# Usages de la Vidéo Protection Intelligente

Pierre Bernas

# Usages de la Vidéo Protection Intelligente

Pierre BERNAS, EVITECH SAS, pbernas@evitech.com

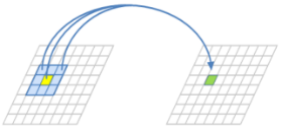

Nous examinons ici en premier lieu les impacts des évolutions de l'IA sur les applications d'**analyse intelligente d'images pour la sécurité globale**, dont Evitech est éditeur de logiciel depuis 2005 [1]. Ces applications analysent les images de caméras de vidéoprotection couleur et thermiques pour y détecter des situations dangereuses, et pour prévenir ou limiter leurs conséquences : intrusion, comportement individuel dangereux ou port d'arme, détection d'hydrocarbures ou de fumée, gestion de foules, présence d'objet suspect, contrôle d'un processus, etc. Nous identifions alors les apports successifs de l'IA pour l'analyse vidéo. Nous identifions alors les apports successifs de l'IA pour l'analyse vidéo, puis nous abordons le processus de conception, et les cas d'usage. On se penchera alors sur la nécessité d'adapter les processus de déploiement de ces fonctions intrinsèquement complexes à des utilisateurs non techniques, et conclurons sur les freins potentiels légaux ou d'acceptabilité de ces solutions.

**Keywords:** Vidéo, IA, Surveillance, Energie, Acceptabilité, Usages légaux.

**Reference** : Bernas,, P. 2024. Usages de la vidéo protection intelligente. Dans IHM'24: Articles Industriels de la 35ème Conférence Internationale Francophone sur l'Interaction Humain–Machine, Mars 25–29, 2024, Paris, France.

## 1 L'IMPACT DE LA SCIENCE DE L'IA SUR LE TRAITEMENT VIDEO

### 1.1 Les débuts de la discipline : détecter et suivre des mobiles



Le traitement d'images est une discipline qui a émergé avec l'image numérique, dans les années 1980 [2] (après que des précurseurs américains en aient posé les bases dès l'apparition des ordinateurs dans les années 60), et qui a progressé jusque dans les années 2015 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 [3], filtrage, statistiques, calculs de gradients et recherche de points caractéristiques, histogrammes, apprentissage (ACP, SVM)... En matière de reconnaissance d'objets, les performances de classification[1] restaient faibles (ex : quelques dizaines de pourcents), sauf dans des environnements très contraints[2].

[1] Reconnaître une forme de personne, de véhicule, d'animal, etc.

[2] Ex : dans une usine, dans la rue, où les classes de mobiles potentiellement présents sont en faible nombre, et leurs caractéristiques faciles à distinguer (couleur, taille, forme).

C'était une époque d'intuitions, de formulations mathématiques d'impressions que l'expert cherchait à automatiser dans un algorithme : « *le traitement d'images était dirigé par l'Homme* »[3]. En 2005, le décodage de la vidéo numérique en temps réel devient possible, et aujourd'hui un ordinateur d'un coût de l'ordre de 1000 €[4], consommant 150 W, peut décoder simultanément 5 à 10 flux vidéo, et leur appliquer un traitement algorithmique classique (détection de changements, suivi des cibles au cours du temps, filtrage des changements parasites, …), 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[5]. On y détecte le mouvement, la forme, la taille, mais on ne reconnaît ni les objets, ni les situations précises. Lorsque deux cibles se croisent dans l'image, elles forment temporairement un groupe tant qu'elles sont en contact « optique », puis retrouvent chacune leur identité (caractéristiques, identifiant de suivi, historique de provenance, etc) à leur séparation. Cette approche est particulièrement adaptée au suivi d'un ou de quelques mobiles isolés, elle n'est pas efficiente sur une foule où tous les mobiles forment ensemble un gros *blob*[6] à l'image, qu'on ne sait pas segmenter en « sous-cibles » individuelles.

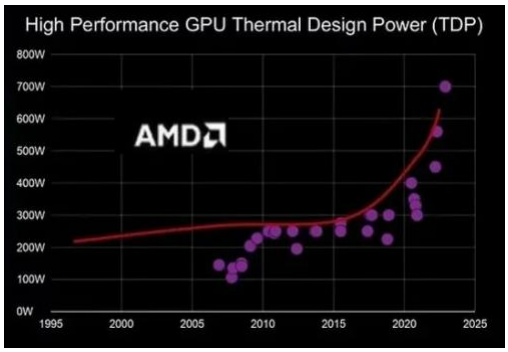
Au niveau de l'interface homme-logiciel, les objets trackés sont des *blobs* de pixels en mouvement, tandis qu'une fois qu'ils sont superposés à la vidéo, celle-ci leur donne tout leur sens (détourage des objets en mouvement). Le filtrage des changements parasites est la clef d'une expérience utilisateur agréable (le système amène à l'utilisateur des informations de valeur), dans le cas contraire l'utilisateur a l'impression de perdre son temps à regarder des animations sur des clips vidéo qui détectent n'importe quoi (le système semble « ne pas respecter » l'utilisateur en lui remontant des « déchets » (fausses détections)).

### 1.2 Les réseaux neuronaux profonds : reconnaître les classes d'objets


Autour de 2013, l'émergence des réseaux neuronaux profonds [4] et des réseaux neuronaux convolutionnels [5], en conjonction avec l'apparition de cartes graphiques puissantes, ont permis de briser le plafond de verre des performances en matière de reconnaissance d'objets dans des images, qui stagnaient depuis plusieurs années. Le modèle AlexNet [6] a notamment illustré ceci sur la compétition de reconnaissance d'objets sur la base ImageNet [7] (14 millions d'images). Ces réseaux élaborent automatiquement, dans leurs données intermédiaires lors de la reconnaissance d'images, des descripteurs bien plus nombreux et riches que ceux que l'on cherchait à utiliser dans le traitement d'images classique. La caractérisation des personnes, des animaux, celle des véhicules avec leurs sous-classes, par exemple, sont devenues possibles avec des scores de reconnaissance élevés (plus de 90% sur des silhouettes vues entières). Ces outils ont permis de préparer des applications destinées à reconnaître s'il y avait tel ou tel type d'objet dans une image, et de le localiser, pour un ensemble bien défini de type d'objets pour lesquelles on les a entraînées. Mais ce n'est pas une panacée : dans les cas « non coopératifs », où une personne par exemple

[3] Roger Mohr (1947-2017†), ex-Directeur du laboratoire Inria GRAVIR et de l'Ensimag, et cofondateur d'Evitech.
[4] Processeur Intel core i7 12700, RAM 16 GB, en Oct 2022.
[5] Une caméra elle-même consomme 5 à 10 Watts.
[6] Blob : Binary large object.

se dissimule derrière un carton ou sous une couverture[7], la classification est inopérante. Dans les cas de foules, seules les premières silhouettes suffisamment visibles entièrement ou assez complètement, à l'avant-plan, seront classifiées, les petits morceaux (sommets des toits de voitures sur une scène de rue, bords de têtes dans une foule) seront totalement ignorés. C'est pourquoi cette approche est inefficiente pour compter des objets dans une situation dense, on utilise alors d'autres techniques.




L'intégration des cartes graphiques est impérative pour des reconnaissances d'images rapides (typiquement 10 à 50 ms). Ces cartes graphiques font plus que doubler l'énergie consommée par caméra, par rapport à la seule analyse du mouvement, et leur puissance augmente (cf. graphique AMD ci-contre).

Cependant, un usage parcimonieux (comme par exemple une détection déclenchée par un événement d'apparition issu de l'analyse du mouvement, information qui sera ensuite conservée dans le suivi ultérieur de l'objet) permet de limiter cet impact.

On peut donc utiliser cette innovation de deux façons : soit pour enrichir les informations d'une détection et d un suivi algorithmique, en l'utilisant parcimonieusement sur des nouveaux objets mobiles détectés, soit à très haute fréquence, comme un détecteur, avec son *tracking* associé.

Il y a des avantages et des inconvénients aux deux approches :

- L'IA comme détecteur de cibles : une carte graphique (GPU) ne pourra traiter que quelques caméras (2 à 6, selon la résolution visée) à une fréquence acceptable (7 à 12 images par seconde). Les usages pour surveiller des situations non-coopératives restent exclus. Le fonctionnement sur des groupes peu denses est meilleur que dans l'approche algorithmique.
- Comme classificateur venant qualifier des cibles détectées et suivies par algorithmie (CPU) : l'impact de consommation électrique reste alors modeste et la carte peut être partagée entre des caméras plus nombreuses, pour être utilisée pour classifier les nouvelles apparitions. Elle sera en veille (retombant typiquement de 300 W à 50 W) lors des périodes « calmes » (la nuit par exemple). Mais les agrégats de cibles sont très difficiles à suivre.

Dans le cadre d'une qualification pour un client industriel pour lequel il était nécessaire de surveiller des situations de coactivités véhicules/personnes, nous avons testé les deux approches, et la seconde s'est finalement avérée supérieure. Le but de l'application était d'identifier les risques, pour une personne, d'être présente à côté d'un véhicule en manœuvre, dans des scènes impliquant une à 5 personnes et un à deux véhicules, vues par une caméra en hauteur : la question posée était de savoir quand il y avait des véhicules et quand il y avait des personnes dans la zone. L'utilisation croisée de la détection de cibles et du *tracking* algorithmique, avec la classification ponctuelle par réseaux neuronaux profonds, ont permis sur une vidéo de coactivité de 12 heures d'atteindre une performance de 99,3%, tandis que sur cette même vidéo l'approche par classification n'atteignait pas 85%, du fait des cas de masquages (personnes derrière véhicules). Il faut cependant reconnaître que c'est la notion de groupes, exposée au

[7] Quelqu'un qui se dissimule pour ne pas être vu ouvrant une porte ou entrant dans un site, par exemple, ou encore un prisonnier qui s'évade…

§1.1 qui permettait de maintenir les informations de présence relatives aux mobiles invisibles car recouverts par des véhicules.

### 1.3 L'évolution vers les VLM : identifier des situations

Très récemment, les VLM (*Vision Language Models*) [8], comme par exemple LLAVA[8] (Modèle ouvert soutenu notamment par Microsoft), ou GPT-4V (OpenAI), entraînés sur des millions d'images de corpus ouverts, annotées par les textes les accompagnant, ou annotées automatiquement par l'IA, permettent d'accéder à une reconnaissance universelle et contextuelle d'images (ex : telle situation[9], sous la pluie, la nuit, dans la foule, …). Celle-ci rend possible la réalisation d'applications permettant notamment de reconnaître une variété quasi illimitée de types d'objets, soit connus, soit constitués à partir de modifications décrites sur des types connus (ex : une girafe avec un chapeau melon violet), ainsi que de caractériser des situations dans lesquelles ceux-ci sont présents.
Ce type de modèles s'active par des prompts (description textuelle d'une requête de détection), et s'appuie sur un réseau neuronal profond de type *Transformeur*, qui peut opérer sur une carte GPU relativement commune (NVIDIA RTX 3070) avec un temps de traitement d'une image de l'ordre de la seconde. Les performances de détection de ces nouveaux modèles sont très élevées. Dans un test récent, on atteint de 90 à 95 % de précision pour une détection d'une conjonction de présences dans une scène, appliquée à une centaine d'images, en partie positives et négatives, et, après quelques minutes d'essais d'un prompt adapté à la situation (test réalisé début 2024 sur LLAVA 1.6 Mistral-7.b).
Même sans disposer encore de retours d'expériences importants sur leur usage, on peut gager que ces modèles permettront bientôt l'extension des outils d'analyse vidéo à des applications « non-envisagées par avance », qui pourront être configurées par des prompts saisis <u>après</u> installation, et activées par des événements (horloge ou détection dans la vidéo d'un événement, etc). Elles pourront donc être modifiées et/ou étendues après intégration et livraison auprès d'utilisateurs pour répondre à des demandes encore inconnues à la date de leur livraison, ce qui débloquera des capacités non envisagées jusqu'à ce jour, et permettra aussi de plus finement circonscrire les conditions d'une situation à détecter. L'intelligence n'est plus statique, elle deviendra une capacité relativement évolutive.

## 2 UTILISATEURS ET CONCEPTION

### 2.1 Les cahiers des charges et la conception


Les demandes des utilisateurs en matière d'analyse de la vidéo consistent le plus souvent à pouvoir détecter des situations dangereuses ou dommageables qui se sont produites ou qui ont failli se produire et qu'ils veulent pouvoir empêcher ou enrayer dans le futur. Ils spécifient ces situations et s'enquièrent d'une solution qui y répond. Pour l'incendie, par exemple, il s'agit de détecter une fumée ou une température excessive. Dans le cas d'hydrocarbures, de gaz ou d'abandon de déchets sur un bord de route (R1), il s'agit de fuites (ci-contre, une citerne), et de dommages à l'environnement, et des risques associés ultérieurs (pollution, explosion, …).

[8] https://llava-vl.github.io/
[9] Exemples : « allume un feu », « ouvre une porte », « creuse un trou avec une pelle », « escalade une clôture », etc.

Pour des comportements humains (vol à la tire, vols dans des véhicules, …) il s'agit de mettre en place un enchaînement associant la détection, l'interpellation et la pénalisation : on va essayer de décrire la situation sur une caméra donnée et d'y associer une détection. Un opérateur ferroviaire va par exemple cibler des personnes qui restent autour de l'entrée d'une gare, avant les guichets et les portillons, au moment où les usagers ouvrent leur sac et accèdent à leur portefeuille pour y chercher leur carte ou leur ticket (R2).

Enfin, pour les dangers importants (trop de personnes sur un quai de gare risquant de tomber sur les voies), on va proposer de compter les personnes sur le quai et de lever une alarme pour réguler les flux en cas de trop forte densité (R3). On peut aussi proposer de détecter un mouvement de foule (ex : le risque sera calculé sur le couple « nombre de personnes * vitesse »).

Le taux de détection des vraies alarmes attendu caractérise des préoccupations utilisateurs pour la gestion du problème : un risque majeur (incendie) doit être détecté dans 100% des cas, tandis qu'un phénomène statistique qui doit être quantifié (franchissement d'un portillon sans payer) peut supporter un taux de détection de 90%. Dans l'autre sens, le taux de détections inappropriées (fausses alarmes, ou situations répondant au scénario recherché, sans pour autant constituer un réel danger, comme par exemple un fumeur sous la caméra, ou un véhicule autorisé mais peu reconnaissable sur une place à accès réglementé, un sanglier grattant au pied d'une clôture pour entrer dans un site, …) est accepté par rapport aux moyens disponibles en matière de levée de doute. Un client sera plus tolérant sur les détections inappropriées si les alarmes sont traitées par un prestataire qui lui facture un forfait, mais il le sera moins si ce sont ses opérateurs qui sont débordés et n'ont pas le temps de regarder les images d'alarmes pour effectuer la levée de doute.

La conception suit alors de façon très proche les demandes des utilisateurs : on va ainsi décrire un mobile ou reconnaître une personne qui stagne durablement dans la zone surveillée par la caméra, ou compter les têtes[10] de personnes dans une zone. Au bord d'une route, on va décrire un véhicule qui s'arrête et qui y dépose des objets.


A l'intérieur des sites privés, les règles applicables sont définies par les propriétaires ou exploitants, et, moyennant le respect du droit à l'image et du droit du travail, un grand nombre de détections sont possibles (intrusion, vitesse, immobilité, …).

Mais hélas le fait de produire une application qui assure cette détection dans l'espace public peut produire des détections inappropriées qui sont pénalisantes pour les libertés publiques, comme les anti-exemples suivants, en reprenant les règles de détection citées plus haut :

- R1 : Des personnes qui installent une table et des chaises pour pique-niquer en bord de route.
- R2 : Une personne qui en attend une autre avec qui elle a rendez-vous, devant l'entrée de la gare,
- R3 : Des habitants d'une cité qui se retrouvent au pied des immeubles pour une fête des voisins.

C'est pourquoi la loi et des mouvements associatifs s'opposent à l'usage généralisé de telles détections dans l'espace public.

[10] Dans une foule dense, une caméra en hauteur ne voit plus que des têtes, et morceaux de têtes.

### 2.2 Interactions utilisateurs

En dehors des besoins de levée de doute, il est nécessaire, comme on le voit, de circonscrire le plus finement possible le contexte de l'alarme, à la fois pour en réduire le nombre et pour empiéter le moins possible sur les libertés publiques.

Deux approches sont alors possibles, et se rencontrent sur le terrain : proposer une interface assez riche et chargée, et assurer, par une prestation, une configuration fine du système avec un engagement de taux de détection et de limite aux fausses alarmes, ou au contraire simplifier le système le plus possible et laisser l'utilisateur final configurer lui-même les algorithmes de détection qu'il veut activer sur les caméras, par exemple en dessinant des directives (zones, directions, gabarits calibrant la perspective au sol) sur l'image de la vidéo, ou en paramétrant quelques valeurs autour de la vidéo (durée, vitesse, …).
En pratique, aucune des deux solutions n'est parfaite, car dans un cas il y a le coût et la contrainte de la prestation, et, dans l'autre cas, la simplicité de l'interface oblige l'utilisateur à se limiter à des scénarios de détection de présence d'une instance de classe (présence de fumée, présence d'une voiture sur une zone interdite, de personne), de délai (arrêt ou stationnement, notion de « dépôt » au bout de quelques dizaines de secondes ), de direction (contre sens), de couleur. Ces notions sont insuffisantes pour circonscrire finement un contexte et atteindre les performances de rectitude (taux de bonnes détections et absence de détections inappropriées) souhaitées. Et la spécification de circonstances complexifierait beaucoup l'interaction.

### 2.3 Obstacles à l'utilisation

Au-delà des récriminations sur la contrainte de devoir lever le doute sur des détections inappropriées, on rencontre alors des obstacles souvent inattendus à l'utilisation. En voici quelques exemples.

Sur un site privé, des gardiens qui ne voulaient pas qu'un système automatique réduise leurs emplois ont joué avec les câbles vidéo analogiques pour intervertir les caméras la nuit et prétendre que l'application intervertissait les flux. D'autres sont venus placer des obstacles devant la caméra pour faire croire à sa défaillance (une main, dont on reconnaissait les doigts, à droite), et essayer de faire rejeter le système. Sur un autre site privé, des opérateurs se sont opposés au système car il allait les empêcher de dormir ( !).
Sur un site public, des opérateurs déjà en charge de surveiller les images des caméras vidéos ont critiqué le système au motif qu'il allait les rendre responsables : tant qu'il ne s'agissait pour eux que de surveiller une vidéo du coin de l'œil, ils pouvaient toujours dire qu'ils avaient manqué un événement pendant qu'ils regardaient une autre caméra ; mais avec un système de gestion d'alarmes les obligeant à acquitter les détections, ils devenaient responsables du traitement des problèmes remontés par ce système !
Sur un site industriel, un processus de gestion de la coactivité entre des personnes et des véhicules qui est rationalisé et surveillé par une application intelligente pour éviter les accidents soulève un antagonisme entre sécurité et productivité : si on contrôle les enchaînements entre les acteurs pour les séquencer, on réduit les risques d'accidents, mais on augmente les temps d'attente… et cela peut aller jusqu'à impacter des primes de productivité !
Enfin, il y a les détections impossibles, qui sont souvent le fait des utilisateurs les plus éloignés de la technologie et qui voudraient que l'on détecte une température corporelle trop élevée avec une caméra couleur, ce qui est

impossible, ou qu'on anticipe ce à quoi pense une personne observée à la vidéo (détecter une personne qui surveille le quai de gare pour prévenir un comparse de l'arrivée d'un contrôleur …) : la caméra n'entre pas dans la tête des personnes présentes…
La capacité de prédiction, ici, comporte aussi ses limites : il est par exemple possible, pour un système surveillant un rond-point, de lui demander de prédire, à la façon dont un véhicule aborde ce rond-point, où il va sortir. Mais ceci se limite à ce type de situations. Contrairement à ce que laissent croire les fictions, c'est simplement une approche de corrélation statistique, entre la voie de circulation utilisée, la position du véhicule sur cette voie, et sa vitesse, qui prédit la sortie de rond-point visée. Cette approche suppose une donnée statistique antérieure abondante.

Ces obstacles caractérisent finalement très souvent un retard entre un projet qui est enclenché, et une doctrine[11] qui n'est pas aboutie et acceptée dans de nombreux projets : on se retrouve en train de configurer ou de déployer le système et on se demande soudain comment telle ou telle situation va être gérée. A titre d'exemple, on voit apparaître depuis quatre ans des demandes d'analyse d'images à bord de trains, du fait de la disparition programmée des conducteurs, mais on ne sait pas encore comment va être traitée une alarme survenant en provenance d'un train en train de rouler...

## 3 LE DROIT D'USAGE EN ESPACES PUBLICS

### 3.1 Impacts sociétaux, risques

L'exemple cité plus haut du comptage sur un quai de gare pourrait sembler justifié par le risque que présente une situation de surpopulation, mais il a été cité en exemple par la CNIL comme une détection qu'elle juge illégale. Ce qui constitue l'illégalité est l'enchaînement entre le traitement (l'analyse de l'image) et l'action répressive immédiate qui porte sur les personnes mêmes qui ont provoqué la détection.
On le comprend mieux si on pense au cas de la fête des voisins au pied d'une cité, l'arrivée d'une patrouille de police à chaque fois que plus de 10 personnes s'y retrouvent ensemble s'apparenterait rapidement à un harcèlement.

### 3.2 Acceptabilité et droit d'usage

Le droit du contrôle routier, conduit par l'Etat, permet une surveillance assez large, même si son acceptabilité n'est pas totale, comme on peut le voir sur l'exemple des radars routiers masqués ou dégradés. On appelle radars de feux des systèmes purement vidéo. Mais ce droit se limite strictement à la doctrine du code de la route, il n'a pas de continuum juridique avec la vidéo surveillance.
La vidéo surveillance sur l'espace public est installée par les opérateurs de sites publics (Etat, Collectivités, Transports, Culture, Commerces, Sites recevant du public…), conformément la loi Sarkozy de 2006, et relève selon les situations du droit à l'image (RGPD), du droit social (droit du travail), et/ou de la loi Informatique et Libertés, qui interdit les traitements massifs de données personnelles pour ficher des informations, et discriminer des individus dans des recherches multicritères. Dans certains cas, un dispositif de consentement[12] peut permettre de diminuer cette interdiction (notamment pour l'enregistrement automatique des images).

[11] Terme militaire utilisé pour décrire l'ensemble des procédures de surveillance, les règles du système automatique, et les actions attendues des hommes en cas d'occurrence d'événements dangereux redoutés.
[12] « N'entrez ici que si vous consentez à être filmé, … »

Lorsqu'elle est installée dans l'espace public, une caméra de vidéo-surveillance 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[13], mais de remplir une mission précise de détection de danger, qui a été validée par avance pour la caméra concernée. Ces missions figurent sur la déclaration d'installation et d'usage de la caméra, et en sont indissociables.

Malgré son caractère qui est évidemment utile dans un ensemble de cas (alerter lorsqu'un bâtiment public fermé prend feu, si une personne tombe d'un pont, etc), 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 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 donc planer un risque sur les libertés[14], et elle est par conséquent systématiquement interdite à ce jour. La loi dite JO 2024 de Mai 2023 autorise temporairement 8 scénarios pour sécuriser les JO et les grands événements, mais elle limite drastiquement cette exception[15] et la rend en pratique d'application très réduite.

Il faudrait pouvoir permettre plus largement certaines applications (mesure du niveau de l'eau d'une rivière, détection de fumée au-dessus d'une école fermée, comptage de véhicules et passants, et détection de certaines situations à risques dans certains endroits bien précis exposés à ces risques, justifiée par des incidents passés), mais aussi interdire les abus comme on peut les voir dans certains pays qui exploitent par exemple une *reconnaissance supposée raciale des visages*[16], ou qui détectent abusivement tout attroupement comme une menace à l'ordre établi. Ceci suppose une configuration prédéfinie, agréée et validée de l'algorithme de détection sur la caméra, qui correspond à sa déclaration, et qui interdit qu'un opérateur configure une détection comme bon lui semble : une interface homme-machine sophistiquée, et réservée à des professionnels assermentés ?

## 4 CONCLUSION

Comme on l'a vu, malgré ses 25 années d'existence, la vidéo-surveillance intelligente reste un sujet d'actualité, qui provoque rejets et peurs (la fiction exagérant la réalité), et qui est en émergence dans l'espace public. Elle utilise des technologies en constant renouvellement. Beaucoup de nouveautés sont encore à venir, et cette technologie soulève des questions d'acceptabilité et de légalité, qui devront être traitées de façon démocratique et apaisée pour un usage plus consensuel et mieux accepté.

## REFERENCES


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

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

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

[4] 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.


[13] Comme suivre à la vidéo des personnes que l'agent choisirait arbitrairement ou qu'il connaît personnellement.

[14] https://www.cnil.fr/sites/default/files/atoms/files/cameras-intelligentes-augmentees_position_cnil.pdf

[15] Selon cette loi, seul l'Etat peut acquérir les traitements et en doter les utilisateurs, après avoir vérifié l'objectivité des traitements et la sécurité informatique des plates-formes.

[16] https://ipvm.com/reports/racial-ethnic-standards, IPVM, John Honovich, 2021.

[5] « Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation », DeepLearning 0.1, LISA Lab (Août 2013).

[6] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017. 60(6): p. 84-90.

[7] 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.

[8] Visual Instruction Tuning, H. Liu, C. Li, Q Wu, YJ Lee, CVPR 2023, Dec 2023. https://arxiv.org/pdf/2304.08485.pdf.

[9] P. Bernas, G. Née, P. Drabczuk, "*Peaceful Monitoring of Crowds*", Conférence WISG 2013, Troyes,