Mauro Barni, Patrizio Campisi, Edward J. Delp, Gwenaël Doërr, Jessica Fridrich, Nasir Memon, Fernando Pérez-González, Anderson Rocha, Luisa Verdoliva, and Min Wu

Information Forensics and Security

A quarter-century-long journey

©SHUTTERSTOCK.COM/TRIFF

Digital Object Identifier 10.1109/MSP.2023.3275319
Date of current version: 14 July 2023

Information forensics and security (IFS) is an active R&D area whose goal is to ensure that people use devices, data, and intellectual properties for authorized purposes and to facilitate the gathering of solid evidence to hold perpetrators accountable. For over a quarter century, since the 1990s, the IFS research area has grown tremendously to address the societal needs of the digital information era. The IEEE Signal Processing Society (SPS) has emerged as an important hub and leader in this area, and this article celebrates some landmark technical contributions. In particular, we highlight the major technological advances by the research community in some selected focus areas in the field during the past 25 years and present future trends.

Introduction

The rapid digitization of society during recent decades has fundamentally disrupted how we interact with media content. How can we trust recorded images/videos/speeches that can be easily manipulated with a piece of software? How can we safeguard the value of copyrighted digital assets when they can be easily cloned without degradation? How can we preserve our privacy when ubiquitous capturing devices that jeopardize our anonymity are present everywhere? How our identity is verified or identified in a group of people has also significantly changed. Biometric identifiers, used at the beginning of the 20th century for criminal investigation and law enforcement purposes, are now routinely employed as a means to automatically recognize people for a much wider range of applications, from banking to electronic documents and from automatic border control systems to consumer electronics.

While the issues related to the protection of media content and the security of biometric-based systems can be partly addressed using cryptography-based technologies, complementary signal processing techniques are needed to address them fully. It is those technical challenges that gave birth to the IFS R&D community. Primarily driven, at their early stage, by the need for copyright protection solutions, IFS contributions were published in various venues and journals that were not dedicated to this area. Although some dedicated

conferences (SPIE/IST Conference on Security, Steganography, and Watermarking of Multimedia Contents; ACM Multimedia and Security Workshop; and ACM Workshop on Information Hiding and Multimedia Security) emerged, this nascent community lacked a well-identified forum where researchers, engineers, and practitioners could exchange the latest advances in the area, which is multidisciplinary by nature. A call for contributions to *IEEE Transactions on Signal Processing* in 2003 attracted enthusiastic responses to fill three special issues on secure media. It was time to create a platform to advance the research and technology development of signal processing-related security and forensic issues.

To foster broader community building and strive for a bigger and lasting impact, a collective effort by a group of volunteer leaders of the SPS charted a road map in 2004 for creating *IEEE Transactions on Information Forensics and Security* (*T-IFS*) and a corresponding IFS Technical Committee, both of which were launched in 2006. It was written in the proposal to IEEE that the new journal would aim at examining IFS issues and applications "both through emerging network security architectures and through complementary methods including, but not limited to: biometrics, multimedia security, audio-visual surveillance systems, authentication and control mechanisms, and other means." A few years later, in 2009, the first edition of the IEEE Workshop on Information Forensics and Security was held, in London, U.K.

The IFS community has established a strong presence in the SPS and is attracting submissions from a variety of domains. In view of the page budget allocated to this retrospective article, rather than surveying, exhaustively but briefly, each individual IFS area, we opt for a more focused review of selected domains that experienced major breakthroughs over the past 25 years and that are expected to be more aligned with the technical background of the *IEEE Signal Processing Magazine* readership. While this choice does imply that some IFS areas will not be covered, it should not be taken as any indication that some IFS contributions are more welcome than others. We hope that the following few pages will give the readers a flavor of what happened in this field as well as the specifics of the mindset underpinning this research area.

Information forensics and security is an active R&D area whose goal is to ensure that people use devices, data, and intellectual properties for authorized purposes.

Digital watermarking

In the late 1990s, MP3 song exchanges on peer-to-peer file sharing networks and DVD ripping increased piracy concerns. In this emerging digital interconnected world, generational copies became perfect clones that could be efficiently distributed worldwide without being burdened by the shipping logistics of the old analog world. Digital watermarking was introduced in this context to complement traditional cryptography-based solutions and provide a second line of defense. The essence of this technology was to introduce imperceptible changes in media content—should it be audio, images, video, text, or something other—to transmit information that could later be recovered robustly, even if the watermarked content had been modified [1].

This new research area rapidly attracted contributions from related domains: perceptual modeling, digital communications, audio/video coding, pattern recognition, and so on. Early watermarking methods used very simple rules, e.g., least-significant-bit replacement, thereby providing almost no robustness to attacks. Significant progress was made when the IFS research community realized that the retrieval of the embedded watermark could be framed as a digital communications problem. A seminal watermarking contribution, coined as spread-spectrum watermarking [2], leverages a military communications model well known for its resilience to jamming. The underlying principle is to spread each watermark bit across many dimensions of the host media content to achieve robustness; for a given bit $b \in \{\pm 1\}$ to be embedded and an input (host) n-dimensional vector $\mathbf{x}$, additive spread-spectrum outputs a watermarked vector $\mathbf{y}$ such that

$$\mathbf{y} = \mathbf{x} + b\mathbf{w} \tag{1}$$

where $\mathbf{w}$ is an n-dimensional carrier secret to adversaries. Spreading is achieved when $n \gg 1$. Due to intentional and inadvertent attacks (e.g., content transcoding), a legitimate decoder (that knows the carrier $\mathbf{w}$) gets access only to a distorted version of $\mathbf{y}$ from which it must extract the embedded bit b with the highest possible reliability. By further exploiting connections with statistical detection and coding, it has been possible to derive optimal ways to extract the embedded watermark information for various hosts and increase robustness using channel coding. It should be kept in mind, however, that watermarking deviates from standard communications theory in that

1) The watermark embedding process must remain imperceptible, so standard power constraints (i.e., $\|\mathbf{w}\|^2/n \ll \|\mathbf{x}\|^2/n$) are not sufficient.
2) The wide range of attacks that the watermark is expected to be resilient to exceeds the typical channel distortions and jammers.

Spread-spectrum watermarking fundamentally assimilates the host media content $\mathbf{x}$ to interference on the underlying low-powered watermark transmission $b\mathbf{w}$, which can, at best, be modeled statistically.

A breakthrough came about by accounting for the fact that the media host is fully known at the time of watermark embedding (but not watermark detection) and that an alternative communications model (with side information) can be used to reject the interference of $\mathbf{x}$ completely. Side-informed watermarking is typically instanced through quantization-based watermarking [3]. In this case, the watermarked vector is obtained as

$$\mathbf{y} = Q_b(\mathbf{x}) \tag{2}$$

where $Q_b(\cdot)$ is a secret (to adversaries) vector quantizer that depends on the embedded bit of information and is designed in such a way that $\mathbf{y} - \mathbf{x}$ meets a perceptibility constraint. The theoretical basis for side-informed watermarking was derived from Gel'fand and Pinsker's random-binning idea [4], which relies on an auxiliary random variable as a proxy for trading off source and channel distortion. Especially influential to the IFS community was Costa's application of the random-binning paradigm to side-informed Gaussian channels that introduced the key element of distortion compensation in his construction of the auxiliary variable [5]. Following Costa's catchy title, the concept of coding with side information at the transmitter came to be known as "dirty paper coding" (DPC). The rescue from the oblivion of DPC, now so prevalent in wireless communications, undeniably owes partly to watermarking research.

While the communications model serves as the core engine, the generic blueprint of a watermarking system typically requires additional components: a signal transformation to map the host media onto a multidimensional feature space that is robust to distortion, a powerful resynchronization framework to compensate for the misalignment experienced by the watermarked content, key-seeded pseudorandom mechanisms to obfuscate inner mechanics of the system to nonauthorized parties, and so on. This final item reveals a salient aspect of watermarking. A hostile adversary who wants to disrupt watermarked communications may be present, especially when watermarking is used for copyright protection. In that case, there is an incentive for pirates to strip the watermark that prevents them from accessing premium content or that somehow encodes their identity.

Therefore, a sizable research effort has been dedicated to characterizing how to prevent such an adversary from detecting, estimating, tampering, and/or removing the watermark signal. For instance, it has been shown that specific measures should be taken to prevent watermark information leakage when the adversary has the opportunity to observe several watermarked assets [6]. On another front, research contributions have highlighted the risks of making the watermark detector publicly accessible. In that case, the adversary could devise powerful strategies to disrupt the watermark, thanks to the availability of a reliable oracle [7]. This is particularly relevant when watermarking is used for copy and playback control, and relevant countermeasures must be implemented. More critically, findings from deployments for live video distribution revealed that pirate operators might blend different sources of the same video stream, each one with its watermark, to generate the video they distribute. Such adversarial behaviors, which have long been thought to be academic mind games, require the introduction of dedicated coding mechanisms for the watermark to survive, such as, for instance, anticollusion codes [8], [9].

Content protection application use cases have historically driven research on digital watermarking. For instance, in the late 1990s, the Content Protection Technical Working Group and the Secure Digital Music Initiative considered watermarking to implement a copy control mechanism for DVDs and music. Still, the adoption of watermarking was hampered in its early days by controversies, e.g., the backlash against protection mechanisms after the U.S. Digital Copyright Millennium Act and European Union (EU) Copyright Directive and bullish marketing that oversold watermarking as a silver bullet against piracy. Nevertheless, forensic watermarking is widely deployed in digital cinemas, postproduction houses, screener systems, and direct-to-consumer video distribution platforms to trace the source of piracy. This commercial success was further recognized in 2016 with a Technology and Engineering Emmy Award for "Steganographic Technologies for Audio/Video for Engineering Creativity in the Entertainment Industry." Digital watermarking is also routinely used to perform audience measurement for radio and TV by companies, such as Nielsen, in North America, and Médiamétrie, in Europe. Besides this core market, the scope of watermarking applications has now expanded beyond content protection, e.g., to convey metadata and media content reliably. For instance, Digimarc is pushing watermarking to replace barcodes in retail stores to speed up checkout time and is currently exploring whether watermarking could be used for plastic packaging to facilitate waste recycling.

Robust hashing

The fundamental requirement of digital watermarking is that multimedia content needs to be modified prior to its delivery to the recipient. In other words, this technology degrades, to some extent, the content, which can appear odd in view of its goal to protect the asset. A parallel line of research in the IFS community, coined *robust hashing* (also known as *perceptual hashing* and *content fingerprinting*), stemmed from this contradiction.

Robust hashing constructs a binary representation of the content in a robust low-dimensional space, aiming at fast and reliable recognition under severe distortions. It is common practice for watermarking techniques to also leverage such low-dimensional spaces to introduce the watermark signal in perceptually significant portions of the signal and thereby achieve robustness. Therefore, it is no surprise that the bodies of work of both research areas share several design patterns, e.g., invariant spaces, pseudorandom projections, differential features, and so on. On another front, there are connections to research on biometrics and content indexing that look for functions that output the same binary value for similar contents. Nevertheless, the IFS community undertook this challenge with a slightly different approach, inspired by cryptographic hash functions. These one-way functions have the property that they produce very different hash values as soon as a single bit of the input changes. They are routinely used to check the integrity of digital assets and to provide efficient inverse lookup mechanisms for large-scale databases. Robust hashing relaxes this "high-sensitivity" property, and the aim is, rather, to tailor one-way functions that yield the same result for perceptually similar pieces of content [10]. The underlying rationale is that a media asset should hash to the same or similar value even after modifications to the content that do not alter its semantics, such as recompression, filtering, resizing, and

more. Thus, robust hashing can be viewed as some quantization scheme in a robust multidimensional space.

A common approach now is to have several such quantizers that produce subhashes, combined with efficient nearest-neighbor search mechanisms. For instance, it has been exemplified that capturing the sign of the first derivative of some robust transform coefficient provides a rather stable hash representation of audio content [11]. Such a mechanism can, then, be exploited to construct song recognition applications, such as Shazam, and to provide means for broadcast monitoring and audience measurement. Robust hashing has also been used to check the integrity of multimedia documents and to identify physical objects, thanks to the hash of their microstructures. Today, research on robust hashing mostly focuses on visual content and is mostly published in content indexing and retrieval venues. A seminal example represents content as a bag of (visual) words [12], each word being a robust hash, and exploits these words to recognize content. For instance, YouTube deploys such techniques to assess whether uploaded user-generated content contains material subject to copyright claims.

Steganography and steganalysis

Steganography is a tool for private covert communication. The sender typically hides a secret message in a host (cover) document by slightly modifying it and then communicating it overtly to the recipient. A steganographic channel is considered secure when an adversary observing the communication cannot detect the fact that steganography is being used. Once the use of steganography is detected, it is considered broken.

Steganography complements encryption for situations when even the existence of the private communication must be concealed and thus finds use in oppressive regimes that ban the use of encryption and for military operations. Statistical undetectability is the main factor distinguishing between steganography and watermarking. In contrast to steganography, the presence of watermarks is often advertised, and they usually have to be robust to distortion while carrying a relatively small payload.

Since detection of steganography amounts to detecting slight modifications of the host signal, steganalysis can be categorized as a forensic technique whose goal is to establish whether the host signal has been modified in a way that is indicative of embedding a secret. Consequently, many techniques developed for steganalysis found applications in forensics and vice versa.

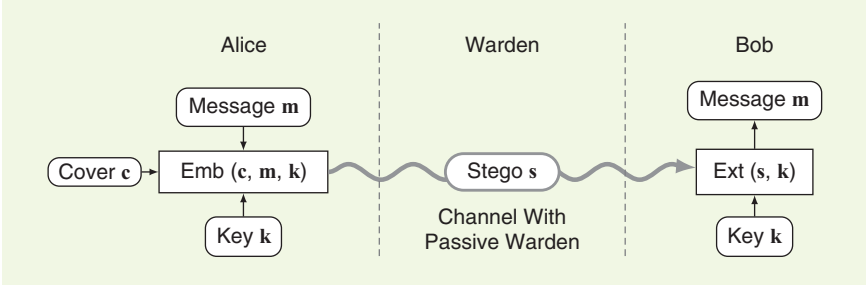

FIGURE 1. The components of the steganographic channel.

Steganographic security

Steganography, in its modern form, is conceptually nested within the fields of information theory, statistical hypothesis testing, and coding. The senders, usually named Alice and Bob, communicate by exchanging objects, which we will assume are digital images. They both agree upon the embedding and extraction algorithms used to embed a secret message in a cover and extract it from the stego-object. Both algorithms make use of a secret shared key, as depicted in Figure 1.

Covers are drawn for communication according to some probability distribution P_c. If Alice uses steganography, her images $\mathbf{s}$ will, in general, follow a different stegodistribution P_s. If Alice is able to make sure that $P_c = P_s$, the stegosystem is considered perfectly secure because it is impossible to distinguish between P_c and P_s by just observing the channel.

Perfect security is achievable only when Alice knows the cover distribution, in which case, she can synthesize images from her secret message by running it through a cover source entropy decoder. Bob reads the message by compressing the image.

For digital media, though, the underlying statistical model is never perfectly known, and thus, all stegoschemes, in practice, are imperfect. A useful measure of steganographic security is the Kullback–Leibler (KL) divergence,

$$D_{\mathrm{KL}}(P_c \| P_s) = \sum_{\mathbf{x}} P_c(\mathbf{x}) \log \frac{P_c(\mathbf{x})}{P_s(\mathbf{x})} \tag{3}$$

because it bounds the performance of the best detector that the warden can build. For imperfect steganography, $D_{\mathrm{KL}}(P_c \| P_s) > 0$, and with n images being sent, Alice needs to scale her payload to be proportional to $\sqrt{n}$ to avoid being caught, with near certainty, by the warden. This asymptotic result is known as the *square root law of imperfect steganography* [13].

Practical steganography

There are two main image formats currently in use: raster formats, such as bitmap, portable network graphics, and TIFF, and the popular JPEG format. Steganographic methods for JPEG files modify the quantized discrete cosine transform coefficients to embed the secret message.

Modern embedding methods are adaptive to content: they take into account the detectability of embedding changes in different parts of the cover image. Intuitively, changes made to a blue sky or water, out-of-focus parts of an image, and overexposed pixels will be more detectable than in textured areas, such as sand, grass, and trees. Alice can guide the embedding by assigning "costs" of changing each cover element and then requesting that the expected total cost of embedding (distortion) be minimal. Alternatively, she can adopt a statistical model and embed while minimizing statistical detectability, usually simplified to the point that it can be expressed using the deflection coefficient.

Today, virtually all steganographic techniques for digital images use some form of the preceding two paradigms. The actual embedding is implemented using linear codes, with the message $\mathbf{m}$ being communicated as the syndrome of the stego-image represented with bits $\mathbf{y} = \mathrm{mod}(\mathbf{s}, 2)$: $\mathbf{m} = \mathrm{Ext}(\mathbf{s}) = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is the code parity check matrix. The parity check matrices of the so-called syndrome trellis codes [14] offer a clever blend of randomness for optimality in terms of the payload distortion (detectability) tradeoff, with enough structure to allow computationally efficient implementation using the Viterbi algorithm.

Steganalysis

Steganalysis detectors can be built using the tools of detection theory as a form of the likelihood ratio test, and they can be data-driven constructed using machine learning. The former usually imposes a statistical model on signals extracted from the image (typically, noise residuals), while data-driven detectors are built by representing images using "features" hand designed to be sensitive to embedding changes but insensitive to image content.

A popular general methodology for designing feature representations for steganalysis is based on the concept of a rich model consisting of a large number of diverse submodels [15]. Rich media models can be viewed as a step toward automatizing steganalysis to facilitate fast development of accurate detectors of emerging steganographic schemes, instead of having to develop a new approach for each new embedding method.

The latest generation of detectors is built in a purely data-driven fashion by presenting a deep convolutional neural network (CNN) with examples of cover and stego images. This constitutes yet another paradigm shift in the field of steganalysis that leads to significant improvements in the detection accuracy of just about every embedding scheme in both the spatial and JPEG domains (see an example of HILL algorithm detection in Figure 2).

The most recent trend in steganalysis (and in forensics in general) with deep learning (DL) is to use CNNs pretrained on computer vision tasks as a good starting point and apply the techniques of transfer learning to refine them for steganalysis. For steganalysis, though, one needs to make sure that the feature map resolution is not decreased via pooling and strides too early in the network architecture, as this form of averaging suppresses the signal of interest, the noise-like stegosignal, while reinforcing the content, which is really the "corrupting noise" for the steganalyst [16].

DL has also advanced the field of steganography in the form of adversarial embedding and fully automatized data-driven learning of embedding costs Backpack [17].

Biometrics

In the past couple of decades, biometric technologies have become increasingly pervasive in our everyday life, due to several inherent advantages they offer over conventional recognition methods, which are based on what a person knows, e.g., passwords and PINs, and what a person has, e.g., ID cards and tokens. However, using biometric data raises many security issues specific to biometric-based recognition systems and not affecting conventional approaches for automatic people recognition.

Biometrics, such as voice, face, fingerprint, and iris, to cite a few, are exposed traits. Therefore, not being secret, they can be covertly acquired and stolen by an attacker and eventually misused, leading, for example, to identity theft. Moreover, contrary to passwords and PINs, raw biometrics cannot be revoked, canceled, and reissued if compromised since they are the user's intrinsic traits and are limited in number.

The use of biometrics also poses many privacy concerns. Biometric data can be used for purposes different than what was intended in the first instance of collection and for what an individual has agreed to. Moreover, when an individual gives out biometric characteristics, information about that person, such as ethnicity, gender, and health conditions, can be potentially disclosed. Some biometrics can be covertly acquired at a distance and, therefore, could be used for surveillance. In addition, using biometrics as a universal identifier across different applications can allow person tracking, thus potentially leading to profiling and social control. To some extent, the loss of anonymity can be directly perceived by users as a loss of autonomy.

In recent years, the need to develop secure and privacy-compliant biometric systems has stimulated industrial and academic research [18] and standardization activities [19]. A biometric system is the interconnection of data capture, signal processing, comparison, decision, and data storage subsystems. Threats against a biometric system are diverse and can be unleashed against its different components, including transmission channels. In Figure 3, major intentional attacks are synthetically illustrated.

Among the depicted attacks, those against templates, hindering both the security and privacy of biometric systems, and the spoofing attack, also known as a *presentation attack*, at the sensor level, have become mainstream in biometric-related research.

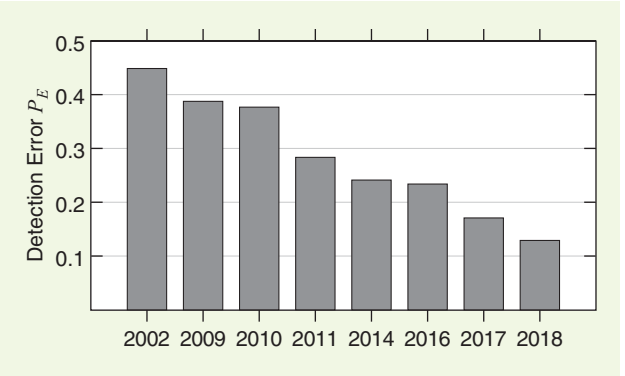

FIGURE 2. The detection error of the HILL stegoalgorithm at 0.4 bits/pixel. The improvements during 2011–2016 were due to the use of rich models, while CNNs were responsible for the advancements during 2017–2018.

Security and privacy requirements of a biometric system

The need to protect biometric templates has emerged as a very stringent requirement for the deployment of secure and privacy-sympathetic biometric systems. However, classical encryption techniques cannot be effectively employed, essentially because of the noisy nature of biometric data that does not allow making comparisons directly in the encrypted domain. To answer the need for secure and privacy-compliant biometric systems, several approaches that treat security and privacy requirements as two sides of the same coin, aiming at enhancing security and minimizing privacy invasiveness, have been designed in the recent past, also triggering significant standardization activity.

The ISO/International Electrotechnical Commission (IEC) 24745 standard [19] makes a clear distinction between the identity reference and the biometric reference (BR), where the first refers to nonbiometric attributes, such as names, addresses, and so forth, uniquely identifying a user, and the second to biometric-related attributes of the individual.

From an ideal standpoint, a secure and privacy-compliant biometric system needs to possess the following properties:

- *Confidentiality*: From a security point of view, the BR can be made available to authorized entities, which have full control of the data, and is protected from nonauthorized ones. From a privacy perspective, the BR can be accessed only by entities that need to access the data and for the purpose for which they were initially collected. The user has full control of the data, with the right to be forgotten.
- *Integrity*: The integrity of the biometric sample of the BR, and the whole authentication process, need to be ensured.
- *Revocability and renewability*: If a BR is compromised, severe security and privacy issues can arise since an attacker can get unauthorized access and the BR as well as other personal data can be revealed. Therefore, it should be possible to revoke a compromised BR and issue a new one based on the same biometric sample.
- *Irreversibility*: To prevent biometric data from being used for purposes different than the originally intended and agreed ones, the BR needs to be irreversibly transformed before being stored.
- *Unlinkability*: BRs should not be linkable, and it should not be possible to infer that they originated from the same biometric data, in a computationally feasible way, for different applications and datasets.

In addition, from an ideal point of view, the performance of a biometric system compliant with the requirements in the preceding should not degrade.

Biometric template protection

According to ISO/IEC 24745, a renewable BR consists of two elements, the pseudonymous identifier (PI) and the auxiliary data (AD), which are generated during the enrollment phase and kept separated either logically or physically.

Biometric template protection schemes can be broadly classified into two different categories, specifically, transformation-based approaches and biometric cryptosystems. It is worth mentioning that hybrid methods combining these two have also been investigated.

Transformation-based approaches rely on using a function whose parameters represent the AD, either invertible or

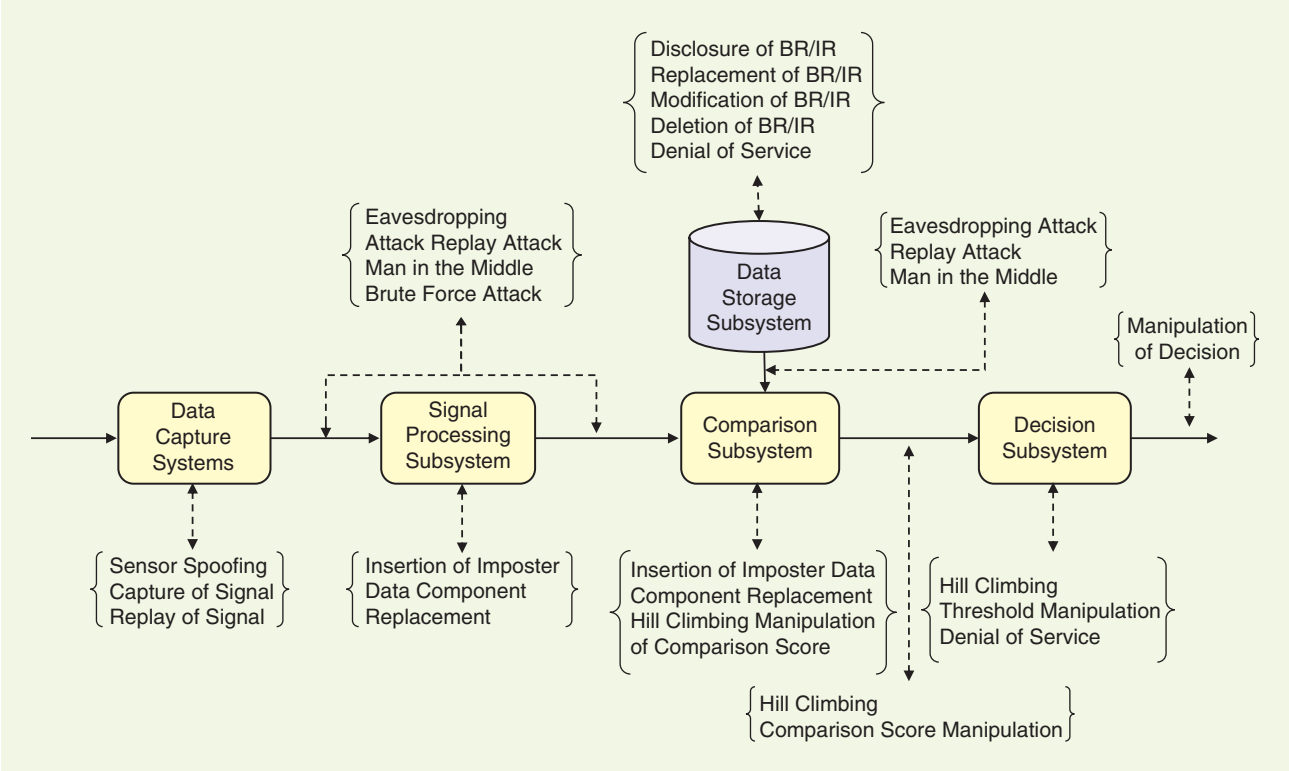

FIGURE 3. The threats to a biometric system. BR: biometric reference; IR: identity reference.

noninvertible, used to transform the BR. The original template is discarded, whereas the transformed one is stored as PI. The match is then performed in the transformed domain. The employed function can be invertible, in which case, a key is needed, and the security of the approach depends on key management. Alternatively, systems relying on the use of noninvertible transformations are more secure, but the design of a noninvertible transformation that does not degrade the recognition performance accuracy is challenging. Several transformations have been proposed and been applied to a variety of biometric identifiers [20], [21].

Biometric cryptosystems are based on adapting cryptographic techniques to the intrinsic noisy nature of biometric data, usually employing error correction coding approaches to protect a BR. Roughly speaking, biometric cryptosystems can be classified into key generation, where a key is obtained from the BR, and key binding schemes, where the BR is bound to a key. In both cases, a secret related to the PI is bound to the BR to generate some public data, namely, the AD, that ideally does not leak information about the BR and that, in conjunction with the BR, can allow retrieving the secret. Within the key binding scenario, the introduction of two constructions, the fuzzy commitment for ordered data and the fuzzy vault for unordered data, have significantly stimulated research in the field and been applied to several biometric identifiers [22]. In [23], an error-tolerant cryptographic primitive, namely, the secure sketch, has been introduced, and in [24], a general framework for analyzing the security of a secure sketch with application to face biometrics is carried out.

Approaches based on secure multiparty computation, employing homomorphic encryption and garbled circuits, have also been proposed for secure and privacy-compliant biometric systems [25]. These architectures rely on the computation of the distance between the stored BR and the probe biometrics in the encrypted domain. The privacy and security of these approaches depend on the computational effort to discover a decryption key by an adversary. However, these methods have shown that they are not mature enough to be deployed in real-life applications, such as those requiring fast identification response, since the involved computational complexity, communication burden, and computational time are significantly higher than other architectures.

Several information-theoretic studies have investigated potential AD information leakage in key binding approaches, that is, the amount of information leaked by the AD about the BR. In [26], the fundamental tradeoff between the secret key and privacy leakage rates in biometric systems is studied for different scenarios. In [27], the findings of [26] are expanded by further discussing the tradeoff among security, privacy, and identification performance. It has been pointed out that as higher identification rates are achieved, more information leakage must be tolerated, and the smallest secret keys can be generated. In addition, the need to provide a quantitative evaluation of unlinkability has been addressed in ISO/IEC 30136 [28], where metrics for security and privacy protection performance assessment have been given.

Presentation attacks

Due to the integration of biometric sensors in almost every smart device and their use in several applications, presentation attacks [29], defined as the presentation of previously stolen human characteristics or fake ones to the acquisition sensor of a biometric system to gain unauthorized access, are receiving increasing interest. Several approaches have been proposed, mainly for fingerprint, iris, and face biometrics. The advent of DL has further fed this line of research, with the development of deepfakes (the "Advent of Deepfakes" section).

Multimedia forensics

The analysis of multimedia evidence has been an essential part of digital forensics since the 1980s. However, only in the late 1990s, with the proliferation of personal digital devices, did it became a full-fledged research field known as *multimedia forensics*, with a focus on source identification (for example, establishing which camera took a given photo) and authenticity verification (for example, detecting the presence and position of manipulated areas in an image). These areas have radically evolved in the past 25 years, following the equally fast evolution of key enabling technologies, such as the hardware and software of imaging devices and new methods for synthetic data generation, and pushed by the massive increase in the volume of audiovisual communications over the Internet and social networks.

The progression of IFS research from digital watermarking to multimedia forensics revolves around the use of "extrinsic" versus "intrinsic" features [30], [31]. These terms were first coined by an interdisciplinary team at Purdue University for electrophotographic printers [32]. The team examined the banding artifacts of printers and treated them as an "intrinsic" signature of a printer that can be identified by appropriate image analysis techniques; it also developed a way to manipulate the banding artifact to embed additional information as an "extrinsic" signature to encode side information, such as the date and time that a document was printed.

From statistical to data-driven approaches

Around the turn of the millennium, the most popular approaches for source identification and forgery detection relied on mathematical and statistical models. A breakthrough in the field was the emergence of methods based on the concept of the device fingerprint, following the seminal 2006 work [33] on the camera photo response nonuniformity (PRNU) pattern. The PRNU is a deterministic sensor pattern, due to tiny imperfections in sensor manufacturing, and can be regarded as a sort of device fingerprint. PRNU-based methods significantly advanced the state of the art in both source identification and image forgery detection and have been extensively used by law enforcement agencies to analyze both physical devices and web accounts. Many other pieces of information have been exploited for forensic investigation, arising from all phases of the digital life of multimedia assets: traces of in-camera operations, such as chromatic aberrations, color filter array artifacts, and double JPEG compression, but also clues related to

out-camera image processing steps, such as image smearing, shadows, and reflections [34].

An inherent limit of model-based approaches is that they mostly fail when the hypotheses do not hold. On the other hand, this is rather the norm in real-world uncontrolled scenarios where data go through unpredictable post-processing operations, as happens on social networks. Another major issue is technological advances. For example, the introduction of computational photography has strongly changed the data acquisition pipeline, and many hypotheses of model-based methods do not hold anymore. Data-driven machine learning methods can partially solve these problems, and, in fact, they have been successfully applied, starting from the first experiments in 2005 [31]. Several tools also took inspiration from work carried out in steganalysis since, despite the obvious differences, both research areas focus on seemingly invisible alterations of the natural characteristics of an image. For example, in the 2013 IEEE Image Forensics Challenge, the winning solutions relied on the rich models [15] developed with great success in steganalysis.

Advent of deepfakes

DL has brought a revolution in multimedia forensics, making available a wealth of simple and powerful tools that allow one to create synthetic content easily. The first deepfake video dates back only to autumn 2017. However, the wide spread of tools based on autoencoders, generative adversarial networks, and, more recently, diffusion models has led to the exponential growth of deepfakes we observe today, which threatens so many areas of our society, from politics to journalism to the private lives of citizens. In particular, artificial intelligence (AI)-powered tools that allow one to generate realistic faces has raised great alarm not only for the diffusion of misinformation but also for the vulnerability of biometric systems.

However, DL also heavily impacted the defense side, making new powerful tools and methodologies available to the forensic analyst. Especially important was the creation of larger and larger datasets of manipulated media, which allowed researchers to train and fine-tune deep NNs [35]. DL-based detectors, trained and tested on such datasets, soon outperformed methods that relied on handcrafted features. In particular, in the most challenging situations of low-quality compressed videos, there is a large gap between a solution based on extracting forensic features and a fully data-driven method based on a very deep CNN [36]. In attribution tasks, DL allows one to learn a camera fingerprint from the available data, gaining independence from a fixed mathematical model and proving more effective in many different situations [37]. The concept of the fingerprint was also extended to synthetic images with so-called artificial fingerprints, related to new types of artifacts introduced in the generation process [35].

Despite their obvious potential, AI-based methods also have well-known weaknesses, from a general lack of interpretability to a limited generalization ability, with poor performance on data generated by manipulation methods and synthetic sources never seen in the training phase. A further major challenge is represented by adversarial attacks, which can easily fool DL detectors. This happens especially when the detector relies on low-level features that can be easily removed by injecting suitable adversarial noise. For this reason, a recent trend is toward the exploitation of semantic artifacts, which are more robust to different signal degradations, such as the biometrics of a specific identity and the geographic information estimated from an image or video.

From single- to multimodal analysis

Another major evolution is represented by a paradigm shift from data processed in isolation to multimodal analyses. With the progress of technology, devices tend to lack unique features that allow easy identification, and manipulations become increasingly sophisticated, evading the scrutiny of expert users. In this context, working on a single media modality may be insufficient, while a joint analysis of all pieces of information associated with a media asset may become key to successful forensics. Accordingly, current methods look for inconsistencies among multiple modalities, such as audio–video and text–image, the latter being especially relevant when unaltered images are used in a new but false context. This trend started in the early 2010s, with the introduction of multimedia phylogeny [38], which aims at investigating the history and evolutionary process of digital objects by automatically identifying the structure of relationships underlying the data. This has led to synergy among different research fields: signal processing, computer vision, and natural language processing. In parallel, the attention to multimedia forensics has moved from forensics labs and law enforcement agencies, as it was 25 years ago, to big tech companies, such as Facebook and Google, and large international research programs. Likewise, research papers once published mostly in specialized forensics venues now find a much wider audience, including major computer vision conferences and satellite workshops.

Adversarial signal processing and machine learning

If there is one thing that researchers trained in the watermarking field had learned by the end of the first decade of the new millennium, it is that security is not robustness [39]. Dealing with random noise and benevolent manipulations is not like dealing with an enemy whose explicit goal is to make the system fail. In the meantime, other security-oriented signal processing applications were emerging, including multimedia forensics [40], biometric security [41], network intrusion detection [42], spam filtering, anomaly detection, and many others. Despite their differences, all these fields were characterized by a unifying feature: the possible presence of one or more adversaries aiming at making the system fail. Prompted by this basic observation, multimedia security researchers started studying the adversarial dynamics describing the interplay between the actions of the system designer and the adversary. Some early works in this area include [43] and [44], where game theory was used to predict behavioral dynamics in traitor

tracing and media sharing applications. As a result of these activities, a broad research area, often referred to as *adversarial signal processing* [45], emerged, whose final goal is to design signal processing tools that retain their effectiveness even in the presence of an adversary.

The peculiar feature of adversarial signal processing is the presence of an informed and intelligent attacker who does not act stochastically since he/she introduces a disturbance optimized to cause the maximum damage to the system. To do so, the attacker exploits the knowledge he/she has about the to-be-attacked system. As a leading example (and without pretending to be exhaustive), let us consider a system responsible for making a binary decision. The binary decision may regard the presence of a watermark within a signal, detecting anomalous behavior, and verifying a biometric trait. Let us assume, for simplicity, that a linear function is used. More specifically, let

$$\phi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle = \sum_{i=i}^{n} x(i) w(i) \tag{4}$$

be a linear combination of the input signal $\mathbf{x}$ and a proper vector of weights $\mathbf{w}$. The system makes a positive decision if $\phi(\mathbf{x}) > T$ and a negative one otherwise. In digital watermarking, for instance, $\mathbf{x}$ corresponds to the observed signal and $\mathbf{w}$ to the watermarking key. As another example, (4) may model an anomaly detector based on Fisher discriminant analysis. In this case, $\mathbf{x}$ contains the features the detector relies on and $\mathbf{w}$ the weights of the linear combination. In a white-box attack, the attacker knows exactly the form of ϕ, including the vector $\mathbf{w}$. In this case, the optimum attack corresponds to adding a perturbation $\mathbf{z}$, defined as follows:

$$\mathbf{z} = (T - \langle \mathbf{x}, \mathbf{w} \rangle - \varepsilon) \frac{\mathbf{e}_w}{\|\mathbf{w}\|} \tag{5}$$

where $\mathbf{e}_w$ is a vector having the direction of $\mathbf{w}$ and ε is an arbitrarily small positive quantity and where we have assumed that the goal of the attacker is to turn a positive decision into a negative one. If the attacker does not know $\mathbf{w}$, attacking the system would require the addition of random noise, with two negative consequences (for the attacker): 1) uncertainty about the result of the attack and 2) the necessity of introducing into the system a larger distortion (on average).

Alternatively, the attacker may not know $\mathbf{w}$ but may have access to the values assumed by ϕ in correspondence with properly chosen inputs. In this case, the attacker may estimate the gradient of ϕ and add a perturbation aligned to the negative direction of the gradient. If ϕ is nonlinear, he/she can use gradient descent to exit the positive decision region with a minimal distortion. In other cases (black-box attacks), the attacker can observe only the final decision of the system in correspondence to chosen inputs. In this case, he/she can apply a so-called sensitivity attack. Let us assume again that ϕ is a linear function. The attacker first chooses a random input resulting in a negative decision. Then, he/she applies a bisection algorithm to find a point on the boundary of the decision region. Finally, he/she repeats the procedure n times, finding n points on the boundary of the decision region. Due to the linearity of ϕ, such n points are enough to estimate $\mathbf{w}$ and compute $\mathbf{z}$ as in (5). If the detection boundary is not linear, then the attack is more difficult yet still possible, as shown in [7].

Let us now turn our attention to the defender. First, the defender may want to keep part of the system secret. By following a common practice in cryptography, secrecy should be incorporated within a key, while the overall form of the system is assumed to be known. In the simple linear case outlined before, this means that the attacker knows the form of ϕ but does not know $\mathbf{w}$. If the defender knows the attacker's strategy, he/she can adopt other countermeasures. For instance, he/she may try to limit the information the attacker can infer by observing the system's output by randomizing the decision function. In systems based on machine learning, the defender may retrain the system by incorporating some attacked inputs in the training set. In this way, the system learns to recognize the attacked inputs and treat them properly.

A problem with most of the defenses described so far is that they assume a static situation, where the attacker adopts a fixed strategy ignoring the possible countermeasures adopted by the defender (it goes without saying that a similar drawback applies to most attack strategies.) When this is not the case, the defender can adopt a worst-case solution, assuming that the attacker has perfect knowledge of the attacked system. However, this tends to be an overly pessimistic approach, given that, in some cases, it may be difficult and even impossible for the attacker to obtain perfect knowledge of the system. Furthermore, the defenses put forward under the worst-case assumption may be too complicated, leading to a significant deterioration of the system's performance. An elegant solution to solve this apparent deadlock and avoid a situation in which new attacks and defenses are developed iteratively in a never-ending loop consists of modeling the interplay between the attacker and the defender by using game theory. Game theory, in fact, provides a powerful way to model the interplay between the attacker and the defender, whose contrast can be defined by the payoff of a zero-sum game, while the constraints they are subject to and the knowledge they have can be modeled by a proper definition of the set of moves they can choose from. Furthermore, it is possible to model both scenarios wherein the attacker and the defender design their systems independently and situations where one of the players acts first and the other adapts his/her move based on the choice made by the first player. Eventually, by computing the payoff at the equilibrium, the achievable performance of the system when both players adopt an optimum strategy can be evaluated. Some examples of works where game theory was successfully used to derive optimal strategies for the attacker and the defender include [42], [43], [46], and [47].

Adversarial AI

DL and AI are revolutionizing the way signals and data are processed. In addition to new opportunities, DL raises new security concerns. When Szegedy et al. [48] pointed out the existence of adversarial examples affecting virtually any

deep NN model, the AI community realized that robustness is not security and that proper countermeasures had to be taken if AI were to be used within an adversarial setting. Such concerns gave birth to a new discipline, usually referred to as *adversarial AI* (or *adversarial machine learning*). Adversarial machine learning has many similarities with adversarial signal processing. When targeting a binary classification network, for instance, adversarial attacks are nothing but a re-edition of the white-box attacks described in (5). More generally, by assuming that the to-be-attacked input is close to the decision boundary and that the boundary is locally flat, in its simplest form, an adversarial example can be computed as

$$\mathbf{x}_{\text{adv}} = \mathbf{x} - \varepsilon \mathbf{e}_{\nabla\phi(\mathbf{x})} \tag{6}$$

where we have assumed that the goal of the attack is to change the sign of the network output, $\mathbf{e}_{\nabla\phi(\mathbf{x})}$ is a vector indicating the direction of the gradient of the output of the network in correspondence of the input $\mathbf{x}$, and ε is a (usually small) quantity ensuring that the network decision is inverted. Even if more sophisticated ways of constructing adversarial examples have been proposed, the similarity to adversarial signal processing is evident, the main peculiarity of attacks against DL architecture being the ease with which $\nabla\phi(\mathbf{x})$ can be computed by relying on back propagation. A distinguishing feature of adversarial machine learning is the possibility of attacking the system during the learning phase. Backdoor attacks are a typical example, where the attacker manipulates the training set to inject into the network a malevolent behavior [49]. Interestingly, this attack presents several similarities with watermarking. Backdoor injection, in fact, can be seen as a way to watermark an NN to protect the intellectual property rights of DL models [50].

Given the striking similarities between adversarial machine learning and security-oriented applications of signal processing, the signal processing community is in an ideal position to contribute to the emerging area of adversarial AI, transferring to this domain the theoretical and practical knowledge developed in the past 25 years.

Additional topics

As mentioned in the introduction, we opted for a focused review of the IFS research areas that experienced the major attention and breakthroughs and that were aligned with the typical technical background of the readership of *IEEE Signal Processing Magazine*. Nevertheless, the scope of IFS is much wider, and this section is intended to provide a glance at other relevant subareas. The most interested readers are invited to check out our flagship journal publication, *T-IFS*, to grasp a comprehensive view of the IFS domain, including areas closer to computer science, information theory, and digital communications.

Due to its security-oriented application domain, a sizable portion of the IFS research has been related to applied cryptography for (multimedia) signals. While the IFS community has not been so much involved with advances for digital rights management and conditional access systems, it has explored alternate encryption mechanisms that would be better suited for multimedia signals compared to bulk encryption. These techniques, routinely coined *selective encryption* [51], consist of encrypting only a small portion of the signal representation to incur unrecoverable degradation while preserving lossy compression capabilities. Unfortunately, the parsing cost of these techniques hampered their adoption. Surviving incarnations of this paradigm today include pattern based, most notably used in Apple's Sample Advanced Encryption Standard and MPEG Common Encryption [52], and encryption limited to a region of interest.

On another front, signal processing in the encrypted domain received increasing attention in the late 2000s. The necessity of processing encrypted signals without first decrypting them arises naturally whenever two or more parties need to cooperate to reach a common goal, without revealing proprietary signals (and data) of a private nature, such as, for instance, medical records [53]. The cryptography community provided baseline tools, namely, homomorphic encryption and multiparty computations, to process encrypted data. Nevertheless, when the data to be processed take the form of signals, it is necessary to exploit synergies between cryptographic tools and signal processing techniques to obtain secure and efficient solutions. Over the years, the IFS community has greatly contributed to developing such solutions for a wide variety of applications domains, including biomedical signal processing, biometrics, smart metering, private recommender systems, and many others.

While forensic techniques for multimedia documents were discussed in the "Multimedia Forensics" section, the IFS community has also explored how to apply similar methodology for other types of signals. For instance, human actions may lead to changes in the surrounding electromagnetic field associated with Wi-Fi systems. These changes can be analyzed to detect a variety of movements, e.g., 1) walking in a building and entering a room, (2) subtle breathing movements, and 3) unexpected sudden movement, such as falling down [54]. The power grid, whose nominal frequency varies over time in a unique manner, also provides a ubiquitous form of ambient signatures. Harnessing the time–frequency properties of grid signatures, which may be revealed in subtle but detectable ways from audiovisual recordings, can enable forensic analysis to determine the capturing time, location, and integrity of these recordings [55].

A final line of research worth mentioning is the so-called security of noisy data [56]. It relates to the ability of reliably recognizing data and signals when their representation slightly differs from one observation to the other. This fundamental problem underpins several IFS subareas, such as biometrics, robust hashing, and sensor-based forensics. Interestingly, the baseline tools developed to tackle this challenge can be revisited to tailor some kind of physically unclonable features. For instance, some IFS contributions have demonstrated that it is feasible to extract some signature from the microscopic

structures of paper and other physical objects to facilitate brand protection and the fight against counterfeiting [57].

Conclusions

All in all, within a quarter century, the IFS research community has widened its focus well beyond its initial goal that revolved, to a large extent, around intellectual property protection. It fully embraced the transition into our new digital world and addresses fundamental societal challenges related to trust, privacy, and protection. While such topics are routinely viewed as owned by computer science, the unique contributions of the IFS community clearly established that signal processing has its own role to play. For instance, the great strides in machine learning are anticipated to raise challenges with the emergence and popularization of generative methods capable of substantiating synthetic realities, such as deepfakes and artificially generated images, text, and videos. This blur of the frontier between natural and synthetic signals is happening right now and will undeniably become an exciting playground for the IFS research community.

While the changes coming with machine learning may be scary, they are also likely to have their own batch of benefits for the IFS research area.

While the changes coming with machine learning may be scary, they are also likely to have their own batch of benefits for the IFS research area, which typically aims at isolating/detecting low-power signals whose characteristics may not be known beforehand. Similar to other domains, technological advances need to be accompanied, on some occasions, by an evolution of the legal framework that governs our lives to diffuse the risk of unmanaged technical advances, including to rule the use of IFS technologies and avoid their misuse. The road to hell is paved with good intentions, and a surveillance technology to protect the safety of people can be abused to invade privacy. It is of equal importance to educate citizens and raise their awareness of some of the dangers inherent to our new digital world, without scaring them. As Descartes said, human senses can be misleading, and one should not take at face value what he sees, reads, or hears (on the Internet).

Acknowledgement

We acknowledge the support of this research by DARPA, under agreement FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. This work has also received funding from the EU, under the Horizon Europe vera.ai project, grant agreement 101070093, and is supported by a Technical University of Munich Institute for Advanced Study Hans Fischer Senior Fellowship and by the Preserving Media Trustworthiness in Artificial Intelligence project, funded by the Italian Ministry of Education, University, and research within the PRIN 2017 program. Anderson Rocha acknowledges the financial support of the São Paulo Research Foundation through the DéjàVu project. Fernando Pérez-González's research is partially funded by the Spanish Ministry for Science and Innovation and NextGenerationEU/PRTR through project Federated Learning with Model Ownership Protection and Privacy Armoring (Grant MCIN/AEI/10.13039/501100011033) and by Xunta de Galicia and the European Regional Development Fund, under project ED431C 2021/47. This work was supported by the National Science Foundation under Grant 2028119.

Authors

Mauro Barni (barni@dii.unisi.it) received his Ph.D. degree in informatics and telecommunications in October 1995. He is a full professor in the Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy. He has been the editor-in-chief of *IEEE Transactions on Information Forensics and Security*, chair of the IEEE Information Forensics and Security Technical Committee, and a Distinguished Lecturer of the IEEE Signal Processing Society (2013–2014). He was the recipient of the 2016 European Association for Signal Processing (EURASIP) Individual Technical Achievement Award. His research interests include multimedia security, media forensics, and artificial intelligence security. He is a member of EURASIP and a Fellow of IEEE.

Patrizio Campisi (patrizio.campisi@uniroma3.it) received his Ph.D. degree in electrical engineering from the Università degli Studi "Roma Tre," Roma Italy. He is a full professor in the Department of Industrial, Electronics, and Mechanical Engineering, Roma Tre University, 00146 Rome, Italy. He is the vice president of publications for the IEEE Biometrics Council (2023–2024) and a member of the IEEE Signal Processing Society (SPS) Nominations and Appointments Committee (2022–2024). He was the SPS director of student services (2015–2017), general chair of the 2015 IEEE International Workshop on Information Forensics and Security, chair of the IEEE Technical Committee on Information Forensics and Security (2017–2018), and editor-in-chief of *IEEE Transaction on Information Forensics and Security* (2018–2021). His research interests include emerging biometrics and biometric security and privacy. He is a Fellow of IEEE.

Edward J. Delp (ace@ecn.purdue.edu) received his Ph.D. degree from Purdue University. He is the Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and a professor of biomedical engineering at Purdue University, West Lafayette, IN 47907 USA. He was the chair of the SPIE/IST Conference on Security, Steganography, and Watermarking of Multimedia Contents (1998–2008) and the IEEE Technical Committee on Information Forensics and Security (2008–2010). He is the recipient of the 2004 IEEE Signal Processing Society (SPS) Technical Achievement Award and the 2008 SPS Society Award for his work in image and video compression and multimedia security. His research interests include image and video processing, multimedia

security and forensics, and machine learning. He is a Life Fellow of IEEE.

Gwenaël Doërr (gdoerr@synamedia.com) received her Ph.D. degree in signal and image processing from the Université de Nice Sophia-Antipolis, Nice, France (2005). She is a principal software architect at Synamedia, 35700 Rennes, France. He is active in standardization efforts related to watermarking, such as the Ultra HD Forum, DASH Industry Forum, and Consumer Technology Association Web Application Video Ecosystem. He co-organized the first IEEE International Workshop on Information Forensics and Security (IFS), in 2009, and served as the IEEE IFS Technical Committee chair in 2014–2015. His research interests include antipiracy solutions, with a focus on forensic watermarking and piracy monitoring techniques. He is a Senior Member of IEEE.

Jessica Fridrich (fridrich@binghamton.edu) received her Ph.D. degree in system science. She is a distinguished professor of electrical and computer engineering at the State University of New York at Binghamton, Binghamton, NY 13850 USA. She was an associate editor of *IEEE Transactions on Information Forensics and Security* and serves on the program committee of the International Workshop on Information Forensics and Security. Her research interests include steganography, steganalysis, digital watermarking, and digital image forensics. She is a Fellow of IEEE.

Nasir Memon (memon@nyu.edu) received his Ph.D. degree in computer science from the University of Nebraska. He is a professor of computer science and engineering at the Tandon School of Engineering, New York University, New York City, NY 11201 USA. He has been on the editorial boards of several journals and was the editor-in-chief of *IEEE Transactions on Information Security and Forensics*. His research interests include digital forensics, biometrics, data compression, network security, misinformation, and security and human behavior. He is a Fellow of IEEE, the International Association for Pattern Recognition, and the Society of Photo-Optical Instrumentation Engineers.

Fernando Pérez-González (fperez@gts.uvigo.es) received his Ph.D. degree in telecommunications engineering. He is a full professor of telecommunications engineering at the University of Vigo, E-36310 Vigo, Spain. He has served as the editor-in-chief of *EURASIP Journal on Information Security* and a senior area editor of *IEEE Transactions on Information Forensics and Security*. He has received three best paper awards at the IEEE Workshop on Information Forensics and Security. His research interests include multimedia forensics, privacy enhancing techniques, and security and privacy in deep learning. He is a Fellow of IEEE.

Anderson Rocha (arrocha@unicamp.br) received his Ph.D. degree in computer science. He is a full professor of artificial intelligence and digital forensics at the Institute of Computing, University of Campinas, Campinas 13083-852, Brazil, where he is the coordinator of the Artificial Intelligence Lab. A Microsoft and Google Faculty Fellow, he is a former chair of the IEEE Information Forensics and Security Technical Committee (2019–2020) and an affiliated member of the Brazilian Academy of Sciences and the Brazilian Academy of Forensics Sciences. His research interests include artificial intelligence, digital forensics, and reasoning for complex data. He is a Senior Member of IEEE.

Luisa Verdoliva (verdoliv@unina.it) received her Ph.D. degree in Information engineering from the University Federico II of Naples. She is a full professor in the Department of Electrical Engineering and Information Technology, University Federico II of Naples, 80125 Naples, Italy. She is a former chair of the IEEE Information Forensics and Security Technical Committee (2021–2022) and is currently the deputy editor-in-chief of *IEEE Transactions on Information Forensics* and a senior area editor of *IEEE Signal Processing Letters*. She is the recipient of a Google Faculty Research Award for Machine Perception and a Technical University of Munich Institute for Advanced Study Hans Fischer Senior Fellowship. Her research interests include image and video processing, with main contributions in the area of multimedia forensics. She is a Fellow of IEEE.

Min Wu (minwu@umd.edu) received her Ph.D. degree in electrical engineering from Princeton University. She is the Christine Yurie Kim Eminent Professor in Information Technology and associate dean of engineering at the University of Maryland, College Park, College Park, MD 20742 USA. She served as past chair of the IEEE Information Forensics and Security Technical Committee (2012–2013) and editor-in-chief of *IEEE Signal Processing Magazine* (2015–2017) and is currently president-elect of the IEEE Signal Processing Society (2022–2023). Her research interests include information security and forensics, multimedia signal processing, and applications of data science and machine learning in smart health and the Internet of Things. She is a Fellow of IEEE, the American Association for the Advancement of Science, and the U.S. National Academy of Inventors.

References

[1] R. Wolfgang, C. Podilchuk, and E. Delp, "Perceptual watermarks for digital images and video," *Proc. IEEE*, vol. 87, no. 7, pp. 1108–1126, Jul. 1999, doi: 10.1109/5.771067.

[2] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997, doi: 10.1109/83.650120.

[3] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001, doi: 10.1109/18.923725.

[4] S. Gel'fand and M. S. Pinsker, "Coding for channels with random parameters," *Problem Control Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[5] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983, doi: 10.1109/TIT.1983.1056659.

[6] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: Theory and practice," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3976–3987, Oct. 2005, doi: 10.1109/TSP.2005.855418.

[7] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind Newton sensitivity attack," *IEE Proc. Inf. Secur.*, vol. 153, no. 3, pp. 115–125, Sep. 2006.

[8] W. Trappe, M. Wu, Z. J. Wang, and K. J. R. Liu, "Anti-collusion fingerprinting for multimedia," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1069–1087, Apr. 2003, doi: 10.1109/TSP.2003.809378.

[9] G. Tardos, "Optimal probabilistic fingerprint codes," *J. ACM*, vol. 55, no. 2, pp. 1–24, May 2008, doi: 10.1145/1346330.1346335.

[10] A. Swaminathan, Y. Mao, and M. Wu, "Robust and secure image hashing," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 2, pp. 215–230, Jun. 2006, doi: 10.1109/TIFS.2006.873601.

[11] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. 3rd Int. Conf. Music Inf. Retrieval*, Oct. 2002, vol. 2002, pp. 107–115.

[12] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vision*, Oct. 2003, pp. 1470–1477, doi: 10.1109/ICCV.2003.1238663.

[13] A. D. Ker, "The square root law of steganography," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2017, pp. 33–44.

[14] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 920–935, Sep. 2011, doi: 10.1109/TIFS.2011.2134094.

[15] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012, doi: 10.1109/TIFS.2012.2190402.

[16] Y. Yousfi, J. Butora, E. Khvedchenya, and J. Fridrich, "ImageNet pre-trained CNNs for JPEG steganalysis," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2020, pp. 1–6, doi: 10.1109/WIFS49906.2020.9360897.

[17] S. Bernard, P. Bas, T. Pevný, and J. Klein, "Optimizing additive approximations of non-additive distortion functions," in *Proc. 9th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2021, pp. 105–112, doi: 10.1145/3437880.3460407.

[18] P. Campisi, *Security and Privacy in Biometrics*. London, U.K.: Springer-Verlag, 2013.

[19] *Information Security, Cybersecurity and Privacy Protection — Biometric Information Protection*, ISO/IEC 24745:2022, International Organization for Standardization, Geneva, Switzerland, Feb. 2022.

[20] E. Maiorana, P. Campisi, J. Fierrez, J. Ortega-Garcia, and A. Neri, "Cancelable templates for sequence-based biometrics with application to on-line signature recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 3, pp. 525–538, May 2010, doi: 10.1109/TSMCA.2010.2041653.

[21] V. M. Patel, N. K. Ratha, and R. Chellappa, "Cancelable biometrics: A review," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 54–65, Sep. 2015, doi: 10.1109/MSP.2015.2434151.

[22] K. Nandakumar and A. K. Jain, "Biometric template protection: Bridging the performance gap between theory and practice," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 88–100, Sep. 2015, doi: 10.1109/MSP.2015.2427849.

[23] Y. Dodis, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," in *Proc. Int. Conf. Theory Appl. Cryptogr. Techn.*, 2004, vol. 3027, pp. 523–540, doi: 10.1007/978-3-540-24676-3_31.

[24] Y. Sutcu, Q. Li, and N. Memon, "Protecting biometric templates with sketch: Theory and practice," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 503–512, Sep. 2007, doi: 10.1109/TIFS.2007.902022.

[25] J. Bringer, H. Chabanne, and A. Patey, "Privacy-preserving biometric identification using secure multiparty computation: An overview and recent trends," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 42–52, Mar. 2013, doi: 10.1109/MSP.2012.2230218.

[26] T. Ignatenko and F. M. J. Willems, "Biometric systems: Privacy and secrecy aspects," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 956–973, Dec. 2009, doi: 10.1109/TIFS.2009.2033228.

[27] T. Ignatenko and F. M. J. Willems, "Fundamental limits for privacy-preserving biometric identification systems that support authentication," *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5583–5594, Oct. 2015, doi: 10.1109/TIT.2015.2458961.

[28] *Information Technology — Performance Testing of Biometric Template Protection Schemes*, ISO/IEC 30136:2018, International Organization for Standardization, Geneva, Switzerland, Mar. 2018.

[29] S. Marcel, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Cham, Switzerland: Springer-Verlag, 2022.

[30] P.-J. Chiang, N. Khanna, A. Mikkilineni, M. O. Segovia, S. Suh, J. Allebach, G. Chiu, and E. Delp, "Printer and scanner forensics," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 72–83, Mar. 2009, doi: 10.1109/MSP.2008.931082.

[31] M. Stamm, M. Wu, and K. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, May 2013, doi: 10.1109/ACCESS.2013.2260814.

[32] G. N. Ali, A. K. Mikkilineni, J. P. Allebach, E. J. Delp, P.-J. Chiang, and G. T. Chiu, "Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices," in *Proc. Int. Conf. Digit. Printing Technol. (IS&T's NIP)*, 2003, vol. 19, no. 2, pp. 511–515.

[33] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 2, pp. 205–214, Jun. 2006, doi: 10.1109/TIFS.2006.873602.

[34] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, Mar. 2009, doi: 10.1109/MSP.2008.931079.

[35] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.

[36] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 1–11.

[37] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 144–159, 2020, doi: 10.1109/TIFS.2019.2916364.

[38] Z. Dias, A. Rocha, and S. Goldenstein, "Image phylogeny by minimal spanning trees," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 774–788, Apr. 2012, doi: 10.1109/TIFS.2011.2169959.

[39] L. Pérez-Freire, P. Comesaña, J. R. Troncoso-Pastoriza, and F. Pérez-González, "Watermarking security: A survey," in *Proc. Trans. Data Hiding Multimedia Secur. I*, 2006, pp. 41–72, doi: 10.1007/11926214_2.

[40] M. Barni, M. C. Stamm, and B. Tondi, "Adversarial multimedia forensics: Overview and challenges ahead," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, 2018, pp. 962–966, doi: 10.23919/EUSIPCO.2018.8553305.

[41] B. Biggio, P. Russu, L. Didaci, and F. Roli, "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 31–41, Sep. 2015, doi: 10.1109/MSP.2015.2426728.

[42] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16–29, Jan. 2009, doi: 10.1109/TSP.2008.2007335.

[43] H. V. Zhao and K. J. R. Liu, "Traitor-within-traitor behavior forensics: Strategy and risk minimization," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 4, pp. 440–456, Dec. 2006, doi: 10.1109/TIFS.2006.885023.

[44] H. V. Zhao, W. S. Lin, and K. J. R. Liu, *Behavior Dynamics in Media-Sharing Social Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[45] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8682–8686, doi: 10.1109/ICASSP.2013.6639361.

[46] P. Schöttle and R. Böhme, "A game-theoretic approach to content-adaptive steganography," in *Proc. Int. Workshop Inf. Hiding*, 2012, pp. 125–141, doi: 10.1007/978-3-642-36373-3_9.

[47] M. Barni and B. Tondi, "Theoretical foundations of adversarial binary detection," *Found. Trends Commun. Inf. Theory*, vol. 18, no. 1, pp. 1–172, Dec. 2020, doi: 10.1561/0100000102.

[48] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014.

[49] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *IEEE Open J. Signal Process.*, vol. 3, pp. 261–287, Jul. 2022, doi: 10.1109/OJSP.2022.3190213.

[50] Y. Li, H. Wang, and M. Barni, "A survey of deep neural network watermarking techniques," *Neurocomputing*, vol. 461, pp. 171–193, Oct. 2021, doi: 10.1016/j.neucom.2021.07.051.

[51] A. Massoudi, F. Lefebvre, C. D. Vleeschouwer, B. Macq, and J.-J. Quisquater, "Overview on selective encryption of image and video: Challenges and perspectives," *EURASIP J. Inf. Secur.*, vol. 2008, Dec. 2008, Art. no. 179290, doi: 10.1155/2008/179290.

[52] *Information Technology – MPEG Systems Technologies – Part 7: Common Encryption in ISO Base Media File Format Files*, ISO/IEC 23001-7:2016, International Organization for Standardization, Geneva, Switzerland, Feb. 2016.

[53] R. L. Lagendijk, Z. Erkin, and M. Barni, "Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 82–105, Jan. 2013, doi: 10.1109/MSP.2012.2219653.

[54] B. Wang, Q. Xu, C. Chen, F. Zhang, and K. J. R. Liu, "The promise of radio analytics: A future paradigm of wireless positioning, tracking, and sensing," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 59–80, May 2018, doi: 10.1109/MSP.2018.2806300.

[55] R. Garg, A. Varna, A. Hajj-Ahmad, and M. Wu, "'Seeing' ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013, doi: 10.1109/TIFS.2013.2272217.

[56] P. Tuyls, B. Škorić, and T. Kevenaar, *Security with Noisy Data*. London, U.K.: Springer-Verlag, 2007.

[57] M. Diephuis, S. Voloshynovskiy, T. Holotyak, N. Stendardo, and B. Keel, "A framework for fast and secure packaging identification on mobile phones," in *Proc. SPIE Media Watermarking, Secur., Forensics*, Feb. 2014, vol. 9028, doi: 10.1117/12.2039638.

SP