Smart Video Protection and A.I.

P. Bernas, G. Née, EVITECH SAS.

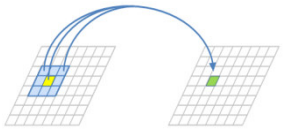
Scientific, technical, economic, and legal considerations

In this article, we examine the impact of AI on intelligent **image analysis applications for global security**, based on images from color and thermal CCTV (video-protection) cameras, which has been the expertise of EVITECH since 2005[1]. These image analysis applications aim to detect dangerous situations[2] in order to prevent or limit their consequences: intrusion into a sensitive site, dangerous individual behavior, detection of a hydrocarbon leak or smoke, crowd control (counting, flow, incidents), presence of suspicious or dangerous objects, etc. After a general presentation of the history of the discipline, we identify the most relevant contributions of AI to video analysis, in theoretical terms, but also in practical terms (AI has facilitated analysis in public space), as the capacities of video surveillance operators are not unlimited either in terms of energy consumed or in terms of price. We cannot conclude without addressing the legal issue, as the European regulation limits these uses in public space.

1- The impact of AI science on image processing in general

1.1 – The discipline

Image processing is a discipline that emerged with the digital images, in the 1980's[3] (after American precursors had laid the foundations for it, once computers had appeared in the 1960s), and which progressed until 2005 by adding successive algorithmic and mathematical techniques allowing to analyze and better discern the properties of the images and hence those of the video: mathematical morphology[4], filtering, statistics, gradient calculations and search for characteristic points, histograms...



It was a time of intuitions, of trials and errors, of mathematical formulations of impressions that the expert sought to automate in an algorithm: *"image processing was directed by Man"*[5]. Classical neural networks, using learning techniques, were feebly developed around character recognition capabilities[6].

---

[1] « Interests and limits of intelligent video for Global Security »,P. Bernas, Conf. AVIRS 2008.

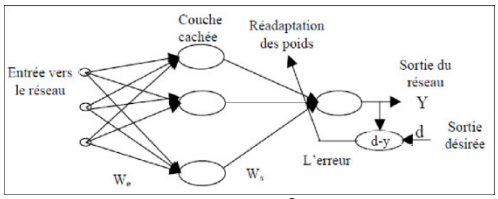
[2] In this picture, a person holding a firearm (source Evitech Lynx).

[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 †), Former Director of Inria GRAVIR laboratory and Ensimag school, and cofounder of Evitech.

[6] The first convolutional network Neocognitron (Kunihiko Fukushima) was proposed in 1979, it is trained to recognize simple visual shapes. It still constitutes recent works basis: 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.

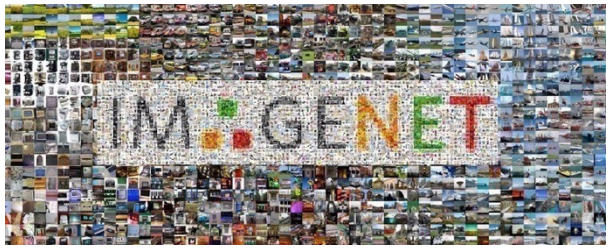


However, the training strategy of the networks has evolved thanks to *"back propagation"* (using the error to correct learning). Used for training neural networks since 1985[7], it was in 1989 that Yann Lecun was the first to implement a network combining *"back propagation"*[8] and a convolutional neural network applied to the recognition of handwritten characters[9].

In parallel, learning techniques have been developed as an automatic alternative, with notably « Principal Component Analysis »[10] and « *Support Vector Machines* »[11]. Based on descriptors taken from classical processing (colors patterns, HOGs, characteristic points, ...) they have shown interesting progress capacities, but without succeeding in revolutionizing the field, because the gains obtained were not spectacular.

It is also around 2005 that the power of the processors made it possible to decode the digital video without necessarily using a parallel processor, and to evolve from the processing of the photographs to that of the video in real time. Today, a computer costing around 1000 €[12], consuming 150 W, can simultaneously decode 5 to 10 video H.264 encoded streams, and apply classic algorithmic processing to them.
This puts the hardware cost of this processing (excluding the cost of the software) in the range of 100€ to 200€ per stream, with an energy consumption of 15 to 30 W for each. Moreover, these processing can operate adapted on parts of the image, at different scales.

Around 2013, the emergence of deep neural networks[13] and convolutional neural networks[14], in conjunction with the emergence of powerful graphics cards, has broken the glass ceiling of performance in object recognition in images, which had been stagnant for several years.



In particular, AlexNet[15] allowed a very significant progress on a competition of object recognition in photographs which was a reference on the color images base ImageNet[16] (14 million images) for years.

1.2 – Neural networks

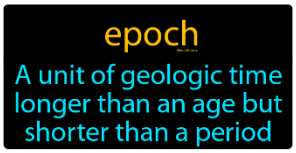
Everything starts with the *neuron*, a kind of elementary micro-computing entity with a weight and able to communicate with its neighbors, during learning and computation operations. The general approach of the processing is to structure a deep convolutional neural network with inputs, outputs, and made of a certain number of layers of neurons (pooling layers, fully connected layers,

[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] Image source : 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 & 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.

convolution layers), linked by operators (mathematical operators, and convolution operators, in particular, involving quantities of neurons).

The learning process makes annotated data (images or others) flow in this network, by applying these operators, and by finely back-propagating by small successive modifications, the weights of the neurons, according to the expected results for each data (annotations). This corrective process is called a learning step. A complete passage of the database of examples (a *dataset*) on the network is called an *epoch*.

In parallel to this learning, we use a second set of annotated data, the test data, to measure the recognition score of the network on this second set. In order not to pollute the process and risk overlearning, these data must imperatively be reserved for measuring the performance of the trained network, and not for training.
When, after an alternating number of *epochs* and tests, the network becomes optimal (it stops improving from one *epoch* to the next), or when it starts to regress, we stop.



We then realize that (in case of image) these networks automatically elaborate, in their intermediate, data descriptors that are much more numerous and richer than those that we were trying to use in the classical image processing.

This learning process modifies at each step the weights of the neurons of the network in a progressive way, by small variations, image after image. We can imagine a notion of distance and continuity between two similar networks, using the distances of the values of their weights. The continuous space defined by all these possible networks, with all the possible values of weights, could thus be seen as a curved surface whose lows would be the optima, for the expected detections. In this paradigm, the learning process would make the network "descend" according to a convergence towards the most accessible through by following the line of slope. It is possible, during training, to reach a local minimum, which is not an absolute optimum. By methods that radically shake up the weights of the neurons, at initialization, we can also change the starting point, restart the whole training, and this time try to converge to another minimum, somewhere else, that we can hope to be more efficient. Sometimes, we issue on the null network (zero weight everywhere).

This observation gives to the discipline an experimental character which can seem quick shocking for a scientist, but which for the moment is the common lot of the actors of the field.

Another subject that also has a strong experimental dimension is the constitution of the training sample data base: which samples to select and how to annotate them. A whole set of rules have emerged from the progressive learning experience to define the right conditions for the constitution of a base of relevant examples for learning. Some of them have been automated, in he case of images, such as mirror inversion of the images, their rotation, or resizing, to artificially increase the number of examples, through a discipline that is rightly called data augmentation[17]. Others are still the object of empirical expertise:

[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.

the variety of positions, of backgrounds, of lightings is sought to balance a dataset. The attempts to use false data[18] in training (copying and inserting "rare" thumbnails of objects in a photo, to try to multiply the examples of them) have shown that they only lead to false data detectors[19].

Once the configuration of the network is obtained, passing an image to it and obtaining a result (e.g. boxes with all recognized objects) is called a *forward*[20].

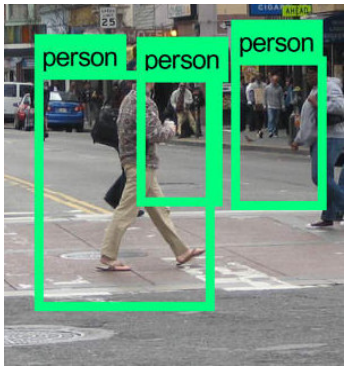
The training time of an *epoch* multiplies the basic time of the *forward* by the number of training cases (e.g., one hundred thousand, or one million). The number of *epochs* needed to converge a network can be large (hundreds). The total training time is the product, that is up to tens or hundreds of millions of *forwards*[21]. However, this time is normally used only once, to build the network.

From this invention, which was extraordinary in terms of the progress achieved in terms of performance, most of the known problems of data processing have been revisited with these techniques and have encountered solutions whose results were often better than with the previous approaches.

2- Opportunities offered by AI on the image

2.1- Recognition of objects, people, faces, ...

The first and most publicized image processing application, for which AI has significantly improved performance, is the recognition of a close-up object in an image or in a color photograph (largely viewed and uncovered). By training a convolutional deep neural network with a large database of annotated images (we are talking about hundreds of thousands, or millions), it is possible today to produce a classification algorithm that reaches performances of more than 90% in the recognition of such “large” objects, in a realistic use domain (close-up, lighting, distance, sharpness, ...).



This is the need for on-board sensors in automatic driving cars, which are expected to have a very high reliability: recognize the objects immediately in front of it (e.g. a person crossing). The progress is then such that they overcome the recognition rate of a human observer[22].

A distinction is made here between photo classification (determining the main object that constitutes the photo) and detection (or recognition), which consists in identifying all the recognizable objects[23] present, even at the background of the photo.

For example, in the case of face detection, on a camera placed above a busy crossing point, we now reach scores closer and closer to 100%, as the recognition distance increases with technology over the years (thus for faces that are more and more distant -smaller- in the in the image).

[18] Here, a false tank.
[19] F. Jurie, for example, on the detection of tanks from a satellite view.
[20] A forecast.
[21] Practically : days weeks.
[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] Here, YOLO : the person in grey in the foreground is detected, but cars and persons in the background are not. Also, the detection of a *building* class (to detect buildings) depends on the existence of the building class in the learning process.

But the recognition rate of the objects in the image is always relative to a certain base of annotated examples (the *dataset*), whose resolution, illumination and quality must be carefully examined to be able to draw conclusions about its future use. A powerful and fast detector today (YOLO v5 here) achieves a **car** recognition (*recall*) performance between 15 and 64% for an *accuracy* (absence of false detections) of 25 to 57% on a subset of the *PascalVOC* color image base database (tests done by Evitech in 2022 on representative images). This database also includes objects in the background, multiple objects, with partial overlaps, which complicates the recognition and decreases the rates. As we can see on these figurs, there is still a lot of room for improvement.

On this same application, other approaches have been and are being massively investigated to allow recognition, even when we have little data (for example only tens, or hundreds of images). Among these, we can mention the *fine tuning* approach: we use a neural network initialized on the millions of images of general classes annotated in a large *dataset* (we have a network that recognizes the objects in this *dataset*), which we then train to learn one or more very specific classes that we are looking for (a new class, hitherto unknown[24]), on the specific data which we have, in lesser number (tens, hundreds). It can converge or specialize on a fairly efficient recognition system: the recognition scores of the new classes are only slightly lower than those of the full training.

In terms of execution time, object recognition in an image is in a way the base unit of image processing by AI. A forward on a 500x500 image on an average GPU card, costing about 1000€[25], with 4 to 6 GB of RAM, consuming about 290 W, mobilizes the card between 10 and 30 milliseconds[26]. Despite the uncertain availability of graphic cards (2021 and early 2022), NVIDIA is the company that is currently providing the most efficient solutions in terms of performance/price ratio. This delay of 10-30 ms is to be compared with the frame rate of the video, 25 to 30 frames per second (fps), so a new frame every 33 to 40 ms. On a very small installation with 4 cameras, one card, at best it could, process all the 4 cameras images at a size reduced from original to 500x500 pixels, at 25 fps, at worst it would only process them at 8 fps, which would mean about one image out of 3, which clearly reduces the performance of an analysis application that would be entirely based on AI to detect and track moving objects in a video.

However, there are more than 4,000 color megapixel cameras in the Paris Video Protection Plan[27] , and more than 500 color cameras in every major European railway station (there is practically no use of thermal cameras in urban/transport environment[28]).

In the field of facial biometrics, behind the recognition of a face shape, pairs of faces are compared to each other to identify an individual: we develop networks dedicated to such a comparison, by training them on pairs of faces of the same person taken with different lighting, at different ages, with different hairstyles, etc. Applications designed as processing cascades (face detection then comparison) can mobilize several successive networks to produce a processing.

[24] The INRIA public research organization has realized a seat detector for ski lifts with this approach.
[25] A NVIDIA RTX 3070 Ti - 8Gb card in october 2022, internet common cost.
[26] For object recognition neural networks of SSD or YOLO type. This time is the forward time. We assume the DNN already loaded in the card.
[27] The Paris CCTV project (PVPP): Image AFP.
[28] There are not yet public reference databases of annotated thermal images like Pascal-Voc or Imagenet for color, neither any know recognition statistics in case of thermal images. We have observed that color images trained networks can sometimes recognize correctly some thermal objects.

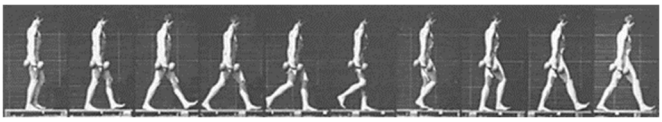
2.2- Postures

Another set of applications that has been the subject of successful research concerns the recognition of parts of the human body, or the superposition in the image of a skeleton or a model of a human body.

Networks such as OpenPose[29] have made it possible to locate all the body parts of a group of dancers in an image. The technology has been used in videos that have since become famous.

However, in terms of execution time, these applications are already 3 to 4 times slower than object recognition. A pose calculation on a 500x500 image on the GPU card mentioned above, takes between 70 and 120 milliseconds.

The recognition of a particular posture of a person (fall on the ground, foot in the air) can be based on this posture detection to recognize an attitude (the arm raised, for example, or to estimate the angle between the legs) or to identify particular gestures (e. g. gripping an object in a store, and immediately after, having hands grouped together at the level of the belly, can suggest a shoplifting).



In the case of gestures, we can also reconstruct a succession of positions and compare it to the successions of positions that have been learned to be to be detected: this leads to solutions based on double neural networks, one to elaborate the postures over the images, and to constitute temporal blocks (for example 50 successive postures of the same person), and the other one to classify the temporal blocks and to recognize a particular gesture, such as a punch.

The second network will therefore be solicited continuously over a sliding window, or punctually in bursts (e. g. once every second, or every two seconds), for each camera to analyze (in order to detect on a horizon of 50 images every 25 to 50 images[30]), while the first network will generally require between one and three GPU cards to process the 25 frames per second of a single camera.

If we increase the power of the GPU card to use both networks on the same card, the GPU memory required for real-time processing is the sum of the memory of the two networks[31] , both of which must be loaded together, and operate synchronously with the video streams. In this type of case, we can see that the need for hardware resources can grow very quickly, even to process a single video stream.

Which CCTV operator, managing hundreds of cameras, will be able to dedicate 3 GPU cards (3000 €, 900W) to each of his cameras?

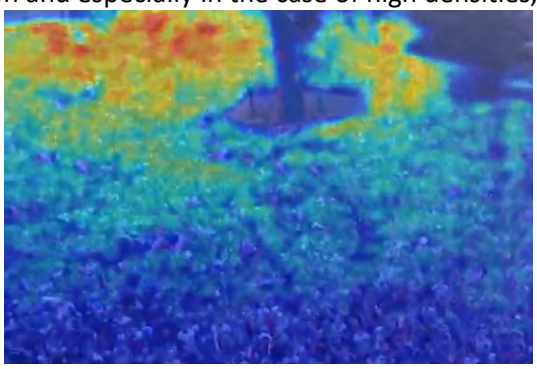
2.3- Density

[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] To avoid a loss of detection in case of cut of the image series just before the event to detect.

[31] A detection network requires in general several GB, the reference GPUcard cited in this paper has 8.

The most fruitful approach, in terms of density estimation and especially in the case of high densities, is not to detect each person or each head, but to dedicate a neural network specialized in the recognition of head features[32] , in order to be able to take into account the heads partially hidden[33].



In terms of execution time, this application is about twice slower than object recognition. A density measurement on a 500x500 image on an average GPU card with 6 to 8 GB of RAM, as seen previously, takes between 20 to 60 milliseconds[34].

2.4 – Motion analysis

The analysis of the relative movement of a set of objects and their background in a video, called optical flow, has been for long the subject of intense research in the field of image analysis. All prior results outside DNN have been completely superseded in quality by FlowNet in 2015 and its successors since then. Indeed, the equations of motion based on the relationship between contrast and motion could not always handle the areas of homogene (non-contrasted) regions well: the segmentation of moving objects lacked precision. Today, with these AI techniques, we can segment an object that moves in a self moving video, without even overflowing when this object passes over an area with the same color as its edge.

Alas, in terms of execution time, this application is much slower than object recognition. The processing of a 500x500 image to the next, on an average GPU card, as seen previously, is far slower than real time (thus far above 40 ms, for a 25 frames per second).

2.5 – Detecting differences between two images

There are still many applications addressed by the discipline, including for example the measurement of differences between two images[35] , which can be used for various purposes in security: to detect an object on an airport runway that could pierce the fuselage of an airplane, measuring the progress of a truck loading/unloading process on a logistic platform, detect a change on the side of a road, or an abandoned object, etc. In these applications, we try to detect differences, but we don't know what they are, so we can't look for a particular class of objects to search or to detect.

In fact, there is hardly any subject of image detection or analysis that is not approached through the prism of deep neural networks, so much so that some international conferences of the discipline international conferences in the field appeared in recent years to eliminate from their selection all studies based on "classical" techniques.

2.6 – Other applications

Many other applications exist which are not currently used in security. We can mention semantic segmentation (classifying pixels according to the nature of the object to which they belong, which allows for a fine clipping of objects instead of a bounding box), image alterations (improving resolution) [36] , or image synthesis[37], or detectors open to learning of new classes by some examples entered by the user etc...

---

[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] Here, a density heat map by Evitech Lynx.

[35] R. Daudt, B. Saux, et A. Boulch, *Fully Convolutional Siamese Networks for Change Detection*. 2018.

[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.

At a more global level, we can also find applications aiming at treating a problem without segmenting it into sub-problems: driving a car from a front camera view, or detecting an abnormality in a scene in general. The approach is delicate, because the link between the application building and its use is embedded, and one must trust the variety of the initial trainings, which one hopes will cover all possible cases. In practice, tests of this type of application in security field have shown that they are efficient on close-up views and simple repetitive actions: human/automaton interaction for example. We do not have good feedback on general situations.

They are therefore not used much in the security domain, where strong semantics are preferred, based on basic detectors with certified rates : an AND operator joining a 90% detector and a 100% reliable position estimator is at the end 90% reliable. Detection of an object in such position fully done by a DNN network requires a large amount of tests with this object at different positions which may be too costly to build to reach this reliability.

2.7 – Image analysis in "non-cooperative" environments

The recognition of a car or a pedestrian in the city is not much debated, in low to medium densities, and in good weather conditions: AI is usually able to easily classify urban targets, and an AI-based detector can be realized, at the cost of the underlying GPU.
However, there are limitations in very frequent use cases such as **umbrellas** when it is raining (the detection of hidden people is then made very difficult), or for dense crowds (where counting by detection is impossible), or when wearing clothes/masks that make the shapes of the object disappear (confusion/non detection), as well as decorative or advertising designs on vehicles which can induce false classifications.

On the other hand, **in applications intended for so-called "non-cooperative" uses, the observed targets will try not to be detected or recognized while crossing the detection zone. The AI is then inoperative**. The following contexts are generally cited:

- Preparation and perpetration of an intrusion in a site, or on a border,
- Escape from a place of detention (prison, asylum),
- Military applications.

On these uses, the targets to be detected will tend to be camouflaged (behind a cardboard box, an umbrella, in military clothes, under a fur), to crawl[38] , to pass at night when the color imagery is strongly degraded: the classification will not give any result in such situations.

In non-cooperative environments, the only contribution of AI is that of measuring the difference between two images, or the analysis of movement which do not make any assumptions about the shape sought.

It is thus well in the **public space**, where one observes multiple and openly visible objects (people, cars) in more complex scenes than the controlled situations of a closed site, that **AI brings new possibilities, and pushes back the limits of classical analysis[39]**.

---

[37] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. *Synthesizing images of humans in unseen poses*. In CVPR, 2018

[38] Or realize a crawling/rolling move, head orented toward camera, « rolling & crawling », to minimize their visible surface whilst slipping slowly on the ground.

[39] For detection of abandoned garbage objects in a street, for example, IA does not allow to classify object types that are too various (bags, mattresses, windows, rubble, …), but help to classify the car or the person from which the object is laid.

3- The economic impact on image analysis

3.1 – The use cases

Since 2005 at EVITECH, we equip sensitive sites with intelligent image analysis solutions from color and thermal cameras, for the detection of various incidents: intrusions, smoke, , oil leaks, behavior analysis, crowd control (counting), etc.

The recent development of urban color camera fleets has led a certain number of communities to source image analysis, first for the protection of municipal or departmental sites during closure periods[40], then for incident management in the city. AI provides the possibility of classifying targets in a more refined way than before, in particular to carry out. We will discuss the economic aspects in this section.

3.2 – The price factor

A color video surveillance camera costs on average between 50 and 1000 €, 50 € for a small fixed camera of entry-level, and 1000 € for a motorized camera adjustable and good quality. These are average prices generally observed on the markets where EVITECH operates. To be installed in a public space, this camera costs between 500 € and 10.000 €, depending on whether it is installed with 3 screws and wired "with a glue gun" to a guard's box in a low-cost housing complex[41] , or fixed by a steel gallows at the top of a mast in the city near a listed building, at the end of a trench that required the sidewalk or roadway to be pierced[42].

When operators are asked about the **perceived value of image analysis**, compared to the cost of the camera, the answers are generally **less than 50%**[43]. This cost includes the price of the software license, generally the number of video streams to be processed, and the cost of the hardware required to necessary to use it, CPU and GPU.

3.2 – The approach to “share for costs saving”

Given these constraints, public space video surveillance operators are not in a position today to dedicate hundreds of watts and thousands of euros per video stream to equip their cameras, even partially with intelligence of the Deep Learning type. As a software publisher, we must aim for a mutualization of computing resources and a parsimonious use of the GPU resources. Solutions must be designed that are capable of sharing it simultaneously between many video streams, to limit the financial impact.

One point deserves explanation here: when a GPU card has to execute the processing of a DNN it must be loaded into the card and this operation usually takes a few seconds. The network, once loaded, occupies part of the card's memory, and, if this is sufficient, it is possible to load two or even more networks (see §2.2) and to use them alternately and instantaneously on the card. When an image analysis configuration is active on a group of cameras, it is therefore preferable not to modify the allocation of the DNN networks to the GPU cards, except in exceptional cases like for start-stop operations, or for redundancy. The *sharing* mentioned above therefore consists of performing the same processing (e.g. recognizing objects) on images from different origins (different streams).

[40] EVITECH solutions analyze hundreds of thermal cameras over firemen sites protection.
[41] Estimation from RIVP social buildings.
[42] Estimation from Paris city.
[43] Qualitative enquire done by Evitech, in 2016, in the framework of the Eurostars Cro-magnon project, for a buying price without installation and configuration.

Some actors use AI in a systematic way (and permanently, as the basis of their analysis software) to continuously analyze the processed camera images (whose formats are commonly 1080x1080 or 1980x1080 pixels, or even larger). As GPU cards generally process formats smaller than or comparable to 500x500 pixels, this approach requires either to reduce the image size, and then to lose valuable information for the classification of distant and smaller objects in the scene, or to cut the image into 4 or 8 parts, and then to glue the pieces and the detections back together, which consumes 4 to 8 times more, and causes problems at the borders. Their applications therefore require, depending on the processing resolution and analysis frequency, relatively high and permanent GPU resources for few cameras, or a sacrifice of the available resolution.

On the other hand, in order to share these resources in the best possible way, our Jaguar software, operating on CPU, acquires the video stream and first detects new targets that appear in the image of a camera. Then, if the configuration allows it, it makes a request to the GPU to classify these new targets, consecutively to their appearance, and, according to the strategy, with potentially the image area at the best possible resolution, analyzing exactly the target area. **It does not require classification at night when there is no target[44] , and stops asking for this classification once it has established it.** It maintains this property with the tracking of the moving object in the image, even when the target becomes very small or is partially hidden by a wall or a car (in situations where the classification would not recognize it). The GPU is no longer required, and it is available to classify another object that appears on another camera. The benefit is twofold: more cameras are processed by the GPU card, and, during low traffic periods, very few calls to the GPU, and therefore, a reduced power consumption from the GPU.

This extraordinary capacity is an essential feature of a parsimonious approach: it relies on the continuous use of tracking at 100-200 € of CPU hardware cost per processed flow, and on the punctual use of classification at the appearance of the target (on GPU card, 1000 €). A silhouette on an urban route in the field of a camera persisting between 2 and 20 seconds, the savings per processed video stream processed video stream represents 98% to 99.8% of the GPU cost (300W, 1000 €) compared to a continuous detection by GPU.

These criteria have not yet emerged in the calls for tenders from market operators, but, considering the constraints of sustainable development, or more prosaically the evolution of energy costs, they will soon become determining criteria.
By sharing a single GPU card between 10 and 100 cameras (10 for mobile classification in a not very busy street, 100 for a crowd density measurement per camera every 20 seconds in a camera park), as we do with our Jaguar and Lynx software, we allow operators to access the benefits of AI with minimal economic cost.

3-3 – A tiny platform to save costs

In the NVIDIA range, we also used the JETSON Nano 4 GB card on its small autonomous module, whose initial price (around 130 €) and consumption (5W) were compatible with processing for one or even a few cameras. For networks of reasonable size, or at the cost of a *pruning* operation of the neural network, it is possible to classify a few images per second (which

[44] The GPU card then turns into sleep mode, and only consumes a few watts.

gives a processing time reference compared to other DNN based functions). Connected to a small computer consuming 15-25W and processing between one and 4 cameras, this card allows to compose, with a network link and a distribution of the processing, a very competitive platform for a small autonomous urban video terminal[45].

3-4 – Run on CPU to save costs

Since 2017, the "open" ONNX[46] format has been standardized and is being developed with the aim of constituting a standardized platform for implementing neural network models. This initiative has been joined by many actors, resulting in a variety of interpreters of this format, and in particular to allow the porting over different platforms.

One of them is simply the CPU, which is not very efficient (7 to 8 times slower than the GPU card mentioned above, on a dedicated half core i9 processor) but which allows to completely abstract from the GPU.

On applications measuring the density of people in a swimming pool by video analysis to control chlorination, or at the bottom of a ski lift to control the Ecodrive[47] mode, a density measurement every 15 or 30 seconds is sufficient and gives perfect satisfaction to our customersfor the control of their process

In these applications, the meter is a black box that uses the image to count and destroys it instantly. It outputs a number at regular intervals, which is then used to control the automatic process (chlorination or lift speed).

4- The right to use in public spaces in continental Europe

When a camera is installed in a public space in Europe, it is declared to the Authorities with a security objective, or a **purpose**, well specified.

It is not a question of providing a toy to the operator of the video center, for unlimited use on demand[48] , but to fulfill a precise mission: detect crowds in a busy place, detect congestion, forbidden stops, deposit of objects (waste), to ensure the quietness of the streets at night (to detect chases, rodeos, people running, ...), or to detect a general event like smoke or measuring flows, etc. These missions appear on the camera's declaration, and are inseparable from it.

However, there is a gap today between the available image analysis technologies and the law concerning the analysis of images in the public space: the analysis of images in the public space is considered from a legal point of view as an automated processing (and indeed it is the processing of a computer, therefore automated) on personal data of people who have not given their consent (and indeed, faces of passers-by in the street that are resolute enough to be recognized in a video are personal data), passers-by who are not asked anything when they cross the street.

This capability poses a risk to freedom considerations[49].

As the potential of image processing is significant, particularly in the area of biometrics, we are expecting changes in the French law that will define the context of use of image processing applications on the installed cameras, in order to allow certain applications (measurement of the water level of a river, smoke detection, counting of vehicles and passers-by, automatic controls and

[45] https://vdsys.fr/vigicam-ii/ : detection of abandoned waste, for example.
[46] https://onnx.ai/index.html
[47] Ecodrive : Automatic lift speed adaptation to attendance.
[48] Like following through the city video cameras specific people selected by the agent, e. g. people his knows personally.
[49] E. g. a feeling of harassment, if a police patrol appears each time more than 6 people gather for a residence.

detection of certain risk situations in certain specific places exposed to these risks) and prohibit abuses as we can see them in some countries that exploit, for example, a *supposedly racial recognition of faces*[50] , or which improperly detect any gathering as a threat to the established order.
In such a context, the image processing application will be declared with the camera, and any camera will not be able to have any kind of processing added to it.
It is likely that there will be a strong link between the declaration to authorities and the applications authorized to use the camera images: a camera placed to detect crowds can be enhanced with counting, whereas a camera placed as a tool to ensure quietness in a street, will be allowed to be improved by adding speed of movement measurement, within the limit of its declaration.

This will allow, as for the uses in the private and military sites, to optimize the configurations to reach configurations to reach 100% detection of the desired situations with a minimal false alarm or error rate[51] , two objectives that are at the heart of our culture since the creation of EVITECH.

[50] https://ipvm.com/reports/racial-ethnic-standards, IPVM, John Honovich, 2021.
[51] EVITECH Jaguar software has the iLIDS certification with the highest grade since 2013.