University of Groningen

AReN

Greco, Antonio; Petkov, Nicolai; Saggese, Alessia; Vento, Mario

Published in:
IEEE transactions on information forensics and security

DOI:
[10.1109/TIFS.2020.2994740](https://doi.org/10.1109/TIFS.2020.2994740)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Greco, A., Petkov, N., Saggese, A., & Vento, M. (2020). AReN: A Deep Learning Approach for Sound Event Recognition Using a Brain Inspired Representation. *IEEE transactions on information forensics and security*, 15, 3610-3624. [9093814]. <https://doi.org/10.1109/TIFS.2020.2994740>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

AReN: A Deep Learning Approach for Sound Event Recognition Using a Brain Inspired Representation

Antonio Greco^{ID}, Nicolai Petkov^{ID}, Alessia Saggese^{ID}, and Mario Vento^{ID}

Abstract—Audio surveillance is gaining in the last years wide interest. This is due to the large number of situations in which this kind of systems can be used, either alone or combined with video-based algorithms. In this paper we propose a deep learning method to automatically recognize events of interest in the context of audio surveillance (namely screams, broken glasses and gun shots). The audio stream is represented by a gammatonegram image. We propose a 21-layer CNN to which we feed sections of the gammatonegram representation. At the output of this CNN there are units that correspond to the classes. We trained the CNN, called AReN, by taking advantage of a problem-driven data augmentation, which extends the training dataset with gammatonegram images extracted by sounds acquired with different signal to noise ratios. We experimented it with three datasets freely available, namely SESA, MIVIA Audio Events and MIVIA Road Events and we achieved 91.43%, 99.62% and 100% recognition rate, respectively. We compared our method with other state of the art methodologies based both on traditional machine learning methodologies and deep learning. The comparison confirms the effectiveness of the proposed approach, which outperforms the existing methods in terms of recognition rate. We experimentally prove that the proposed network is resilient to the noise, has the capability to significantly reduce the false positive rate and is able to generalize in different scenarios. Furthermore, AReN is able to process 5 audio frames per second on a standard CPU and, consequently, it is suitable for real audio surveillance applications.

Index Terms—audio surveillance, deep learning, CNN, gammatonegram, brain inspired representation.

I. INTRODUCTION

THE size of the global video surveillance market is expected to strongly grow up in the next years. According to a study conducted by the *BIS Research* [1], the global video surveillance market is expected to reach \$62.6 billion by 2023, with a compound annual growth rate (CAGR) of 9.3%. As a matter of fact, audio analysis can be used as a standalone solution or combined with video analytic tools. When used standalone, it can be considered as the solution

for those situations that cannot be reliably detected by only analyzing the video stream [2]; a scream or a gunshot are two examples of events of interest, potentially abnormal, having a very distinctive acoustic signature, and that cannot be easily detected by visual analysis. Vice versa, when combined with video analytic tools, audio analysis may significantly help to understand what is currently happening in a scene, by adding the information provided by video analytics that, among the others, suffers of adverse weather conditions, sudden lighting changes, reflections and so on [3].

Although audio analysis plays a very relevant role in surveillance applications, many challenging tasks must be adequately faced for obtaining results so good to be exploited in real applications. To this concern, in the last years, several scientific contributions have been made available, confirming the wide interest of the scientific community in this field, but the problem is still open [4], at least if we consider application contest in the wild that are characterized by unsatisfactory performance in complex scenarios. Moreover, it's important to consider that, before the beginning of the deep learning era, the effort of the scientific community has been devoted to find out the best feature set and the best classifier for representing and then classifying patterns regarding the problem at hand; so, also in the case of audio event classification, several features sets have been proposed and used, mainly defining them directly on raw data, e.g. the input signal in the time domain, or over its time-frequency representation; examples ranges from temporal based features, as the cepstral ones, to frequency based features, so as from energy-based features to biologically or perceptually driven ones [5]–[10]. A detailed review of the different typologies of features proposed along the years for sound event recognition can be found in a recent survey by Crocco *et al.* [4].

In the last years, the deep learning methods are revealing their effectiveness also in applications of audio analysis; although a taxonomy of the published papers has not yet been definitely settled, we may recognize two main emerging trends: (i) some methods analyze directly the raw audio data, in the time domain, by using Restricted Boltzman Machines (RBM) [11], [12] or, more generally, by exploiting the paradigm of deep belief networks (DBN) [13]. These approaches are somehow related with the use of temporal based features; as a matter of fact, the features are not hand-crafted but instead automatically optimized and obtained directly by the deep

Manuscript received September 11, 2019; revised February 23, 2020 and April 8, 2020; accepted April 30, 2020. Date of publication May 14, 2020; date of current version July 6, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. William R. Schwartz. (Corresponding author: Alessia Saggese.)

Antonio Greco, Alessia Saggese, and Mario Vento are with the Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy (e-mail: agreco@unisa.it; asaggese@unisa.it; mvento@unisa.it).

Nicolai Petkov is with the Faculty of Science and Engineering, University of Groningen, 9712 CP Groningen, The Netherlands (e-mail: n.petkov@rug.nl). Digital Object Identifier 10.1109/TIFS.2020.2994740

networks, during the training procedure; (ii) a second trend refers to papers that use pre-computed sound representations obtained by CNNs, starting from the raw data. A good example is offered by the various time-frequency representations of the input signal, such as the spectrogram or the MFCC spectrogram [14], [15]. Among the most recent and noteworthy contributions in this field, we can cite AENet [16], an example of CNN fed by spectrogram images achieving very promising results for the problem of sound event recognition.

Experimental analyses known in the literature [17], [18], aimed at comparing time-based with frequency-based representation, seem to confirm that the methods that somehow adopt frequency-based representations of the audio perform generally better than the ones based on other representations, including the temporal one. This conclusion, although cannot be considered as a definitive one, encourages, in this historical moment, the adoption of approaches based on suited audio data representations processed by deep architectures, coherently with ii); to this concern, it is worth noting that the spectrogram is experimentally revealing an effective representation of audio signals as for its main property of considering all the frequencies of the input equally relevant [14], [16]; this means that, in the spectrogram, all the frequential components are equally weighted and adequate bandwidth filters are used for obtaining the quantity of energy of the input signal falling within a window of a given fixed size centered on the considered frequency.

As better clarified in the following of the paper, experiences done up to now as for the problem of detection of audio events in the wild, as the ones occurring in smart cities environments, are confirming that the discriminative power of the different classes of audio events mainly concentrate in the low frequencies of the spectrum and progressively decreases as the frequency increases [19]. Coherently with this evidence, it would be better suited a representation able to stretch the low frequencies of the spectrum, compressing at the same time the higher ones, as the gammatonegram. A relevant advantage deriving from the adoption of representations like the spectrogram or the gammatonegram is the possibility of obtaining, as output, an image representing the input signal and this aspect open the way of adopting the wide variety of classification methodologies derived from the computer vision field.

Another important aspect when dealing with audio event recognition is that audio events to be recognized are often superimposed to a significant background noise. This condition is particularly frequent in the audio surveillance applications, especially those “in the wild environments”, where the microphones are typically installed outdoor; in these cases, the signal to noise ratio (SNR) can be in many cases very low, depending on the power of the sound source, not always sufficiently high, and on the great distance between the source and the microphone. Another aspect to be taken into account is the highly variable duration of the sound to be classified; in these cases, the system has to deal at the same time with impulsive events, as gunshots, and with sustained events as screams, being the latter much longer in time. Both the high variability of SNR ratio and the event duration, suggest that the

training of a deep system for sound event recognition should be based on a suited dataset of training events, sufficiently wide to adequately represent all the real situations; this aspect is very relevant and the use of the currently available datasets in the literature, even apparently sufficiently big, demonstrate their inadequate representativeness of the situations “in the wild”, being responsible of significant drops when we pass from the performance on the training set and the one achieved on real data [19].

A way for facing this issue consists in avoiding the problem of training a network from the scratch by using, for example, traditional convolutional neural networks already pre-trained over standard computer vision datasets, such as ImageNet; in this case the available data set, even not so wide, when are insufficient for training a completely new deep network, may instead result sufficient to achieve good results by apply fine tuning [16] to pre-trained networks; in this case the data are used by the system designer for tuning the system, adapting its performance to the specific data at hand.

An alternative is the adoption of method for increasing data representativeness, based, for instance, on data augmentation; techniques widely adopted in the literature ranges from standard image augmentation techniques (random cropping, rotation, random noise and so on) [20] to more advanced techniques applied to the image, aiming at masking the visual representation in specific frequency channels and specific time steps [21]. Of course when images are obtained by a time-frequency sound signal representation, the data augmentation techniques must be specifically devised for this context; in particular the crucial point is the definition of a data augmentation technique effective for taking into account big amounts of noise overlapped to the input signals, so as to simulate real environments with the presence of many different kinds of source of noise, even with very low SNR. Data augmentation can be used for enlarging the representativeness of the input data and should not be considered as a means for augment small datasets; its application in any case allows the system designer to improve data significance, no matter how many data he has in his hands.

At the light of the above considerations, we may draw the main contributions of this paper:

- We adopt a brain inspired representation of the audio signal, namely the gammatonegram [19], simulating the way in which the cochlea membrane in the human auditory system works; the main advantage lies in the fact that this representation has not a constant bandwidth across all frequency windows, but instead wider filter windows for lower frequencies, where the audio events of interest are supposed to lie, so increasing the capability to distinguish the background from the events of interest. This representation allows to adopt deep networks devised for computer vision tasks, inheriting many relevant scientific results on image classification;
- We propose a novel convolutionary neural architecture, called AReN (Audio Event Recognition Network), designed with the main goal to guarantee an high representative capability that can be obtained by using a high number of layers and at the same time to keep

low the total number of its parameters, so avoiding the well-known overfitting problems occurring when the amount of training data is limited. This is done with a network composed by 21 layers (12 convolutional, 4 pooling, 5 fully connected), designed with the following attentions: since the number of the net parameters grows quadratically with the kernel size used in the convolutional layers, AReN adopts only kernels of size 3×3 ; moreover, the fully connected layers are arranged according to a pyramidal structure [22] which substantially reduces the number of net parameters. An enjoyable effect is that the above mentioned pyramidal structure, achieves, even with less parameters, a higher accuracy with respect to the standard structure when a limited amount of training samples is available [23].

- We define a novel problem-driven data augmentation so as to augment the training set with visual representations obtained from noisy versions of the original audio signal; in other words, the noise is not added to the obtained image representation of the sound, but, viceversa, to the original sound, from which we re-obtain the noisy representation, so as to adequately simulate the effect obtained in real environments. We demonstrate the effectiveness of the proposed technique applied on three different datasets.
- The experimental phase has been carried out with the aim to analyze the robustness of the proposed system in real environments; in particular the drop of the performance when moving from “lab conditions” (this is what it is typically done when testing over benchmarking datasets) to live environments; this aspect, only partially investigated in the scientific literature, is crucial for judging the applicability in real scenarios. This is done with respect to a problem of three event classification (broken glasses, screams and gunshot) with a wide range of SNR values, representing challenging real situations.

The proposed architecture has been widely experimented, and the main findings, extensively described in the paper, can be summarized in the following:

- as for the representation issues, the use of the gammatonegram, with respect to spectrogram, revealed to be more effective; it allowed at the same time to increase the recognition rate (recall) and to decrease the false positive rate (precision), with a significant overall performance improvement;
- the proposed system achieves, on standard public datasets, widely used in the literature for audio surveillance in smart cities, performance at the state of the art; the proposed network shows a relevant robustness with respect to environmental noise, thanks to the use of the proposed problem-driven data augmentation. It also reaches a notable generalization capability, being able to achieve the best performance, at the same time, on three different dataset, with a limited drop in the performance when passing in a real-time experiment.
- the system is also effective by the computational point of view; in fact AReN requires 200 ms for the classification on a standard CPU; so, it can be run on traditional hardware, even embedded, and provide five audio clas-

sifications per second, getting a high responsive feeling from the user.

The paper is organised as follows: Section II details the proposed approach and highlights the rationale inspiring our choices; Section III introduces the dataset used in the experimentation and details the experimental protocol and the achieved results, also compared with state of the art methodologies. Conclusions and future works are finally drawn in Section IV.

II. THE PROPOSED METHOD

The length of a sound of interest may vary depending on the type of event; for instance, in the context of audio surveillance we need to recognize both a scream, having a sustained duration, and a gunshot, which is an impulsive sound. In order to deal with both the typologies of sound, we decide to segment the audio stream in consecutive overlapped frames by means of a 3 seconds mobile Hamming window. This choice is supported by a preliminary experimental analysis with different frame size, which showed a decrease of around 10% and almost 20% of the recognition rate with 1 second and 2 seconds windows, respectively.

In order to guarantee that during the operating phase an event is fully included in a single window and thus to avoid border effect due to the presence of events split into two consecutive window, we introduce a 75% overlap between two consecutive windows. This is a particularly important feature especially for short events (such as gunshots), that risk to be too short to be reliably recognized if partitioned into more windows.

Each frame is represented through a time-frequency based representation. The latter has been widely used in the scientific literature, since as already discussed it proved to perform better than the still temporal-based one [17], [18]. The typical approach used in the literature is to adopt a representation based on a Short-Time Fourier Transform (STFT), also known as spectrogram. The spectrogram is a representation of the spectrum of frequencies of a signal (computed by means of the STFT) along the time, depicted as an image.

In this paper we propose to use instead a representation based on Gammatone filterbank; the response of such filterbanks generates an image called gammatonegram [24]. This is a brain inspired representation of the sound emulating the way in which the human auditory system works, and in particular how the *cochlea membrane* responds to incoming sound waves. Indeed, depending on the frequency and the energy of the input sounds, the cochlea membrane has a different vibration, which in turns determines the firing activity of the so called inner-hair cells, posed behind the cochlea, and transfers the electrical stimuli to the auditory nerve. It has been experimentally demonstrated that the gammatone filter is a linear approximation of the impulsive response of the cochlea membrane [25]–[27]. We chose it instead of the full cochleagram [28] for its computational efficiency. Indeed, the gammatonegram requires only a Discrete Fourier Transform (DFT) and a matrix multiplication, since the bins of the DFT are weighted proportionally to the magnitude gain

of a gammatone bandpass filter [29]; it is more efficient than the full cochleagram, which requires the implementation of a parallel filter bank composed by multiple filters with differing central frequencies.

The main difference with respect to the traditionally used spectrogram lies in the fact that in the latter the input signal is processed by a bank of band-pass filters having the same bandwidth; viceversa, in the Gammatone filterbank the bandwidth of the various band-pass filters proportionally increases with their central frequency. It implies that the resolution in the perception of frequency differences is not constant, since the minimum frequency difference that the ear is able to resolve is proportional to the base frequency of the sounds. In other words, a given frequency difference is perceived as much stronger at low frequencies than at high frequencies [19]. This is a particularly important property when dealing with audio surveillance, since the events are typically located at lower frequencies, thus implying that the most important frequencies for audio surveillance (namely the lower) are highlighted and somehow stretched in the gammatonegram image, since represented with more details if compared with the less important frequencies (namely the higher).

In a more formal way, we filter the audio signal with a bank of Gammatone band-pass filters, whose response $g(t, f_c)$ is defined as the product of a gamma function with a sinusoidal tone:

$$g(t, f_c) = at^{n-1}e^{-2\pi bt} \cos(2\pi f_c t + \phi) \quad (1)$$

where t is the time (in seconds), f_c is the central frequency of the filter (in Hz), ϕ is the phase (in radians) which is usually equal to 0.

The constant a is the amplitude and controls the gain, while n is the order of the filter. Finally b is the decay factor, which determines the bandwidth of the filter and the duration of the impulse response. The values of b and a have been inherited by the literature, with the aim to approximate in the best possible way the human auditory system [30].

In more details, we consider a fourth-order filter (namely, $a = 4$), and compute b as follows:

$$b = 1.09 \left(\frac{f_c}{9.26449} + 24.7 \right) \quad (2)$$

The central frequencies of the filters are distributed across the spectrum, in proportion to their bandwidth by following the Equivalent Rectangular Bandwidth (ERB) scale:

$$ERB(f_c) = 24.7 + 0.108f_c \quad (3)$$

Finally, the gammatonegram image is built by concatenating the response of the frame $x(t)$ to the Gammatone filterbank $g(t, f_c)$, where each column of the gammatonegram is the response of the gammatone filterbank at the time instant t :

$$y(t, f_c) = x(t) * g(t, f_c) \quad (4)$$

An example of Gammatone filter, compared with the spectrogram computed over the same frame, is shown in Fig. 1. It is evident how the events of interest, localized in the low

Fig. 1. Representation of a sample for any of the three classes of events of interest and of background using the spectrogram (right) and the gammatonegram (left); the y-axis of the spectrogram is linear, in the range 0-32 KHz; viceversa, the y-axis of the gammatonegram reports frequencies equally spaced between 50 Hz and 10 KHz on the ERB scale. Note how the gammatonegram, by progressively stretching the scale of the lower frequencies and compressing the higher ones, allows to better visually evaluate how the energy of the input signal is spread over its frequential components.

frequencies in the spectrogram images, are spread over the whole gammatonegram image, thus confirming the representational power of the proposed approach.

Given the gammatonegram image (50×200 RGB image) and inspired by the wide literature in computer vision applications, we use a Convolutional Neural Network (CNN) for the classification task. This choice is motivated by the consolidated evidence that the deep learning paradigm has a significant superiority with respect to the classic methods based on handcrafted features. CNNs have been used with success in several audio analysis tasks, such as music classification [31], affective computing [32] and sentence classification [33]. Anyway, as discussed in Section I, it has not yet widely adopted in sound event recognition tasks. We will show that the popular deep networks designed for dealing with very complex image analysis problems are also effective for the problem at hand. Furthermore, according to our knowledge, the proposed network is the first example available in the literature which combines CNNs with gammatonegram based representation.

We call our network *AReN* (Audio Event Recognition Network); its architecture, summarized in Table I, is inspired by VGG-19 and is composed by a set of twelve convolutional layers (instead of 16) with an increasing size of the feature maps (from 64 to 512); each convolutional layer employs a very small receptive field, 3×3 , which is the smallest size to capture the notion of left/right, up/down, center. Four max pooling layers are also used in order to reduce the volume size.

Differently from VGG-19, we introduce a set of fully connected layers organized in a pyramidal way: four different layers whose size ranges from 4096 to 512 (4096, 2048, 1024, 512) are introduced in order to reduce the number of necessary parameters. The network is completed by a last layer whose size is equal to the number of output classes (4 is our context) and a softmax function. The softmax function squashes the outputs of each unit to be between 0 and 1 and also divides each output so that the total sum of the outputs is equal to 1. Given the generic k -th class, with $k = 1, \dots, K$ (K is the number of classes, 4 in our application), the softmax activation function σ works as follows on the input vector z :

$$\sigma(z)_k = \frac{e^{z_k}}{\sum_{k=1}^K e^{z_k}} \quad (5)$$

The main advantage deriving from the proposed deep network lies in the fact that the small kernel size, combined with the pyramidal architecture of the fully connected layers, allows to manage deep networks (like the proposed one, composed by 21 layers) without paying in terms of generalization capabilities, even if the training set is rather small. This is a very important and not negligible feature in the audio surveillance domain, where the datasets available in the literature are not so wide as in video analysis domain.

Each neuron in the hidden layers is modeled by a Rectified Linear Units (ReLU) [34]; thus, given the input x , the neuron's output function $f(x)$ is computed as follows:

$$f(x) = \max(0, x) \quad (6)$$

This choice is justified by the fact that, as experimentally proved in [35], the ReLU shows better convergence performance than sigmoid function and allows to train a network

TABLE I
THE ARCHITECTURE OF AReN: TWELVE CONVOLUTIONAL LAYERS WITH 3×3 KERNELS ARE COMBINED WITH FOUR POOLING LAYERS. FINALLY, A PYRAMIDAL SET OF FIVE FULLY CONNECTED LAYERS IS INTRODUCED

AReN		
(50x200x3) input image		
#Fmap	Kernel size	Layer
64	3x3 kernel	conv(3,64) conv(64,64) pool 1x2
128	3x3 kernel	conv(64,128) conv(128,128) pool 2x2
256	3x3 kernel	conv(128,256) conv(256,256) conv(256,256) pool 2x1
512	3x3 kernel	conv(256,512) conv(512,512) conv(512,512) pool 2x1
FC	-	FC4096 FC2048 FC1024 FC512 FC4
softmax		

several times faster than their equivalents with sigmoid units; as evident, this is a very desirable property when dealing with deep CNNs. Indeed, the ReLU just needs to compute the maximum between two values, without performing expensive exponential operations as in the sigmoid. Furthermore, it is not affected by the problem of the vanishing gradients since the gradient of the ReLU function is always 0 (for $x < 0$) or 1 (for $x > 0$), thus it is possible to add several layers without risking that the multiplication between gradients (layer by layer) vanishes or explodes.

The training of the network has been performed by using a mini-batch gradient descent optimization algorithm based on the adaptive learning rate RMSprop proposed by Ruder [36]. The mini-batch gradient descent takes the best of both worlds and performs an update for every mini-batch of n training examples; in this way the variance of the parameters updates is reduced, so leading to a more stable convergence; moreover, the mini-batch makes the gradient computation very efficient.

In a more formal way, the mini-batch gradient descent optimization algorithm minimizes an objective function $J(\theta)$ parametrized by $\theta \in \mathbb{R}^d$; it works by updating the parameters in the opposite direction of the gradient of the objective function.

The RMSprop divides the learning rate by an exponentially decaying average of squared gradients $E[g^2]$, so that the

running average $E[g^2]$ at time t only depends on the previous average $E[g^2]_{t-1}$ and the current gradient g_t^2 as follows:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + 0.1 g_t^2 \quad (7)$$

The update rule thus becomes:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \quad (8)$$

As suggested by Hinton, we set $\gamma = 0.9$ and the learning rate $\eta = 10^{-6}$.

A data augmentation has also been introduced in order to increase the size of the training set. The traditionally used approaches in computer vision includes flip, rotation, scale, translation and so on. It would be equivalent, in our context, to apply such transformations over the gammatonegram images. However, in this paper we introduce instead a problem-driven data augmentation. We first produce noisy versions of the original sounds of interest, by attenuating or amplifying the energy of the foreground sound according to a specific value of the Signal to Noise Ratio (SNR); then, we generate additional gammatonegram images by using these noisy samples. In other words, the noisy versions of the training data are not built without any knowledge on the problem at hand, but instead are problem-driven.

III. EXPERIMENTAL RESULTS

In order to prove the effectiveness of the proposed approach, we perform an extensive experimental analysis, aiming at evaluating both the efficiency and the effectiveness of our method. The dataset used in our experimentation and the experimental protocol are detailed in Subsection III-A and III-B respectively, while the achieved results, together with a comparison with several state of the art methodologies and a comprehensive discussion, are reported in Subsections III-C and III-D.

A. Dataset

Differently from several video-based applications, there are only a few challenging datasets available in the literature for audio analysis problems. This is particularly true for audio surveillance applications, where the number of available datasets in real environments is very limited. Starting from the above consideration, we decide to use a freely available dataset, widely adopted by the scientific community working in this field, namely the Mivia Audio Events Dataset [37]: about 30 hours of audio recording including four classes of interest, namely glass breaking (GB), gunshots (G) and screams (S). Furthermore, the dataset also includes background (B) samples, corresponding to those sounds acquired both from indoor and outdoor environments, such as silence and Gaussian noise, rain, whistles, crowded ambiance, vehicles, household appliances, bells, applauses and claps. The composition of the dataset is detailed in Table II.

The dataset has been recorded with an Axis P8221 Audio Module and an Axis T83 omnidirectional microphone for audio surveillance applications. It is composed by WAV clips, already partitioned into training (about 70%) and testing (about

TABLE II

COMPOSITION OF THE MIVIA AUDIO EVENTS DATASET: NUMBER OF EVENTS AND TOTAL DURATION OF BACKGROUND (B), GLASS BREAKING (GB), GUNSHOTS (G) AND SCREAMS (S)

Type	Training set		Test set	
	Events	Duration (s)	Events	Duration (s)
B	-	58,372	-	25,037
GB	4,200	6,025	1,800	2,562
G	4,200	1,884	1,800	744
S	4,200	5,489	1,800	2,445

30%). The clips have been PCM sampled at 32 kHz and quantized at 16 bits per sample.

It is important to note that in the era of deep learning we are used to datasets with thousands of classes, accompanied by millions of image data. Viceversa, in audio surveillance applications there are not so big datasets. Anyway, although the MIVIA Audio Event Dataset is quite small and composed by four classes, the number of different types of sounds is significantly higher, thus making the dataset very challenging. It is also important to highlight that some of the background sounds are similar to the classes of interest: for instance, some clips containing the voices of people in a crowded environment could be confused with screams.

The dataset has been augmented so that each clip has been recorded again at different signal-to-noise ratio (namely 5dB, 10dB, 15dB, 20dB, 25dB and 30dB), in order to simulate different distances from the microphone. In particular, mixing the event of interest in the environmental sound, the energy of the former is attenuated or amplified with a coefficient λ computed as follows:

$$\lambda_i^j = 10^{\frac{SNR_j}{20}} * \frac{RMS(B_i(t))}{RMS(E_i(t))} \quad (9)$$

Therefore, the coefficient λ_i^j for the j -th SNR value and the i -th event of interest depends on the specific SNR value and on the root mean square (RMS) of the i -th background sound $B_i(t)$ and event of interest $E_i(t)$.

Such clips, recorded with different SNRs, have been used to perform the problem-driven data augmentation described in the previous section.

B. Experimental Protocol

In order to perform a fair comparison with state of the art methodologies, we decide to use the same experimental protocol originally proposed in [37]. Each frame is labelled with one of the considered classes, namely B, GB, G or S, by adopting the same protocol described in [37]. In particular, we say that the audio frame contains an event of interest (GB, G, S) if the latter occupies at least 50% of the frame or if at least 50% of the event duration is inside the frame.

In the whole, we extracted about 83,000 gammatonegram images: 50,000 of them have been used for training the network and the remaining 33,000 for testing. The same number of spectrogram images have been also extracted for comparison purposes. Both gammatonegram and spectrogram

TABLE III

COMPARISON OF THE PROPOSED APPROACH WITH OTHER CNNs AND STATE OF THE ART METHODOLOGIES IN TERMS OF RECOGNITION RATE (RR), FALSE POSITIVE RATE (FPR), MISS RATE (MR) AND ERROR RATE (ER) APPLYING THE FRAME BY FRAME AND THE EVENT BASED PROTOCOLS. THE METHODS ARE ORDERED BY DESCENDING EVENT BASED RECOGNITION RATE

Method	Ref.	Representation	Deep	Training	Frame by frame (%)				Event-based (%)			
					RR	FPR	MR	ER	RR	FPR	MR	ER
AReN	Proposed	Gammatonegram	Yes	FT	94.65	0.78	4.97	0.38	99.42	0.78	0.40	0.18
VGG-19	[38]	Spectrogram	Yes	FT	92.09	2.89	7.54	0.37	99.14	2.89	0.83	0.03
AReN	-	Spectrogram	Yes	FT	92.49	4.19	6.93	0.58	99.04	4.19	0.52	0.44
MobileNet	[39]	Gammatonegram	Yes	FT	92.46	0.39	7.25	0.29	98.51	0.39	1.35	0.14
VGG-19	[38]	Gammatonegram	Yes	FT	92.06	0.58	7.80	0.14	98.26	0.58	1.63	0.11
MobileNet	[39]	Spectrogram	Yes	FT	92.97	2.21	6.77	0.26	98.04	2.21	1.85	0.11
AENet	[16]	Spectrogram	Yes	TS	86.45	8.64	7.92	5.63	97.57	8.64	0.13	2.30
MobileNet	[39]	Spectrogram	Yes	TS	82.94	4.89	12.11	4.95	97.33	4.89	1.17	1.50
VGG-19	[38]	Gammatonegram	Yes	TS	91.83	1.60	7.62	0.55	97.15	1.60	2.56	0.27
AReN	-	Spectrogram	Yes	TS	84.31	2.55	14.24	1.45	97.10	2.55	1.70	1.20
AENet	[16]	Gammatonegram	Yes	FT	88.73	2.59	8.84	2.43	97.05	2.59	1.28	1.67
AReN	-	Gammatonegram	Yes	TS	90.15	2.85	8.90	0.95	97.04	2.85	2.54	0.42
NASNet	[40]	Gammatonegram	Yes	FT	83.42	0.25	16.29	0.29	96.55	0.25	3.12	0.33
HRNN	[41]	STFT + Mel	Yes	-					96.55			
MobileNet	[39]	Gammatonegram	Yes	TS	80.85	2.43	13.27	5.88	96.28	2.43	1.87	1.85
VGG-19	[38]	Spectrogram	Yes	TS	89.93	3.27	7.32	2.75	96.09	3.27	1.24	2.67
AENet	[16]	Gammatonegram	Yes	TS	91.18	3.10	4.86	3.96	96.02	3.10	2.24	1.74
COPE	[42]	Gammatonegram	No	-					96.00	4.30	3.10	0.90
AENet	[16]	Spectrogram	Yes	FT	78.82	5.35	16.69	4.48	95.66	5.35	0.74	3.60
SoundNet	[43]	Raw audio	Yes	-					93.33	22.34	0.67	6.00
NASNet	[40]	Gammatonegram	Yes	TS	68.12	0.26	30.81	1.07	88.98	0.26	10.21	0.81
GT	[7]	Gammatonegram	No	-					88.60	1.40	9.90	1.40
HF+BoW+SVM	[37]	HF	No	-					86.70	3.10	10.70	2.60
NASNet	[40]	Spectrogram	Yes	TS	60.28	0.53	37.61	2.11	85.04	0.53	12.82	2.13
NASNet	[40]	Spectrogram	Yes	FT	66.43	0.00	33.54	0.03	84.96	0.00	15.01	0.03

images are available upon request for future benchmarking purposes.¹

Applying the classifiers to these samples and analyzing the correspondence between the response and the groundtruth in terms of true positives, false positives, misses (false negatives) and classification errors, we are interested in computing the following performance metrics:

- Recognition Rate (RR), computed as the ratio between the number of true positives, namely the correctly classified events of interest, and the total number of events.
- False Positive Rate (FPR), computed as the ratio between the number of false positives, namely the events detected when only the background is present, and the total number of events.
- Miss Rate (MR), computed as the ratio between the misses, namely the number of events of interest not detected, and the total number of events.
- Error Rate (ER): computed as the ratio between the classification errors, namely the number of events of interest detected but misclassified, and the total number of events.

The protocol typically used for determining the correct detections on this dataset is *event-based*, in the sense that the evaluation is not done by considering the single audio frame

but the whole event. In particular, an event of interest, namely GB, G and S, is considered correctly detected if it is identified for at least one of the consecutive frames in which it appears.

In this paper, we also report the *frame by frame* results; according to this protocol, the recognition rate, the miss rate and the error rate are computed by considering the total number of audio frames, thus in our case the total number of gammatonegram (or spectrogram) images.

Finally, as usual in the literature, for both the frame by frame and the event-based protocols only one occurrence is counted in case a false positive is detected in consecutive frames. Therefore, the FPR assumes the same value applying both the experimental protocols.

C. Experimental Analysis

the capability to break down the falsepositive rate thanks to the gammatonegram, the low sensitivity to the noise, the generalization capability We design our experimental analysis with the aim of pointing out the capability of the proposed network to achieve a good trade-off between sensitivity (high recognition rate) and specificity (low false positive rate), its robustness to environmental noise, its generalization capabilities in different environments and its usability in real applications. In particular, we evaluate the impact on the performance of three main aspects: the input representation, the training strategy and the network architecture.

¹<http://mivia.unisa.it>

As for the first aspect, we represent the audio signal through the gammatonegram and the spectrogram, in order to evaluate the behaviour of the network by using the two input representations and to demonstrate the effectiveness of the gammatonegram.

Regarding the training, we investigate two different strategies: (1) *training from scratch* (TS) and *fine tuning* (FT). The TS requires that the weights of both the convolutional and the fully connected layers are randomly initialised and that the whole network is trained by using the whole augmented training set. The FT uses instead the ImageNet dataset, widely exploited in image processing tasks for training the deep networks before fine tuning, for learning the model and initialising the weights of the convolutional and the fully connected layers; then, a fine tuning is performed over both the convolutional and the fully connected layers with the Mivia Audio Events dataset. Note that ImageNet has been already used in the literature also for audio analysis tasks (for instance in [16]).

The choice of the network architectures in our analysis has been guided by a recent benchmarking paper [44], where the deep networks have been compared on the problem of image recognition. NASNet [40], although being among the slowest networks, is able to reach the highest accuracy on the problem of image recognition. Viceversa, MobileNet [39] is the fastest network, since it is thought for mobile and embedded devices. VGG-19 [38] is widely used in the community of image recognition, since it guarantees a good trade-off between accuracy and computational burden due to its depth. It is also important to note that VGG-19 can be considered as a kind of baseline approach, since it inspires the proposed AReN.

AENet is, according to our knowledge, the only network architecture proposed in the literature for sound event recognition whose weights are freely available for benchmarking purposes. Differently from the previously mentioned deep architectures, in this case we used the weights provided by the authors. Note that the network has been firstly pre-trained with ImageNet and, then, fine tuned with the dataset introduced by the authors, containing 41 categories of sound events represented as spectrograms; thus, the derived representation should be well enough for describing spectrogram images. In our experiments with the FT strategy, we fine tuned AENet with different types of sound events to be recognized; indeed, we started from the weights made available by the authors and, then, we evaluated the performance of the network; as for the TS strategy, we have randomly initialized the weights using the same network architecture.

For the sake of completeness, we also include in the comparison some other approaches whose results are publicly available and represent the state of the art on the considered dataset; note that in this case, differently from the other networks that we have re-trained and tested using the Mivia Audio Events dataset, we only report the results already published in the literature. Among these methods, [41] and [43] are based on CNNs, while the others, namely [7], [42] and [37], are not based on deep networks. In particular, the audio data given in input to the hierarchical recurrent neural network (HRNN) proposed in [41] is pre-processed in the spectral domain by

using a Short-Time Fourier Transform (STFT) and a Log-Mel filtering. The obtained feature vector is then used to detect the presence of an event with a first-level LSTM; whether this network identifies an event of interest, a second-level LSTM recognizes its category. SoundNet [43] is instead a 8-layer CNN fed directly with raw audio. The results for both the networks have been reported in [42], where the trainable COPE feature extractors are proposed to learn an effective sound representation starting from a visual based representation. Haar features and Adaboost is used in [7] in combination with an SVM classifier. Finally, in [37] a long time sound descriptor is computed by using a bag of words approach applied on aural words, obtained by extracting a set of spectral, temporal and energy features from short-time audio frames; the event detection and classification is then performed with a SVM.

D. Results

The results of our experimental analysis are reported in Table III. The methods are ordered according to the RR achieved on the frame by frame experimentation (from the best one to the worst one). The experimental results have undergone a t-test, which confirms their statistical significance with a confidence of over 99%.

The results achieved by the proposed approach are reported in the first row; we obtain an impressive 94.65% of RR with a frame by frame analysis and a still more impressive 99.42% RR with an event based evaluation. We can note that the proposed approach is the best one available in the literature in terms of RR with both the experimental protocols.

From the table we can also note that the best recognition rate has been obtained performing the fine tuning instead than the training from scratch (99.42% vs 97.04%). It means that a proper initialization of the weights, even if conducted over a set of images related to a completely different problem (namely image recognition) is better than a completely random initialization and allow an increasing of about 2.4%.

Moreover, the gammatonegram confirms to be a very good representation, since it allows not only a small increase in the recognition rate of 0.38 points percentage with respect to the use of the spectrogram (99.42% vs 99.04%), but especially a remarkable reduction of the false positive rate (0.78% vs 4.19%). This is due to the high representation capability of the gammatonegram, which allows to well distinguish events of interest from background events, even if very challenging.

It is also worth noting that the results achieved frame by frame by AReN are even more encouraging. The obtained recognition rate is almost 2% higher than the second best performance (94.65% vs 92.97%), while the miss rate is slightly outperformed only by a version of AENet (4.97% vs 4.86%) and is better than the third top rank of about 2%. This result confirms the effectiveness of the proposed solution, which achieves a very good trade-off between sensitivity and specificity.

In addition, the proposed architecture achieves a very good trade-off between false positive rate (0.78%) and miss rate (0.40%); these values, which are among the best both if

TABLE IV

FRAME BY FRAME MISCLASSIFICATION MATRIX OF THE PROPOSED ARCHITECTURE FINE TUNED WITH GAMMATONEGRAMS (A), FINE TUNED WITH THE SPECTROGRAM (B), TRAINED FROM SCRATCH WITH GAMMATONEGRAMS (C) AND TRAINED FROM SCRATCH WITH SPECTROGRAM (D). THE DETECTED CLASSES ARE REPORTED IN THE FIRST THREE COLUMNS, WHILE THE ROWS CONTAIN THE GROUNDTRUTH CLASSES; THE LAST COLUMN REFERS TO THE MISS RATE ON THE SPECIFIC CLASSES

	GB	GS	S	MR
GB	94.11	0.13	0.06	5.70
GS	0.57	97.96	0.17	1.30
S	0.13	0.11	91.83	7.93

(a) Gammatonegram + FT

	GB	GS	S	MR
GB	91.54	0.19	0.19	8.09
GS	0.46	97.39	0.22	1.93
S	0.33	0.35	88.52	10.80

(b) Spectrogram + FT

	GB	GS	S	MR
GB	94.17	0.44	0.09	5.30
GS	0.96	98.22	0.28	0.54
S	0.59	0.48	78.06	20.87

(c) Gammatonegram + TS

	GB	GS	S	MR
GB	87.15	0.15	0.00	12.70
GS	3.15	93.52	0.17	3.17
S	0.63	0.26	72.26	26.85

(d) Spectrogram + TS

analyzed separately and if considering their sum, point out the capability of the network to be at the same time sensitive in detecting events (low MR) and sufficiently careful in avoiding false alarms (low FPR). As evident, this is a very desirable and not negligible property when dealing with audio surveillance applications.

The frame by frame confusion matrices reported in Tables IV (a-d) allow to characterize the errors done by AReN. As already evident from Table III, most of them are misses; in addition, we can note that the system misses more glass breaking and scream events than gun shots. Analyzing the samples of the dataset, we have noticed that this result is reasonable, since a lot of background noises consist of crowd screaming, cheering and whistles, making it easy to generate confusion with glass breaking events and isolated scream sounds. Think, for example, to a music concert where people are already screaming and cheering; in such environment, recognizing isolated screams due to a dangerous situation would certainly be a more challenging task. In addition, it is also worth noting that the screams have some frequential components very similar to those of the speech; therefore, recognizing such kind of events is definitely not trivial and even more challenging for generic audio surveillance systems that must recognize also different events of interest. This observation is further confirmed by the experimental evidence that the scream is the event category generating the higher number of false positives. Anyway, as evident from Table IV, the proposed

AReN fine tuned with gammatonegrams is able to achieve a scream events miss rate lower than the other solutions (7.93% vs 10.80%, 20.87% and 26.85%), thus confirming its effectiveness.

1) *Analysis of the Sensitivity to the Noise*: Another assessment that it is worth making to characterize the performance of AReN is the analysis of the sensitivity to the environmental noise. Indeed, the proposed problem-driven data augmentation has been introduced for increasing the robustness of the network to the environmental noise. The considered dataset is particularly suited for this evaluation, as it consists of audio samples recorded with an increasing SNR (six versions of the same sample with SNR equal to 5, 10, 15, 20, 25 and 30 dB, respectively).

Ideally, the network should maintain the same recognition rate independently from the noise, but actually a performance degradation is expected decreasing the SNR, especially if the value is lower than 10 dB. Such degradation typically depends on the distance from the audio source and the energy of the signal, that in real environments are surely very variable. To reproduce this variability, we plan two experiments: i) we separately apply AReN, trained as described with a problem-driven data augmentation, on the audio samples with different SNRs (5, 10, 15, 20, 25, 30 dB) to evaluate the performance degradation with low SNRs; ii) we perform several experimentations varying both the training and the test set in order to demonstrate the positive impact of the problem-driven data augmentation.

As for the first experiment, the trend of the frame by frame recognition rate varying the SNR, depicted in Fig. 2, shows a very good capability of AReN to reject the noise. As a matter of fact, with an average recognition rate of 94.65%, the network achieves an above-average performance with a SNR greater than 10 (with a peak of 96.29% at 30 dB), paying a significant degradation (88.85%) only with a SNR equal to 5 dB. This result is exactly the same also when the network is fed with spectrograms and using different training strategies; the average recognition rate of these networks is above 90% up to 10 dB, reaching around 83% at 5 dB. The cause of the recognition rate drop is glaring looking at Fig. 3; in fact, the signal distortion starts appearing with a SNR equal to 15 dB, but it becomes very evident at 10 and even more at 5 dB.

As for the second experiment, the results reported in Fig. 4 confirm that the problem-driven data augmentation allows to achieve better robustness towards the varying levels of SNR. Indeed, as expected, the performance on the test set are higher when audio samples with high SNRs are used as test data and lower when the SNR of the sounds in the test set decreases. However, using the problem-driven data augmentation the drop is of course still present, but it is definitely less prominent and the overall performance, obtained by averaging the results on the whole range of SNRs, is higher.

In addition, we perform a third experiment in order to demonstrate the effectiveness of the proposed data augmentation problem driven with respect to the standard image based data augmentation. The results are shown in Fig. 4; in the last column, we report the results obtained with traditional image based data augmentation. In particular, we extensively

Fig. 2. Frame by frame recognition rate of ARēN, fed with gammatonegrams (G) or spectrograms (S) and trained with FT or TS, at different SNRs, from 5 to 30 dB; as expected, in all the cases the recognition rate increases with the SNR, being the sound less noisy, and remains above-average with SNR > 10dB.

Fig. 3. Gammatonegram examples of a glass breaking event occurring at different SNRs, from 5 to 30 dB; with a small SNR (≤ 10), the event is not easily recognizable.

experiment several combinations: the reported results refer to the best combination, namely the augmentation with samples obtained with vertical and horizontal flip, horizontal and vertical shift and rotation in the range $[-10, 10]$ degrees. The peak of the recognition rate (94.81%) is achieved at 30 dB; this value is anyway 1.84% lower than the one achieved with the proposed problem-driven data augmentation. However, even if we can note an improvement of the performance with respect to a training with 30 and 25-30 dB (first and second column in the figure), the performance dramatically

decreases with a SNR lower than 20 dB, so obtaining an average recognition rate of 76.49%. It is evident that this value is significantly lower than the best result achieved by the proposed solution. This experimental evidence confirms that a standard data augmentation, even if very promising when dealing with traditional image based tasks, is not able to successfully deal with the distortions introduced by the environmental noise. Viceversa, a data augmentation driven by the specific problem is definitively more effective.

Therefore, we can conclude that the proposed network demonstrates a low sensitivity to the noise and a good capability to maintain a high recognition rate whether the signal distortion remains within reasonable limits. Furthermore, we can state that the problem-driven data augmentation is more effective than the standard image-based data augmentation for audio events classification, especially in noisy environments.

2) *Considerations About the Network Architecture:* The result of our network and the second best recognition rate achieved by VGG-19 (99.14%) confirms that the depth of the network, combined with a small kernel size of the filters is important to achieve good performance in a context where the amount of data for the training is relatively small. According to this observation, the introduction of the pyramidal set of fully connected layers in our ARēN allows to further increase the recognition rate (+0.4%), preserving at the same time the performance in terms of ER and FPR.

The small amount of data available for training also justifies the lower recognition rate achieved by MobileNet (98.51%) and NASNet (96.55%). Indeed, although so promising in image recognition tasks, such networks seem to be very sensitive with respect to the amount of data required for training and thus seem to be a bit less suited for managing audio analysis tasks. Especially the latter, even more looking at the results achieved frame by frame, suffers the lack of a large amount of samples. It is also important to highlight that AENet achieves in the best case a smaller recognition rate (97.57%) with respect to ARēN and a very high false positive rate (8.64%).

Anyway, all the considered deep architectures achieve at least in one case a recognition rate higher than all the methods previously proposed in the literature. On one hand, these results demonstrate the superiority of the deep networks with respect to classical approaches based on handcrafted features [7], [42] [37]; on the other hand, the effectiveness of the time-frequency representations based on images, namely gammatonegram and spectrogram, in combination with the greater depth of the network emerges from the comparison with other CNNs proposed in the literature [41], [43], which rely instead on STFT and Mel filters or raw audio, respectively.

3) *Fine Tuning vs Training From Scratch:* The effectiveness of the fine tuning as training strategy clearly emerges from the experimental results. All the considered network architectures, namely ARēN, VGG-19, MobileNet and NasNet, except AENet, achieve the best recognition rate by applying the fine tuning; moreover, it is not a case that the six top-rank approaches adopt such training strategy. The exception of AENet, which obtains a higher recognition rate trained from scratch (97.57% vs 95.66%) is probably due to the fact that

Fig. 4. Frame by frame recognition rate of AReN, with different SNR values; each group of columns refers to a specific training set TS_X_30 , including training samples with SNR in the range $[X, 30]$ dB, while the last group is a set obtained with standard data augmentation. TS_5_30 is the full training set, obtained with the proposed problem-driven data augmentation, which allows to achieve the best performance.

the starting weights, obtained after a double pre-training with ImageNet and with a dataset containing 41 sound categories, are too specialized for the sound event recognition task for which the network has been designed.

4) *The Positive Effect of the Gammatonegram*: The gammatonegram is in principle an audio input representation that should simplify the distinction between events and background. According to this observation, Fig. 1 points out the difference between the image that represents the background from the ones which depict a glass breaking, a gun shot and a scream; such difference is surely more evident in the gammatonegram representation than in the spectrogram one. The qualitative analysis is confirmed by the experimental results, since the average false positive rate achieved by the evaluated deep networks which adopt the gammatonegram (1.48%) is significantly lower than the corresponding FPR obtained with the spectrogram (3.45%). This result is evident for all the network architectures, since the gammatonegram representation breaks down the false positive preserving a high recognition rate not only using AReN (0.78% vs 4.19%), but also with VGG-19 (0.58% vs 2.89%), MobileNet (0.39% vs 2.21%) and AENet (3.10% vs 8.64%).

Such experimental evidence broadly justifies the adoption of this representation for the problem at hand.

5) *Generalization Capabilities*: We perform experiments on other datasets to demonstrate the capability of the proposed network to generalize in similar but different audio surveillance environments. In particular, we chose two datasets recorded for smart city surveillance, namely the Mivia Road Events [8] and the Sound Events for Surveillance Applications (SESA) [49]. The first is focused on audio events for road surveillance applications, namely tire skidding and car crash, in addition to the background. The audio files are divided into 4 folds containing 100 events each, and the recommended experimental protocol is a 4-fold cross validation. The second includes three unbalanced sound categories associated to security threats, in addition to a casual (background) class: gunshot, explosion and siren (which also includes car alarms and building alarms). The dataset is challenging especially for the similarity between the events of interest and the casual class, that is also composed by audio recordings with fireworks, thunderstorms, hammering, and many others, easily confusable with gunshots, explosions and sirens. The samples are partitioned into training and test set and the results are

TABLE V

COMPARISON OF THE PROPOSED APPROACH WITH OTHER CNNs AND STATE OF THE ART METHODOLOGIES ON THE MIVIA ROAD EVENTS DATASET IN TERMS OF RECOGNITION RATE (RR), FALSE POSITIVE RATE (FPR), MISS RATE (MR) AND ERROR RATE (ER) APPLYING THE EVENT BASED PROTOCOL. THE METHODS ARE ORDERED BY DESCENDING RECOGNITION RATE

Method	Ref.	Representation	Deep	Event-based (%)			
				RR	FPR	MR	ER
AReN	Proposed	Gammatonegram	Yes	100.00	2.01	0.00	0.00
MobileNet	[39]	Gammatonegram	Yes	100.00	2.96	0.00	0.00
VGG-19	[38]	Spectrogram	Yes	100.00	3.18	0.00	0.00
AReN	-	Spectrogram	Yes	100.00	3.89	0.00	0.00
VGG-19	[38]	Gammatonegram	Yes	100.00	4.17	0.00	0.00
MobileNet	[39]	Spectrogram	Yes	99.50	3.72	0.00	0.50
NasNet	[40]	Gammatonegram	Yes	99.25	2.83	0.25	0.50
AENet	[16]	Spectrogram	Yes	99.25	5.57	0.25	0.50
AENet	[16]	Gammatonegram	Yes	96.00	4.44	3.50	0.50
COPE	[42]	Gammatonegram	No	94.00	4.00	4.75	1.25
BoF	[45]	Spectral features and aural words	No	82.00	2.85	17.75	0.25
BoF	[5]	MFCC	No	80.25	7.69	19.00	0.75

TABLE VI

COMPARISON OF THE PROPOSED APPROACH WITH OTHER CNNs TRAINED USING GAMMATONEGRAM AND STATE OF THE ART METHODOLOGIES ON THE SESA DATASET. THE METHODS ARE ORDERED BY DESCENDING EVENT BASED ACCURACY

Method	Ref.	Representation	Deep	Accuracy (%)	
				Frame by frame	Event based
AReN	Proposed	Gammatonegram	Yes	85.63	91.43
Xception	[46]	Gammatonegram	Yes	83.49	90.48
ResNet-50	[47]	Gammatonegram	Yes	82.81	90.48
VGG-19	[38]	Gammatonegram	Yes	84.68	89.52
MobileNet	[39]	Gammatonegram	Yes	81.93	89.52
Inception v3	[48]	Gammatonegram	Yes	80.26	88.57
Bagging	[49]	Handcrafted features + PCA	No		73.05
SGD					72.13
Gradient Boosting					71.88
SVM					71.73
Random Forest					70.96
KNN					70.12
Ridge					69.51
Passive Aggressive					69.40
Perceptron					68.20
AdaBoost					65.46
Decision Tree					65.38

measured in terms of accuracy. All the samples have been recorded with the same configuration: WAV, mono, 16 kHz, and 16-bit lasting up to 33 seconds each.

The results of these experiments are reported in Table V and Table VI. AReN achieves state of the art results on both the datasets, so demonstrating its effectiveness in this field. The 100% of recognition rate on the Mivia Road Events dataset is impressive, even if shared with MobileNet and VGG-19; it means that the proposed network is able to detect and correctly classify all the events of interest. In addition, the result is even more relevant considering that at the same time AReN obtains the state of the art false positive rate (2.01%, better than the 2.83% of NasNet and the 2.96% of MobileNet), confirming the excellent trade-off between detection sensitivity and specificity achieved by the proposed network. As for the experiment on the SESA dataset, we also include in the comparison other

modern CNNs, namely Xception [46], ResNet-50 [47] and Inception v3 [48], trained with gammatonegrams in order to evaluate the performance of different networks by using such audio representation. In addition, we perform a frame by frame analysis, even if the original protocol does not include such evaluation; in particular, we divide the audio recordings in frames of 900 ms, since this size allows to achieve the best performance on the validation set. AReN outperforms all the other methods, achieving a frame by frame accuracy of 85.63% (vs 84.68% of VGG-19) and an event based accuracy of 91.43% (vs 90.48% of Xception and ResNet-50), confirming once again its effectiveness and generalization capability for the problem at hand.

6) *Real-Time Experimentation*: To the best of our knowledge, the literature in the field of audio surveillance does not include experiments carried out in real-time and in real

TABLE VII

OFFLINE AND ONLINE PERFORMANCE OF AREN IN TERMS OF RECOGNITION RATE (RR), FALSE POSITIVE RATE (FPR), MISS RATE (MR) AND ERROR RATE (ER) APPLYING THE EVENT BASED PROTOCOL

Microphone	Setup	Event-based (%)			
		RR	FPR	MR	ER
Axis T83	Offline	99.42	0.78	0.40	0.18
Axis T83	Online	91.52	0.29	3.30	5.18
Sony ECM-PC50	Online	90.64	4.52	2.08	7.28

environments. To overcome this lack, we have designed an experiment for evaluating the performance of AREN in a setup that is closer to the real conditions, namely in a laboratory of the University of Salerno with people working and speaking. Such experimentation has been carried in a non-anechoic room with a double aim: i) to verify the capability of AREN to achieve a high recognition rate even on audio samples recorded in noisy environments; ii) to demonstrate that the proposed network is able to recognize the events of interest in real-time.

First of all, a multi-threaded architecture is strictly necessary to perform the continuous acquisition of the audio stream and the real-time processing simultaneously. To this aim, the first thread is dedicated to the acquisition of the audio events of the dataset, that are reproduced by the speakers of a workstation and recorded using another microphone positioned 1 meter far from them. In order to verify whether the use of different microphones may affect the performance of the proposed network, we experiment not only the Axis T83, adopted for recording the dataset, but also a Sony ECM-PC50 microphone. The input audio stream is then packed in 3 second frames and then fed as input to the second thread. The latter receives the raw audio, extracts the gammatonegram representation and applies the proposed network to recognize the events of interest.

In order to compare such version, hereinafter *online*, with the original one, called in this experiment *offline*, we have reproduced all the sounds available in the audio surveillance test set, considering the SNR from 5 to 30 dB. To mitigate the effect of the uncontrolled environmental sounds recorded in a non-anechoic chamber, we have repeated five times the online experiment; the results are finally obtained by averaging the performance metrics of the five experiments. Table VII reports the recognition rate, the false positive rate, the miss rate and the error rate achieved in the online setup.

As expected, the performance with this setup is lower than the one achieved in the classic experiment, since the audio samples acquired in real-time are more likely to be affected by additional noise. What is surprising is that, even though the microphone is quite close to the audio source, there is a significant decrease of the recognition rate (99.42% vs 91.52% vs 90.64%, with a corresponding increasing of FPR, MR and ER) that does not depend on the type of microphone used for the test.

We have analyzed the same audio samples available in the dataset and recorded by the microphone; Fig. 5 reports some relevant examples. It is evident that the sounds recorded online are significantly more noisy than the ones used for training the network. A possible explanation for such decrease is that

Fig. 5. Differences between the same events available in the dataset (offline) and recorded in the online experiments.

the bass frequencies of the recorded sounds are much more emphasized than the ones of the original dataset and there is also the presence of a significant amount of reverberation. This is probably caused by the acoustic characteristics of the room in which the tests have been done and generally leads to a loss of the audio quality and, consequently, of the classification accuracy.

Such experimental findings point out the importance of a deep characterization of the audio surveillance systems not only on standard datasets, but also in real environments.

7) *Processing Time*: The real-time experimentation has been also useful for evaluating the processing time required by the proposed network on two different processors.

We notice that AREN spends 200 milliseconds to elaborate an audio frame of 3 seconds on an Intel Core i7-6700HQ CPU @ 2.60GHz and 900 milliseconds on a less powerful AMD A6-4455M APU @ 2.1 GHz. It means that our network is able to process 5 audio frames per second, which are definitely sufficient for the use in real audio surveillance applications.

IV. CONCLUSIONS

The wide range of applications where audio surveillance can be profitably used has surely contributed to the growing interest of the scientific community towards this topic.

Motivated by this trend, in this paper we propose a method for identifying events of interest by analyzing the audio stream acquired by surveillance microphones. The audio stream is represented by means of gammatonegram, a brain inspired representation of the sound emulating the way in which the human auditory system works. Such images are used in order to feed AREN, the proposed 21 layers CNN with the following main features: a small receptive fields of the convolutional layers and a pyramidal structure of the fully connected layers. Such choices have the main advantage of guaranteeing a good generalisation capability also in presence of a relatively small training set.

The system achieves 99.42% recognition rate on the MIVIA Audio Events dataset, overcoming the best result obtained so far in the state of the art (96.55%). The experimental analysis shows that the gammatonegram representation and the adopted architecture, combining the advantages of VGG-19 and the use of a pyramidal set of fully connected layers, allow to break down the false positive rate and to achieve a low sensitivity to the noise. The fine tuning strategy demonstrated a higher reliability than a training from scratch, while the generalization capability on the MIVIA Road Events dataset, on the SESA

dataset and in a simulation of the real environments, together with the satisfactory processing speed, confirm the effectiveness of the proposed solution. The excellent achieved results encourage the use and the evaluation of ARen in real audio surveillance applications.

Future works may include the possibility to extend the proposed approach to other audio analysis problems, including ambient assisted living and sentiment analysis.

REFERENCES

- [1] B Research. (2017). *Global Video Surveillance Market-Analysis and Forecast (2017–2023) Focus on Ecosystem (Camera, Monitor, Storage, Software and Service) and Application (Infrastructure, Commercial, Residential, Industrial, Institutional and Others)*. [Online]. Available: <https://www.reportbuyer.com>
- [2] L. Brun, A. Saggese, and M. Vento, "Dynamic scene understanding for behavior analysis based on string kernels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1669–1681, Oct. 2014.
- [3] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: A review," *IEE Proc.-Vis., Image Signal Process.*, vol. 152, no. 2, pp. 192–204, Apr. 2005.
- [4] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surveys*, vol. 48, no. 4, pp. 52:1–52:46, May 2016, doi: [10.1145/2871183](https://doi.org/10.1145/2871183).
- [5] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, Jan. 2016.
- [6] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools Appl.*, vol. 75, no. 17, pp. 10407–10439, Sep. 2016, doi: [10.1007/s11042-015-3105-4](https://doi.org/10.1007/s11042-015-3105-4).
- [7] A. Saggese, N. Strisciuglio, M. Vento, and N. Petkov, "Time-frequency analysis for audio event detection in real scenarios," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 438–443.
- [8] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and N. Petkov, "Car crashes detection by audio analysis in crowded roads," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6.
- [9] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An ensemble of rejecting classifiers for anomaly detection of audio events," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 76–81.
- [10] K. Lopatka, P. Zwan, and A. Z. Czyzewski, "Dangerous sound event recognition using support vector machine classifiers," in *Advances in Multimedia and Network Information System Technologies*. Berlin, Germany: Springer, 2010, pp. 49–57, doi: [10.1007/978-3-642-14989-4_5](https://doi.org/10.1007/978-3-642-14989-4_5).
- [11] C.-Y. Wang, J.-C. Wang, A. Santoso, C.-C. Chiang, and C.-H. Wu, "Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1336–1351, Aug. 2018.
- [12] F. Guo, D. Yang, and X. Chen, "Using deep belief network to capture temporal information for audio event classification," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP)*, Sep. 2015, pp. 421–424.
- [13] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, Jan. 2017.
- [14] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.
- [15] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 559–563.
- [16] N. Takahashi, M. Gygli, and L. Van Gool, "AENet: Learning deep audio features for video analysis," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 513–524, Mar. 2018.
- [17] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3407–3411.
- [18] A. Mesaros *et al.*, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018.
- [19] A. F. Pour, M. Asgari, and M. R. Hasanabadi, "Gammatonegram based speaker identification," in *Proc. 4th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2014, pp. 52–55.
- [20] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, 2019.
- [21] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*. [Online]. Available: <http://arxiv.org/abs/1904.08779>
- [22] S. L. Phung and A. Bouzerdoum, "A pyramidal neural network for visual pattern recognition," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 329–343, Mar. 2007.
- [23] I. Ullah and A. Petrosino, "About pyramid structure in convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 1318–1324.
- [24] M. Slaney, "Auditory toolbox," Interval Res. Corp., Palo Alto, CA, USA, Tech. Rep. 10, 1998.
- [25] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. London, U.K.: Academic, 1986.
- [26] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, K. Horner, and L. Demany, Eds. New York, NY, USA: Pergamon, 1992, pp. 429–446. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/B978008041847650054X>, doi: [10.1016/B978-0-08-041847-6.50054-X](https://doi.org/10.1016/B978-0-08-041847-6.50054-X).
- [27] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: How the brain analyzes speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 4, Apr. 2007, p. IV-765.
- [28] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–5.
- [29] D. P. Ellis. (2009). *Gammatone-Like Spectrograms*. [Online]. Available: <http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram>
- [30] L. Marchegiani and P. Newman, "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals," in *Towards Autonomous Robotic Systems*, M. Giuliani, T. Assaf, and M. E. Giannaccini, Eds. Cham, Switzerland: Springer, 2018, pp. 247–259.
- [31] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.
- [32] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.
- [33] X. Ouyang, K. Gu, and P. Zhou, "Spatial pyramid pooling mechanism in 3D convolutional network for sentence-level classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2167–2179, Nov. 2018.
- [34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*. Madison, WI, USA: Omnipress, 2010, pp. 807–814. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3104322.3104425>
- [35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2999134.2999257>
- [36] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: <http://arxiv.org/abs/1609.04747>
- [37] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognit. Lett.*, vol. 65, pp. 22–28, Nov. 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167865515001981>
- [38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [39] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: <http://arxiv.org/abs/1704.04861>

- [40] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," 2017, *arXiv:1707.07012*. [Online]. Available: <http://arxiv.org/abs/1707.07012>
- [41] F. Colangelo, F. Battisti, M. Carli, A. Neri, and F. Calabr, "Enhancing audio surveillance with hierarchical recurrent neural networks," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
- [42] N. Strisciuglio, M. Vento, and N. Petkov, "Learning sound representations using trainable COPE feature extractors," 2019, *arXiv:1901.06904*. [Online]. Available: <http://arxiv.org/abs/1901.06904>
- [43] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 892–900.
- [44] S. Bianco, R. Cadene, L. Celona, and P. Napolitano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.
- [45] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2013, pp. 81–86.
- [46] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [49] T. Spadini, D. L. de Oliveira Silva, and R. Suyama, "Sound event recognition in a smart city surveillance context," 2019, *arXiv:1910.12369*. [Online]. Available: <http://arxiv.org/abs/1910.12369>

Antonio Greco received the Ph.D. degree in computer science and computer engineering from the University of Salerno in 2018. He is currently an Assistant Professor with the University of Salerno. His research interests include computer vision and machine learning techniques for video and audio surveillance applications. He serves as a referee for many journals and international conferences.

healthcare, finance, surveillance, manufacturing, robotics, and animal breeding. He is a member of the editorial boards of several journals.

Alessia Saggese received the Ph.D. degree in computer engineering from the University of Salerno, Italy, and the University of Caen Normandy, France, in 2014. She is currently an Assistant Professor with the University of Salerno. Her research interest includes basic methodologies and applications in computer vision and pattern recognition. She has been a member of the International Association for Pattern Recognition Technical Committee 15 on Graph-Based Representations in Pattern Recognition since 2012.

Mario Vento received the Ph.D. degree in computer engineering from the University of Naples Federico II in 1989. He is currently a Full Professor of computer science and artificial intelligence with the University of Salerno, Fisciano, Italy, where he is the Coordinator of the Artificial Vision Laboratory and the Dean of Department of Information and Electrical Engineering and Applied Mathematics. His research activities cover real-time video analysis and interpretation for video surveillance applications, classification techniques, either statistical, syntactic and structural, exact and inexact graph matching, multiexpert classification, and learning methodologies for structural descriptions. He is a Fellow Scientist of the International Association Pattern Recognition (IAPR). He is a fellow of IAPR. He served as the Chairman of the IAPR Technical Committee 15 on Graph-Based Representation in Pattern Recognition from 2002 to 2006.