Video Surveillance and Privacy: A Solvable Paradox?

R. Cucchiara
University of Modena and Reggio Emilia, Italy

L. Baraldi
University of Modena and Reggio Emilia, Italy

M. Cornia
University of Modena and Reggio Emilia, Italy

S. Sarto
University of Modena and Reggio Emilia, Italy

***Abstract*—Video Surveillance started decades ago to remotely monitor specific areas and allow control from human inspectors. Later, Computer Vision gradually replaced human monitoring, firstly through motion alerts and now with Deep Learning techniques. From the beginning of this journey, people have worried about the risk of privacy violations. This article surveys the main steps of Computer Vision in Video Surveillance, from early approaches for people detection and tracking to action analysis and language description, outlining the most relevant directions on the topic to deal with privacy concerns. We show how the relationship between Video Surveillance and privacy is a biased paradox since surveillance provides increased safety but does not necessarily require the people identification. Through experiments on action recognition and natural language description, we showcase that the paradox of surveillance and privacy can be solved by Artificial Intelligence and that the respect of human rights is not an impossible chimera.**

VIDEO SURVEILLANCE concerns models, techniques, and systems for acquiring and processing videos about the external world, detecting targets along time and space, recognizing interesting or dangerous situations, generating real-time alarms, and recording meaningful data about the controlled scene. While the target of surveillance systems can be the whole environment, *e.g.* natural events or moving vehicles, the most complex and addressed target is surely the human being: where people are, what movements or actions they are performing, what is their behavior, and if it is affecting security or safety.

For most of the public debate, Video Surveillance and privacy are incompatible and form an intrinsic paradox. A *paradox*, in logic, refers to a statement claiming something that goes beyond (or even against) the "common opinion". Surveillance means control; control recalls Orwell's Big Brother and many modern examples of mass control against human rights. As a consequence, the debate created prejudicial vetoes on technological achievements and biased public opinion. At the same time, however, the debate has promoted good scientific practices oriented toward privacy preservation. Nowadays, regulations of many countries – though with varying emphasis – are taking the direction of harmonizing technology and privacy rights, as we will briefly see in the next section.

Privacy concerns the possibility to recognize and use the data of a single individual. Instead, Video Surveillance, in itself, does not concern singular people identification. The need for individual identification can rise as a consequence of an alarm, but if no alarm situations are detected the privacy of individual identities and the right of each individual to be free in his behavior must be preserved. Nevertheless, as disentangling vision and individual recognition is not straightforward, Video Surveillance has been, and still is, involved in privacy issues. This was unavoidable in the XX century when surveillance systems were governed by human inspectors only and stored video frames that could be later retrieved and used against the privacy of the depicted persons. At the beginning of the XXI century, instead, Computer Vision advancements made automatic processing effective both in surveillance and in privacy-preserving solutions. Most of these technological advancements were achieved after the September 11 tragedy, when many research centers and big companies put their effort into human detection, tracking, re-identification, and action recognition.

Most privacy-preserving approaches were oriented to *visual anonymization*, achieved by covering face appearances on pictorial data. Also, *pseudo-anonymization* techniques were developed, such as encryption and data scrambling, to store privacy-concerning information in a way impossible to be retrieved by the human eyes. Only in the last decade of the AI season, with the happy wedding between Computer Vision and Deep Learning, many scientific results have been carried out in the direction of human behavior understanding without affecting privacy rights.

In this paper, we propose an overview of the main results of this fifty-year journey of Video Surveillance and, at the same time, of the concern and the request for privacy. After a historical review, we concentrate on some issues in two scenarios. The former is when the training dataset of Machine Learning algorithms is known and can be anonymized or substituted with synthetic data: we show some state-of-the-art results which provide privacy-by-design and privacy-by-default solutions in detection and action recognition. The latter, instead, concerns the use of foundation or large-scale models in which training data is not accessible, and we discuss how it is possible to avoid privacy violations in new tasks such as the automatic description of human scenes.

ABOUT PRIVACY REGULATION

The concept of privacy is a cultural trait and, in accordance with S. Rodotà, "there is a constant relationship between changes in information technologies and changes in the concept of privacy, which is, in fact, a subjective concept that varies according to subjects, historical moments, and places." [1].

A milestone in privacy definition has been, in the late XIX century, the Warren and Brandeis document on "The Right to Privacy" [2] in which, for the first time, private right was constituted as fundamental in civil society. However, only in 1970, the US presented "the Privacy Act" as a still-used federal reference. In the European Union, the "Charter of Fundamental Rights of the European Union" (Nice Charter) has been declared in 2000 as an essential reference point for the constitutional framing of the right to privacy in Europe. In the Charter, it is protected not only the more generic right to privacy (regarding private life) but also the more specific right to protection of personal data. It was substantially similar in all occidental countries, but in the aftermath of the events of September 11, 2001, the US dealt a severe blow to the freedoms and civil rights of American citizens due to the so-called "War on Terror" emergency.

The debate between individual and social security and privacy has never ended since, and it has become more acute in the last 15 years when, in addition to possible government control, a very large control is being carried out by American and Asian corporations through the (although legal) use of social data of the entire planet. This is true for all types of personal data, but it is even more critical for visual data, which contains immediate personal information about human activity, interaction, and behavior.

Nowadays, the debate on privacy is very active: on one hand, the improvements in Computer Vision technologies made automated surveillance so accurate to be practically usable in many public and private areas. On the other hand, privacy is becoming an essential right that is assured by many legislations such as the European GDPR (General Data Protection Regula-

Table 1. International regulations.

Regulation	Summary
EU GDPR 2018	The General Data Protection Regulation regards all private and public entities acting in Europe and concerns data collection (principle of minimization), the requirement of consent, the access and erase of private data, the restriction of processing, and the concepts of "privacy-by-design" and "privacy-by-default" (*i.e.* the need of privacy in designing state-of-the-art technologies, and the fact that privacy must be a standard, a default, during their use).
EU AI ACT 2021	The Artificial Intelligence Act is an EU proposal since 2021, now under approval, which regards the regulation of all AI-based systems used in Europe, when adopted for "high risk" applications which must provide requirements of human oversight, privacy, rights respectful, transparency and accountability.
US CCPA 2020	The California Consumer Privacy Act is the first privacy law in the US to give consumers control over personal information. Similar to GDPR, it regards business companies in California and concerns opt-out, data access, and non-discrimination for rights.
CHINA CSL 2017 and PIS 2020	The Cybersecurity Law and the Personal Information Security specification regulate private information in China. They regard all types of organizations and concern with the right to be informed of rights before the use and collection of private data.
JAPAN APPI 2017	The Act of Protection of Personal Information regulates individual interests in privacy. Since 2019 it has formal recognition from the European Commission and regards all business operators in Japan, forced to public data collection purposes and data minimization criteria.

tion), the Californian law, and also to some extent the China legislation (see Table 1). Meanwhile, events like the use of facial recognition during Black Lives Matter protests in 2020 have raised concerns associated with the recording of images in public places. Following a strong backlash from society, companies like IBM, Microsoft, and Amazon stopped selling software to government authorities.

Many worldwide regulations concern the principle of data minimization (*i.e.* the need of using a minimal amount of data), of opt-out (*i.e.* the option to delete data if required by the data owner), the limitation of data storage (*e.g.* in GDPR) and their inappropriate use. New proposals such as the AI ACT, recently approved by the European Commission, instead, make explicit reference to Machine Learning technologies and regulate the deployment and usage of high-risk applications (*e.g.* health, security, enrollment, education, and finance) in order to assure the key-points of *Trustworthy AI – i.e.* the principles of human oversight, transparency, accountability, compliance with human rights and privacy. They pose several limitations on the design of new AI-based applications such as Video Surveillance.

Finally, we shall underline the difference between anonymization and pseudo-anonymization, according to GDPR: *anonymization* is a process that transforms personal data into anonymous data in such a manner that the data subject is not or no longer identifiable; *pseudo-anonymization* is "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific subject without the use of additional information, provided that such additional information is kept separately and in a lawful manner". According to these definitions, data encryption and data scrambling are procedures of pseudo-anonymization, which are not considered in this work.

A JOURNEY IN VIDEO SURVEILLANCE AND PRIVACY

Video Surveillance has been and is a very active research area, with more than 18.000 scientific papers written on the topic from 2000, according to Google Scholar. In the following, we outline the most important achievements in Video Surveillance and privacy from a historical perspective, ranging from early analog cameras to state-of-the-art Deep Learning techniques. This is also done by screening some of the most relevant scientific works, selected according to the relevance of the scientific venue and impact on the community.

Before 2000

Looking back in hindsight, Video Surveillance as it is today thanks to AI technology was really an impossible challenge in the last century: no computer power for efficient real-time video processing; no pattern recognition algorithms nor effective neural models to detect people or understand their movements; no enough annotated data to allow any Machine Learning-

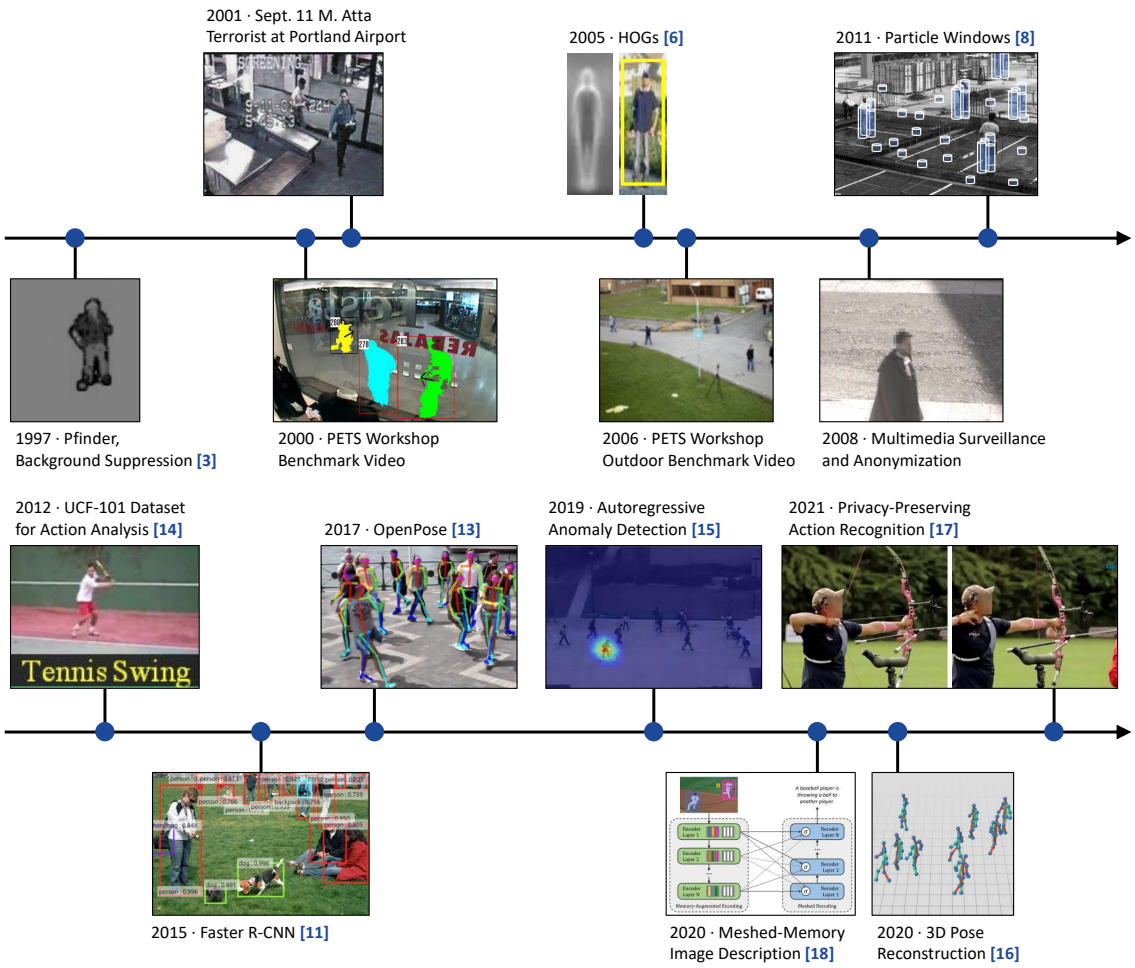

Figure 1. Timeline of the major surveillance and privacy-preserving approaches.

based parameter tuning. However, the problem of surveillance and privacy was already under the focus of discussion since camera systems were available to allow remote human-based control of a scene in CCTV (Closed Circuit Television) systems. A first generation started early: in 1969, police analog cameras were already installed in the New York City Municipal Building near City Hall and connected with VCRs (Video Cassette Recorders). In the '80s, the first military and space research on image processing created synergies with the incoming Digital CCTV systems: in 1985 the first DVR (Digital Video Recorder) was adopted for public place surveillance with some image processing algorithms for noise-cleaning and data enhancement.

After the first attack on the World Trade Center in Feb. 1993, the New York Police Department, FBI, and CIA installed surveillance cameras throughout the area. All over the world, analog interlaced CCTV cameras and computers with video capture cards started to record and store compressed images and make them available on LAN via built-in web servers. At that time, Video Surveillance regarded only the capacity of digital systems to show remote video to human controllers and this, of course, was possibly affecting privacy. In 1998, for instance, New York Civil Liberties Union (NYCLU) published the results of a study on public surveillance cameras, counting 2,397 cameras in Manhattan. In Europe, several research efforts in Computer Vision were made to monitor vehicles on roads (for the so-called Intelligent Transport Systems), and initial experiments were performed to automatically find people in videos. One of the first

key projects has been Pfinder [3] in 1997. In that decade several studies achieved important results in geometry (*e.g.* for camera homography reconstruction), optical flow analysis, face detection, and recognition. The possibility to detect faces and recognize their identities started discussions about privacy issues, but relatively few were related to "automated" surveillance, as this was still far from being considered a real application.

The 2000s: The Boom of Computer-Vision Video Surveillance

In the first years of the XXI century, Video Surveillance spread out for three concurrent reasons: hardware availability, Computer Vision improvements, and the need for social security created by the terrorist attacks in the US and Europe. From the hardware side, a third generation of surveillance systems was born and equipped with IP cameras, LAN video servers, and RISC processors for embedding the pre-processing on "smart cameras", even with PTZ (Pan, Tilt, and Zoom) features. JPEG and then MPEG4 and H264 compression allowed a fast video data transfer. In 2006, Chicago launched the first extensive Video Surveillance network (the "Chicago Virtual Shield") and then in 2011 Chongquin did the same (the "Chongquin Shield") with more than 500,000 cameras and 200 million RFID tags on cars. The social need for Video Surveillance increased the social acceptance of privacy-related issues and accelerated the transfer between scientific results and real systems. In 2001, after terrorists passed through US airports without being recognized by any system, massive experimentations with prototypical person-detection and facial-recognition systems began; this fueled the debate on the privacy-surveillance paradox, and Benjamin Franking's famous phrase about the trade-off between security and freedom [4] became the talk of the town.

Computer Vision-based (also called *smart*) Video Surveillance Systems become widespread for providing 24-hour monitoring of camera video streams, which would have been ineffective and expensive to be done manually. The goal was threefold: to alert security officers of thefts and suspicious persons, to support post-crime forensics, and at least to achieve deterrence. In the meantime, Computer Vision topics that are still scientifically open problems, such as people detection and tracking, and multiple target surveillance under occlusion, shadows, and other artifacts become popular.

Video Surveillance needs have driven two important Computer Vision achievements: (a) the development of *background suppression* techniques and (b) the formulation of *people detection as a classification problem*.

The approaches for real-time surveillance from fixed installed cameras were (and still are) based on the "background suppression" paradigm [5]: it detects targets as moving visual objects, creating a dynamic reference background and segmenting them out, possibly with a distinction from artifacts and shadows. Special efforts were also devoted to multiple people tracking in outdoor spaces, also with multiple-camera acquisition and PTZ cameras. Famous in these years were the PETS benchmark datasets released at the participants of the homonymous PETS (Performance Evaluation of Tracking and Surveillance) workshop[1] since 2000 (see Fig. 1).

We could conceive these methods as the archetypes of current *bottom-up* approaches where data was processed from pixels to regions or shapes: then shapes were processed to distinguish people (*e.g.* by head detection) from other foreground shapes. Concurrently, *top-down* approaches for people detection started to be formulated, by considering people as target models recognizable by a two-class classifier (human presence vs. non-human presence). The birth of people detection started probably with the seminal work of Dalai and Triggs in 2008 [6]. Since that work, hundreds of approaches concerning people were developed. All can be categorized according to three different aspects:

1) Which (hand-crafted) features to employ? Several general purpose descriptors were proposed, such as HoGs [6], covariance and structured part-based descriptors [7];

2) Which classifier? Detectors should be coupled with suitable classifiers such as Neural Networks, SVMs (Support Vector Machines), LogitBoost, AdaBoost, showing a true boom of pattern recognition techniques for people detection.

[1] http://www.cs.cmu.edu/~vsam/Conferences/PETS2000/pets2000.html

3) Which search space? Often searching everywhere is not necessary; thus many proposals focused on improving both efficiency and precision/recall, *e.g.* with *sliding-window* or hierarchical (pyramidal or multi-resolution) windows. As well, the idea of *region proposals* started to be defined, in order to look at regions according to a probability density function $p(X|Z)$ being X the state of the person and Z the observation. For instance, in [8], a multi-stage particle window provided fast and accurate multi-stage probabilistic sampling for boost and SVM classifiers.

These methodologies, also combining background suppression and people detection, allowed the implementation of Video Surveillance systems that proved their effectiveness in many real implementations. They concerned real-time monitoring but also a-posterior analysis for summarization in handling large security issues. An example was the synopsis approach used in the Boston marathon bombing of 2013 [9].

In this decade of the Video Surveillance boom, the paradox of coping with privacy and security was always debated. In 2008 there was more than 4.2 million CCTV cameras installed in London, and a famous interview in the Guardian reported that the Chief Inspector of London police defined surveillance as an "utter fiasco" which was effective only as a preventative measure. The quality of images was too poor for humans and automated algorithms to identify abandoned backpacks and people action analysis was not effective at all. In Italy, the *FreeSURF* project[2] funded in 2006 by the Ministry of Research, involved several universities to develop "a paradigm for the new generation of Video Surveillance systems, free from the control by human operators, and completely respectful of the privacy". In this case, an efficient pipeline for surveillance, motion analysis, people detection, and tracking was followed by the state of the art of classifiers and face detectors to anonymize faces and identities. A large effort started to be devoted to the development of measures and benchmarks since people and face detection were promising but far from today's results.

[2]https://aimagelab.ing.unimore.it/freesurf/

The Last Decade: Deep Learning, Surveillance, and Privacy

In the latest decade, the paradox between Video Surveillance adoption and privacy was exacerbated by the re-birth of Neural Networks and the incredible results of Deep Learning after the ImageNet challenge [10] and the development of Convolutional Neural Networks capable of classifying, recognizing and detecting targets. While Video Surveillance systems become spread and widely used, the debate on privacy and human rights shifted towards social data collection and dual-use of data, culminating in scandals like the Cambridge Analytica one in 2018. The technology of people detection, tracking, recognition, and action analysis become very sophisticated thanks to several results such as:

- *The use of learned features coupled with learned classifiers*. Some past ideas such as region-proposals, low-level and high-level feature representations, and human models were used with success. From Faster R-CNN [11] to YOLO [12], now people can be detected with very high precision also in large areas and crowded scenes. Recent architectures provide compelling results, although being based on supervised learning, they are very dependent on the training dataset.
- *Accurate pose estimation*, after the appearance of OpenPose in 2018 with learned keypoint heat maps [13], which allows finding people in both bottom-up or top-down approaches.
- *Action analysis* become a very large research field, with algorithms looking at single segmented and detected targets and also at the image or video as a whole. Results in action analysis become impressive, especially after the availability of annotated datasets such as UCF-101 [14].
- *Anomaly detection in video for human analysis* becomes doable, again thanks to annotated datasets and the improvement of auto-regressive and autoencoder-based architecture [15].
- *Human and scene text descriptions* become available with the automatic generation of captions: captioning focused on Deep Learning-based visual feature extraction followed by recurrent networks and then on self-attentive ar-

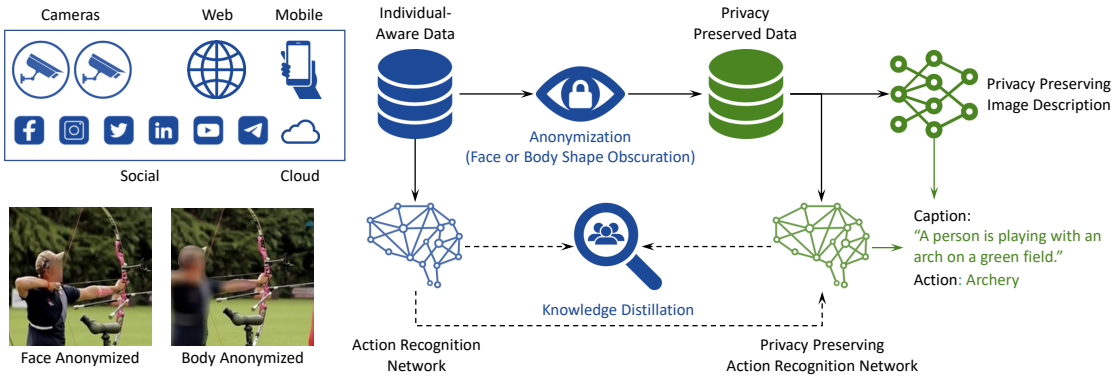

Figure 2. Privacy-preserving action recognition and image description pipeline. Networks are trained on anonymized data collected from multiple sources, and by distilling knowledge from networks trained on sensitive data.

chitectures, also enriched with memories [18].

- *The shift towards 3D space reconstruction, in people detection and tracking*, by extracting knowledge in a three-dimensional space where occlusions and perspective errors can be avoided [16].

On the other side, and this is the interesting paradox, public opinion, especially in US and Europe, is becoming exacerbated against the use of Machine Learning for real-time monitoring in public and private areas. The term "Video Surveillance" has been almost banished in recent research works, not to recall possible privacy or human rights concerns. Surely the effectiveness of Video Surveillance systems has been proved in recent years and also adopted against the democratic expression of citizens. After the Hong Kong 2019 protests of students using umbrellas to protect themselves from surveillance and face identification systems, in many parts of the world the use of AI-based recognition systems has been banned. For instance, the next European AI Act regulation prohibits the use of remote surveillance for actions against democracy.

ABOUT PRIVACY-PRESERVING ACTION RECOGNITION

Have we therefore been working in vain for decades of research? No. The challenge is, indeed, to provide useful tasks in surveillance, from people detection to action analysis in dangerous situation recognition without affecting singular individual privacy. Hundreds of papers have been proposed about pseudo-anonymization, *i.e.* the task of modifying images to be not visible by human experts, leaving inside useful semantic content for a task. Examples of image processing approaches are pixelization, pixel scrambling, or shape hiding in video. This is one side of the problem, which relates to the human controller who should not be aware of the identities of the people being monitored. Clearly, though, the raw footage might be saved before anonymization for authorized personnel. These methods will allow to work on data using standard methods and avoid privacy issues during their use in a privacy-by-default approach. The main and more critical challenge is to understand if new neural networks systems can fully comply with privacy-by-design principles, *i.e.* if they can: (a) understand people's activity in visual data without exploiting information regarding individual recognition; (b) learn to recognize actions and provide surveillance tasks without being trained on sensitive data.

The first step has been recently addressed for object and people detection [20] and by exploring how trained systems can deal with activity recognition by working on anonymized data only [19]. The second step is more critical: one possible solution is to work with synthetic data. This has been provided for single tasks such as people detection and tracking [16], but it is harder for action recognition since it requires costly synthetic datasets coping with the large variety of actions, that instead are currently collected on real data. The alternative to synthetic data is working on previously anonymized videos. In principle,

Table 2. Action recognition and image description performance on anonymized (blurred) images, using Knowledge Distillation (KD) to reduce performance drop. Results are reported on Kinetics-400 and the subset of COCO containing humans, showing the performance gap in terms of top-1 accuracy for action recognition and CIDEr score for captioning with respect to the same model trained without anonymization.

			Action Recognition			Image Description					
Anonymization		KD	top-1	top-5	$\Delta_{\text{top-1}}$	B-1	B-4	M	R	C	Δ_{C}
□	none	-	69.2	88.1	-	80.3	41.2	30.3	60.1	124.6	-
◩	faces	-	68.6	87.7	-0.6	80.2	40.8	30.0	59.7	123.0	-1.6
◩	faces	✓	70.3	88.8	+1.1	80.7	41.3	30.3	56.0	124.8	+0.2
■	full body	-	65.7	85.5	-3.5	77.9	37.9	28.9	57.8	115.2	-9.4
■	full body	✓	68.5	87.9	-0.7	78.4	38.6	29.1	58.2	116.7	-7.9

recognition of a person's actions in a scene is independent of the person's face, and possibly other aspects that might be useful for identification or re-identification, such as clothing. A question that is not completely answered is how much trained systems can learn from such covered data. An example of an answer is given by our experiments with knowledge distillation approaches in Table 2 (see also Fig. 2), where we experiment by blurring faces or full bodies. In this case, we train a standard R(2+1)D action recognition network on obfuscated videos from the Kinetics-400 dataset[3], while distilling knowledge from a network which has been trained on non-obfuscated data [17].

The lesson learned is that, although an ultimate solution is not fully available, it is possible to recognize actions without loss of accuracy even when faces or full bodies are obfuscated.

ABOUT PRIVACY-PRESERVING IMAGE TEXTUAL DESCRIPTION

The way of working on privacy compliant data and providing useful tasks for people surveillance and action analysis could be cumbersome when systems are starting to deal with foundation models, or in general large pre-trained systems that provide textual descriptions of a scene. Image captioning is a new way of generating image descriptions in a natural language way. Also in this case, networks can be trained on anonymized visual data and reduce the loss in performance using knowledge distillation. The results we propose in Table 2, where we employ a regular Transformer-based encoder-decoder and the same setting previously outlined, show that there is a very small decrease in performance when faces are obscured during training, measured with standard captioning evaluation metrics [18] such as BLEU-1 (B-1), BLEU-4 (B-4), METEOR (M), ROUGE (R), and CIDEr (C) on the subset of the COCO dataset[4] containing humans. A larger decrease in performance is measured when most of the people's shapes are blurred. Again, the gap can be reduced by using knowledge distillation from a captioning model trained on non-obfuscated images.

Regarding the textual output, instead, there is no manner to fully control the output that is generated in a decoding step of self-attentive architectures. Few attempts have been provided to control the input and focus on some details of the scene. For example, some works [18] have proposed solutions to describe the activity of people in a controlled manner after a detection step. In this case, the network is trained to focus only on the detections given in the control signal and to describe them in a sequential way. But what happens if now also the identity of a person could be recognized by the amount of data used on training (which is often uncontrollable in pre-trained foundation models)? An approach that could be explored in the next future is to impose not only the controllability of the input and encoding step but also provide controllability for privacy preservation in the generation, by teaching the network to be privacy-compliant (*e.g.* to not mention proper names and identity-disclosing items). This is an important direction of future research for adding some constraints in text generation in a generation of trustworthy, fair,

[3] https://www.deepmind.com/open-source/kinetics

[4] https://cocodataset.org/#captions-2015

and privacy-compliant AI.

DISCUSSION AND FUTURE ISSUES

In this article, we presented an overview of the main achievements that Computer Vision has made in Video Surveillance, focusing on its privacy-related aspects. Drawing from a review of the privacy regulations, we have outlined how Surveillance can be achieved without employing people's identities and showcased an experimental study on privacy-preserving action recognition and natural language description.

In conclusion, the paradox of surveillance and privacy could be solved, actually. First, now video understanding can be effectively provided by Machine Learning approaches and there is no need for continuous human monitoring. Human oversight is necessary only in case of dangerous situations and only for those individuals that could be considered to have dangerous behavior. Second, to understand what people do we do not need information about their identity, their face, or their appearance. Third, we can start discussing the controllability of privacy for pre-trained networks, constraining them to give answers in both privacy-by-design and privacy-by-default methods. New attempts show that this way could be achievable and we hope that this will be the future of AI-based systems: to be designed for human well-being and thus for human security and safety too, without affecting human rights and in particular the freedom of privacy.

REFERENCES

1. S. Rodotà, "La privacy tra individuo e collettività," *Politica del Diritto*, vol. 5, pp. 548, 1974.
2. S. Warren, and L. Brandeis, "The right to privacy," *Killing the Messenger, Columbia University Press*, pp. 1-21, 1989.
3. C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. on PAMI*, vol. 19, n. 7, pp. 780-785, 1997.
4. K. W. Bowyer, "Face Recognition Technology: Security versus Privacy," *IEEE Tech. and Soc. Mag.*, vol. 23, n.1, pp. 9-19, 2004.
5. A. Elgammal, D. Harwood, and L. Davis, "Non-Parametric Model for Background Subtraction," in *Proc. of ECCV*, pp. 751-767, 2000.
6. N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of CVPR*, pp. 886-893, 2005.
7. P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," in *Proc. of BMVC*, pp. 1-11, 2009.
8. G. Gualdi, A. Prati, and R. Cucchiara, "Multistage Particle Windows for Fast and Accurate Object Detection," *IEEE Trans. on PAMI*, vol. 34, n. 8, pp. 1589-1604, 2011.
9. E. Goralnick, P. Halpern, S. Loo, J. Gates, P. Biddinger, J. Fisher, G. Velmahos, S. Chung, D. Mooney, C. Brown, B. Barnewolt, P. Burke, A. Gupta, A. Ulrich, H. Hojman, E. McNulty, B. Dorn, L. Marcus, and K. Peleg, "Leadership During the Boston Marathon Bombings: A Qualitative After-Action Review," *Disaster Medicine and Public Health Preparedness*, vol. 9, n. 5, pp. 489-495, 2015.
10. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. on Comp. Vis.*, vol. 115, n. 3, pp. 211-252, 2015.
11. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. on PAMI*, vol. 39, n. 6, pp. 1137-1149, 2017.
12. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. of CVPR*, pp. 779-788, 2016.
13. Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *Proc. of CVPR*, pp. 7291-7299, 2017.
14. K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild", 2012.
15. D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent Space Autoregression for Novelty Detection," in *Proc. of CVPR*, pp. 481-490, 2019.
16. M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, and R. Cucchiara, "Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation," in *Proc. of CVPR*, pp. 7204-7213, 2020.
17. M. Tomei, L. Baraldi, S. Bronzin, and R. Cucchiara, "Estimating (and Fixing) the Effect of Face Obfuscation in Video Recognition," in *Proc. of CVPR Workshops*, pp. 3263-3269, 2021.
18. M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *IEEE Trans. on PAMI*, vol. 45, n. 1, pp. 539-559, 2023.
19. I. R. Dave, C. Chen, and M. Shah, "Spact: Self-

supervised privacy preservation for action recognition," in *Proc. of CVPR*, pp. 20164-20173, 2022.
20. B. Ma, J. Wu, E. Lai, and S. Hu, "PPDTSA: Privacy-preserving deep transformation self-attention framework for object detection," in *IEEE Global Communications Conference*, 2021.

Rita Cucchiara is a full professor at the University of Modena and Reggio Emilia and Director of the Artificial Intelligence Research and Innovation Center, coordinating the AImageLab research lab and the ELLIS unit of Modena. For more than 30 years works in Computer Vision, pattern recognition and Machine Learning, mainly for human behavior understanding, coauthoring more than 500 publications in the field. Contact her at rita.cucchiara@unimore.it.

Lorenzo Baraldi is a tenure track assistant professor with the University of Modena and Reggio Emilia. He was a Research Intern at Facebook AI Research (FAIR) in 2017. He has coauthored more than 80 publications in scientific journals and international conference proceedings, mainly working on video understanding, Deep Learning, and Multimedia. Contact him at lorenzo.baraldi@unimore.it.

Marcella Cornia is a tenure track assistant professor with the University of Modena and Reggio Emilia. In 2022, she received the ECVA PhD Award for the results achieved during her PhD. She has coauthored more than 50 publications on vision-and-language integration, attentive and saliency models, and Computer Vision solutions for fashion. Contact her at marcella.cornia@unimore.it.

Sara Sarto is currently pursuing the PhD degree in Information and Communication Technologies at the University of Modena and Reggio Emilia. Her research interests include image captioning, cross-modal retrieval, and attentive models. Contact her at sara.sarto@unimore.it.