tandis que le premier réseau nécessitera en général entre une et trois cartes GPU pour traiter les 25 images par seconde d'une même caméra.

Si on augmente la puissance de la carte GPU pour utiliser les deux réseaux sur la même carte, la mémoire GPU nécessaire pour un traitement en temps réel est la somme de la mémoire des deux réseaux[31], qui doivent être tous deux chargés ensemble, et opérer de façon synchronisée avec les flux vidéo. Dans ce type de cas, on

[29] Z. Cao, T. Simon, S.-E. Wei, et Y. Sheikh, « Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields », in CVPR, 2017.

[30] Pour éviter une perte de détection qui serait due à une coupure au mauvais moment.

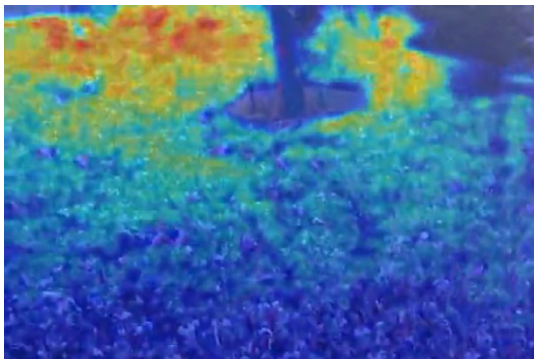
[31] Un réseau de détection requiert quelques GB, la carte GPU citée en référence en a 8.

voit que le besoin de ressources matérielles peut croître très rapidement, même pour simplement traiter un seul flux vidéo.

Quel opérateur CCTV, gérant des centaines de caméras, va pouvoir dédier 3 cartes GPU (3000 €, 900W) à chacune de ses caméras ?

2.3- Densité

L'approche la plus fructueuse, en terme d'estimation de la densité, et notamment dans les cas de densités élevées, consiste non pas à détecter chaque personne ou chaque tête, mais à dédier un réseau neuronal spécialisé à la reconnaissance de caractéristiques de têtes[32], afin de pouvoir prendre en compte les têtes partiellement cachées[33].



En termes de temps d'exécution, cette application est environ 2 fois plus lente que la reconnaissance d'objets. Une mesure de densité sur une image 500x500 sur une carte GPU moyenne, dotée de 6 à 8 GB de RAM, telle que vue précédemment, prend entre 20 et 60 millisecondes[34].

2.4 – Analyse du mouvement

L'analyse du mouvement relatif d'un ensemble d'objets et du fond de la scène dans une vidéo, qui a été l'objet de recherches intenses dans le domaine de l'analyse d'images appelé *flot optique*, a été complètement surpassé par le réseau FlowNet en 2015 et par ses successeurs depuis lors. En effet, les équations de mouvement basées sur la relation entre le contraste et le mouvement ne pouvaient pas toujours bien traiter le cœur des régions homogènes (sans contraste) : la segmentation des objets en mouvement manquait de précision. On peut aujourd'hui, avec ces techniques d'IA, segmenter quasiment au pixel près un objet qui bouge dans une vidéo, sans même déborder lorsque cet objet passe au-dessus d'une zone représentant la même couleur que son bord.
Mais, en termes de temps d'exécution, cette application est beaucoup plus lente que la reconnaissance d'objets. Le traitement d'une image 500x500 à la suivante, sur une carte GPU moyenne, telle que vue précédemment, est très au-delà du temps réel (donc très supérieur à 40 ms, pour une vidéo à 25 images par seconde).

2.5 – Mesures de différences entre deux images

Il y a encore de très nombreuses applications adressées par la discipline, dont par exemple la mesure de différences entre deux images[35], qui peut servir à des usages variés en sécurité : détecter un objet quelconque sur une piste d'aéroport, qui pourrait percer le fuselage d'un avion, mesurer la progression du déchargement d'un camion sur un quai logistique, détecter un changement au bord d'une route, ou un objet abandonné, etc. Dans ces applications, on cherche à détecter les différences, mais on ne sait pas en quoi elles consistent, donc on ne peut pas chercher une classe d'objets particulière.

Il n'y a en fait quasiment plus de sujet de détection ou d'analyse de l'image qui ne soit abordé par le prisme des réseaux neuronaux profonds, à tel point que certaines conférences internationales de la

[32] *Estimation of crowd density in surveillance scenes based on deep convolutional neural network*, Shiliang Pu, Tao Song, Yuan Zhang, Di Xie, 8th Intnl Conf on Advances in Info. Tech., in PCS 111, 2017, pp. 154-159.
[33] "*Peaceful Monitoring of Crowds*", Conférence WISG 2013, Troyes, P. Bernas, G. Née, P. Drabczuk.
[34] Ici, carte de chaleur de densité par Evitech Lynx.
[35] R. Daudt, B. Saux, et A. Boulch, Fully Convolutional Siamese Networks for Change Detection. 2018.

discipline éliminent d'emblée de leur sélection toutes les études reposant sur les techniques « classiques ».

2.6 – Autres applications

De nombreuses autres applications existent dont on n'a pas l'usage pour l'instant en sécurité. On citera par exemple la segmentation sémantique (classer les pixels selon la nature de l'objet auquel ils appartiennent, ce qui permet un détourage fin des objets au lieu d'une boîte englobante), les altérations d'images (améliorer la résolution)[36], ou la synthèse d'images[37], ou les détecteurs ouverts à l'apprentissage de nouvelles classes par quelques exemples saisis par l'utilisateur etc...

A un niveau plus global, on pourra aussi trouver des applications visant à traiter un problème sans le segmenter en sous-problèmes : conduire un véhicule à partir d'une vue d'une caméra frontale, détecter une anormalité dans une scène en général. L'approche est délicate, car le lien entre l'application et l'utilisation est direct, et on doit faire confiance à la variété de l'entraînement initial dont on espère qu'il couvrira tous les cas possibles. En pratique, des tests de ce type d'applications dans la sécurité ont montré qu'elles étaient performantes sur des vues rapprochées et des actions répétitives simples : interaction homme/automate par exemple. On n'a pas de bons retours sur des situations trop générales.

Elles sont donc peu utilisées dans le domaine de la sécurité, où on préfère une sémantique forte, s'appuyant sur des détecteurs avec des taux certifiés : un opérateur logique « ET » s'appliquant entre un détecteur d'objet qui serait fiable à 90%, et un estimateur de position lui-même fiable à 100% possède une fiabilité globale de 90%. Une telle détection faite par un réseau neuronal requerrait des quantités de tests de l'objet à différentes positions, qui pourraient être trop coûteuses à établir pour atteindre cette fiabilité.

2.7 – Analyse d'images en ambiance « non-coopérative »

La reconnaissance d'une voiture ou d'un piéton en ville fait peu de débat, en densité faible à moyenne, et par beau temps : l'IA permet en général facilement de classifier les cibles urbaines, et un détecteur basé sur l'IA peut être réalisé, au prix du GPU sous-jacent.
On notera cependant des limites dans des cas d'usage très fréquents tels que des **parapluies** quand il pleut (la détection de personnes cachées est alors rendue très difficile), ou pour des foules denses (comptage par détection impossible), ou le port de vêtements/masques faisant disparaître les formes de l'objet (confusion/non détection), ainsi que les dessins décoratifs ou publicitaires sur les véhicules qui peuvent induire des fausses classifications.

En revanche, **dans des applications destinées à des usages dits « non-coopératifs », les cibles observées vont chercher à ne pas être détectées ni reconnues pendant leur traversée de la zone de détection. L'IA est alors inopérante**. On cite en général les contextes suivants :

- Préparation et perpétration d'une intrusion dans un site, ou sur une frontière,
- Evasion d'un lieu de rétention (prison, asile),
- Applications militaires.

Sur ces usages, les cibles à détecter vont avoir tendance à se camoufler (derrière un carton, un parapluie, en tenue militaire, sous une

[36] *Identifying Human Edited Images using a CNN*, Jordan Lee, Willy Lin, Konstantinos Ntalis, Anirudh Shah, William Tung, and Maxwell Wulff, Jan 2021, open source paper.
[37] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. *Synthesizing images of humans in unseen poses*. In CVPR, 2018

fourrure), à ramper[38], à passer de nuit quand l'imagerie couleur se dégrade fortement : et la classification ne donnera aucun résultat.

En ambiance non-coopérative, le seul apport de l'IA est celui de la mesure de différence entre deux images, ou l'analyse du mouvement, qui ne font pas d'hypothèses sur la forme recherchée.

C'est donc bien **dans l'espace public**, où l'on observe des objets multiples et ouvertement visibles (personnes, voitures) dans des scènes plus complexes que les situations maîtrisées d'un site fermé, que **l'IA apporte de nouvelles possibilités, et repousse les limites de l'analyse classique[39]**.

3- L'impact économique sur l'analyse d'images

3.1 – Les usages

Depuis 2005 à EVITECH, nous équipons les sites sensibles en solutions d'analyse intelligente d'images issues des caméras couleur et thermiques, pour la détection d'incidents variés : intrusions, fumées, fuites d'hydrocarbures, analyse de comportements, maîtrise des foules (comptage).
Le développement récent des parcs municipaux de caméras couleur a conduit un certain nombre de collectivités à se doter aussi d'analyse d'images, d'abord pour la protection des sites municipaux ou départementaux en période de fermeture[40], puis pour la gestion d'incidents en ville. L'IA apporte la possibilité de classer les cibles de façon plus fine, notamment pour réaliser des mesures statistiques (fréquentation, vitesses). Nous abordons dans cette partie les aspects économiques.

3.2 – Le facteur prix

Une caméra couleur de vidéo surveillance coûte en moyenne entre 50 et 1000 €, 50 € pour une petite caméra fixe d'entrée de gamme, et 1000 € pour une caméra motorisée orientable et de bonne facture. Il s'agit de prix moyens généralement constatés sur les marchés sur lesquels EVITECH intervient. Une fois posée dans un espace public, cette caméra coûte entre 500 € et 10.000 €, selon qu'elle se trouve posée avec 3 vis et câblée « au pistolet à colle » jusqu'à la loge d'un gardien dans un hall de HLM[41], ou fixée par une potence en acier en haut d'un mât en ville auprès d'un édifice classé, au bout d'une tranchée qui a nécessité le percement du trottoir ou de la chaussée[42].

Lorsqu'on questionne les opérateurs **sur la valeur perçue de l'analyse d'images**, par rapport au coût de la caméra, on obtient des réponses en général assez **inférieures à 50%**[43]. Ce coût inclut le prix de la licence logicielle, en général rapporté au nombre de flux vidéo à traiter, et celui du matériel nécessaire pour l'utiliser, CPU et GPU.

3.2 – Partager pour économiser

Compte-tenu de ces contraintes, les opérateurs de vidéo-surveillance de l'espace public ne sont pas en mesure aujourd'hui de dédier des centaines de Watts et des milliers d'euros par flux vidéo pour doter leurs caméras, même partiellement, d'une intelligence de type Deep Learning. Du côté des éditeurs logiciels, il faut viser une mutualisation des moyens informatiques et un usage parcimonieux de la ressource GPU. Il faut concevoir des solutions capables de la *partager*

[38] Voire effectuer un mouvement de « ramper-rouler », tête vers la caméra, pour minimiser leur surface visible tout en glissant lentement au niveau du sol.
[39] Sur l'application de détection de dépôts sauvages, par exemple, l'IA ne permet pas de classifier les classes d'objets déposés qui sont trop variés (sacs, matelas, fenêtres, gravats, …), mais par contre on classifie bien la voiture ou la personne qui effectue le dépôt.
[40] Les solutions EVITECH équipent par exemple des centaines de caméras thermiques sur la protection des SDIS.
[41] Estimation RIVP.
[42] Estimation Ville de Paris.
[43] Enquête qualitative réalisée par Evitech, en 2016, dans le cadre du Projet Eurostars Cro-magnon, pour un prix d'acquisition, hors installation et configuration.

simultanément entre de nombreux flux vidéo, pour en limiter l'impact financier.

Un point mérite explication ici : lorsqu'une carte GPU doit exécuter les traitements d'un réseau DNN, celui-ci doit être chargé dans la carte et cette opération prend en général quelques secondes. Le réseau une fois chargé occupe une partie de la mémoire de la carte, et, si celle-ci est suffisante, il est possible de charger deux, voire plusieurs réseaux (cf. §2.2) et de les utiliser alternativement et instantanément sur la carte. Lorsqu'une configuration d'analyse d'images est active sur un parc de caméras, il est donc préférable de ne pas modifier l'allocation des réseaux aux cartes GPU, sauf exceptionnellement pour les démarrages-arrêts, ou pour la redondance. Le *partage* dont il est question ci-dessus consiste donc à effectuer le même traitement (ex : reconnaître des objets) sur des images d'origines différentes (des flux différents).

Certains acteurs utilisent l'IA de façon systématique (et permanente, c'est la base de leur logiciel d'analyse) pour analyser continument les images des caméras traitées (dont les formats sont couramment 1080x1080 ou 1980x1080 pixels, voire plus gros). Comme les cartes GPU traitent en général des formats de taille inférieurs ou similaire à 500x500 pixels, cette approche nécessite soit de réduire l'image, et de perdre des informations précieuses pour la classification des objets lointains, plus petits, soit de découper l'image pour la passer en 4 ou en 8 parties, puis de recoller les morceaux et les détections, ce qui consomme 4 à 8 fois plus, et pose des problèmes aux frontières. Leurs applications requièrent donc, selon la résolution de traitement et la fréquence d'analyse, des ressources GPU relativement élevées et permanentes pour peu de caméras, ou bien un sacrifice de la résolution disponible.

A contrario, pour partager au mieux ces ressources, notre logiciel Jaguar, opérant sur CPU, acquiert le flux vidéo et détecte tout d'abord les nouvelles cibles qui se présentent à l'image d'une caméra. Puis, si la configuration le prévoit, il effectue une requête vers le GPU pour classifier ces nouvelles cibles, consécutivement à leur apparition, et, selon la stratégie visée, avec potentiellement la zone de l'image à la meilleure résolution possible, en analysant exactement la zone de la cible. **Il ne sollicite pas de classification la nuit quand il n'y a pas de cible**[44], et **il cesse de demander cette classification une fois qu'il la possède**, et il entretient cette propriété avec le suivi de l'objet en mouvement à l'image, même lorsque la cible devient toute petite ou se trouve partiellement cachée par un muret ou une voiture (dans des situations où la classification ne la reconnaitrait pas). Le GPU n'est plus sollicité, et il est disponible pour classifier un autre objet apparu sur une autre caméra. Le bénéfice est double : plus de caméras traitées par carte GPU, et, en période de basse fréquentation, très peu d'appels au GPU et donc une consommation électrique réduite de la part du GPU.

Cette capacité extraordinaire est une caractéristique essentielle dans une approche parcimonieuse : elle repose sur l'utilisation continue du tracking à 100-200 € de coût matériel CPU par flux traité, et sur l'utilisation ponctuelle de la classification à l'apparition de la cible (sur carte GPU, 1000 €). Une silhouette sur un parcours urbain dans le champ d'une caméra persistant généralement entre 2 et 20 secondes, l'économie réalisée par flux vidéo traité représente de 98% à 99,8% du coût GPU (300 W, 1000 €) par rapport à une détection continue par GPU.

[44] La carte GPU passe alors en veille et ne consomme pas plus de quelques watts.

Ces critères n'ont pas encore tellement émergé, dans les appels d'offres des opérateurs du marché, mais, compte-tenu des contraintes de développement durable, ou plus prosaïquement de l'évolution des coûts de l'énergie, ils ne sauraient tarder à devenir des critères déterminants.
En partageant une même carte GPU entre 10 et 100 caméras (10 pour la classification des mobiles dans une rue pas très fréquentée, 100 pour une mesure de densité de la foule par caméra toutes les 20 secondes sur un parc de caméras), comme nous le faisons avec nos logiciels Jaguar et Lynx, nous permettons aux opérateurs d'accéder aux bénéfices de l'IA avec un coût économique minimal.

3-3 – Une toute petite plate-forme pour économiser

Dans la gamme de NVIDIA, nous avons aussi exploité la carte JETSON Nano 4 GB sur son petit module autonome, dont le prix initial (autour de 130 €) et la consommation (5W) étaient compatibles avec des traitements pour une, voire pour quelques caméras. Pour des réseaux de taille raisonnable, ou au prix d'une opération d'amaigrissement du réseau neuronal (*pruning*), il est possible de classifier quelques images par seconde (ce qui donne une référence de temps de traitement par rapport aux autres fonctions). Raccordée à un petit ordinateur de type NUC consommant 15-25W et traitant entre une et 4 caméras, cette carte permet de composer, moyennant un lien réseau et une distribution du traitement, une plate-forme tout à fait compétitive pour une petite borne vidéo urbaine autonome[45].

3-4 – Exécuter sur CPU pour économiser

Depuis 2017 le format « ouvert » *ONNX*[46] a été normalisé et se développe dans le but de constituer une plateforme normalisée pour l'implémentation de modèles de réseaux neuronaux. Cette initiative a été rejointe par de nombreux acteurs, ce qui a permis de disposer d'interpréteurs variés de ce format, et en particulier de permettre des portages sur différentes plates-formes.

L'une d'elles est le CPU, tout simplement, qui n'est pas très efficace (7 à 8 fois plus lent que la carte GPU mentionnée précédemment, sur un demi processeur core i9 dédié) mais qui permet de s'abstraire complètement du GPU.
Sur des applications de mesure de densité de personnes dans une piscine pour commander la chloration, ou en pied de remontée mécanique pour commander le mode Ecodrive[47], une mesure de densité toutes les 15 ou 30 secondes suffit largement et donne parfaite satisfaction à nos clients.
Dans ces applications, le compteur est une boîte noire qui utilise l'image pour compter et la détruit instantanément. Il sort un chiffre à intervalle régulier, qui sert ensuite à commander le process.

4- Le droit d'usage en espaces publics

Lorsqu'elle est installée dans l'espace public, une caméra est déclarée en Préfecture avec un objectif de sécurité, ou encore une **finalité**, bien spécifiée. Il ne s'agit pas de fournir un jouet à l'opérateur du centre vidéo, pour des usages à la demande et illimités[48], mais de remplir une mission précise : détecter des attroupements dans un lieu de passage, détecter des encombrements, des arrêts interdits, des dépôts d'objets sauvages (déchets), s'assurer de la tranquillité des rues la nuit (détecter des poursuites, des rodéos, des personnes qui courent, …), mesurer des flux, etc. Ces missions figurent sur la déclaration de la caméra, et en sont indissociables.
Or, il y a un écart aujourd'hui entre les technologies d'analyse d'images disponibles, et la loi sur l'analyse d'images dans l'espace public : l'analyse d'images sur l'espace public est considérée d'un

[45] https://vdsys.fr/vigicam-ii/ : détection de déchets abandonnés par exemple.
[46] https://onnx.ai/index.html
[47] Ecodrive : adaptation automatique de la vitesse de la remontée mécanique par rapport au nombre de personnes en attente.
[48] Comme suivre à la vidéo des personnes que l'agent choisirait arbitrairement ou qu'il connaît personnellement.

point de vue juridique comme un traitement automatisé (et effectivement c'est le traitement d'un ordinateur, donc automatisé) sur des données personnelles de personnes n'ayant pas donné leur consentement (et effectivement des visages de passants dans la rue assez résolus pour qu'on puisse les reconnaître dans une vidéo sont des données personnelles, de passants à qui on ne demande rien pour traverser la rue).
Cette capacité laisse planer un risque sur les libertés [49].

Les potentialités du traitement d'images étant importantes, notamment en matière de biométrie, la Loi, même si elle s'est assouplie récemment et temporairement pour la « fenêtre » des JO 2024, limite le droit d'usage des applications de traitement d'images sur les caméras installées, de façon à permettre certaines applications (mesure du niveau de l'eau d'une rivière, détection de fumée, comptage de véhicules et passants, contrôles automatiques du respect du code de la route, et détection de certaines situations à risques sur des personnes dans certains endroits bien précis exposés à ces risques), et interdire les abus, comme on peut les voir dans certains pays qui exploitent par exemple une *reconnaissance supposée raciale des visages*[50], ou qui détectent abusivement tout attroupement comme une menace à l'ordre établi.
En matière de responsabilité du traitement d'images, on peut citer également les écarts de performances de reconnaissance de personnes de couleur aux USA, sur des caméras embarquées (écarts non voulus par les concepteurs des applications, mais dus au fait que les bases d'apprentissage en contiennent moins d'exemples), qui ont conduit à des fausses reconnaissances, fatales lors de certains contrôles de police.
En France, l'application de traitement d'images sera déclarée avec la caméra, et toute caméra ne pourra se voir adjoindre n'importe quel traitement. Par ailleurs, les traitements devront justifier de leur objectivité (une absence de biais anthropologique).
On retrouve un lien fort entre la déclaration préfectorale et les applications autorisées à exploiter les images de la caméra : à une caméra placée pour détecter des attroupements, on peut adjoindre un algorithme de comptage ou de détection d'attroupement, tandis qu'à une caméra placée dans la rue et assurant la tranquillité, on peut adjoindre une mesure de vitesse des déplacements, mais dans la limite de sa déclaration.
Ceci permet, comme pour les usages dans les sites privés, et militaires, d'optimiser les configurations pour atteindre 100% de détection des situations recherchées avec un taux de fausses alarmes ou d'erreur minimal[51], deux objectifs qui sont au cœur de notre culture depuis notre création.

[49] https://www.cnil.fr/sites/default/files/atoms/files/cameras-intelligentes-augmentees_position_cnil.pdf
[50] https://ipvm.com/reports/racial-ethnic-standards, IPVM, John Honovich, 2021.
[51] Le logiciel Jaguar d'EVITECH possède la certification iLIDS avec le grade le plus élevé depuis 2013.