Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration

Ohini Kafui Toffa and Max Mignotte

***Abstract*—This paper presents a new approach to classify environmental sounds using a texture feature local binary pattern (LBP) and audio features collaboration. To our knowledge, this is the first time that the LBP (or its variants), which has a proven track record in the field of image recognition and classification, has been generalized for 1D and combined with audio features for an environmental sound classification task. To this end, we have generalized and defined LBP-1D and local phase quantization (LPQ)-1D on the 1-dimensional (1D) audio signal and have applied the original LBP, the variance LBP (VARLBP) and the extended LBP (ELBP) thus generated to the spectrogram of the audio signal in order to model the sound texture. We have also extensively compared these new LBP-based features to the classical audio descriptors commonly used in environmental sound classification, such as MFCC, GFCC, CQT, chromagram, STE and ZCR. We have evaluated our algorithm on ESC-10 and ESC-50 datasets using classical machine learning algorithms, such as support vector machines (SVM), random forest and k-nearest neighbor (kNN). The results showed that the LBP features outperform the classical audio features. We mix the LBP features with the audio descriptors, and our best mixed model achieves state-of-the-art results for environmental sound classification: 88.5% on ESC-10 and 64.6% on ESC-50. Those results outperform the results of methods that used handcrafted features with classical machine learning algorithms and are similar to some convolutional neural network-based methods. Although our method is not the cutting edge of the state-of-the-art methods, it is faster than any convolutional neural network methods and represents a better choice when there is data scarcity or minimal computing power.**

***Index Terms*—Environmental sound classification, local binary pattern, local phase quantization, machine learning, ESC-50, audio signal spectrogram, SVM, random forest, kNN.**

I. Introduction

ENVIRONMENTAL sound classification (ESC) is the identification of daily sounds generated by the activities of humans or by nature, including a dog barking, fire crackling, baby crying, etc. Unlike music and speech, environmental sounds do not have a common structure since they actually have various origins and are very diverse. Their recognition and classification is one of the most important domains of audio signal processing, offering various applications: robot hearing, objectionable content detection, road surveillance, home automation and monitoring, and gunshot detection [1]–[5].

As with speech recognition, audio segmentation and other topics of audio signal processing, ESC relies on the extraction of specific and efficient audio features from time or frequency domains [6]–[8]. Some of those features are the short-time fundamental frequency (SFuF) [7], Gabor filters [4], [5], short-time energy (STE) [7], zero-crossing rate (ZCR) [7], constant-Q transform (CQT) [9], gammatone frequency cepstral coefficients (GFCC) [10], [11], chromagram [12], and mel-frequency cepstral coefficient (MFCC) [13]. The latter is the most commonly used for ESC. Though the spectrogram of the 1D audio signal is a 2D (time $\times$ frequency) image, it is not common to use image features to classify audio contents. With the growth and popularity of deep learning in image classification [14], numerous works have started using convolutional neural networks (CNNs), or a mix of spectrogram, MFCC and cross-recurrence plot (CRP) [16], to classify the spectrograms of sounds [15]. Nevertheless, to our knowledge, few research endeavors exploiting image descriptors have been published.

LBP is an operator or a texture analysis and characterization method based on the pixel neighborhood introduced by Ojala *et al.* [17]–[19]. Although this operator was originally developed for texture analysis and classification, it has become one of the most prominent and efficient texture descriptors used in many fields of image processing and computer vision, including image classification, biomedical image analysis, facial (and gender and family) recognition, and motion recognition [20]–[25], to name a few. Despite its popularity, few works using LBP have been dedicated to the ESC task. Kobayashi et *al.* [26] and Ren et *al.* [1] are among the rare authors to classify acoustic sounds using LBP. The authors in [26] applied LBP to the spectrogram, while others applied the LBP to the gammatone-like spectrogram. Our work falls within that category. We believe that we can proceed further than these previous works by directly applying the LBP to the 1D signal instead of the 2D spectrogram. First, we generalize the LBP for 1D by defining the local binary pattern-1D and local phase quantization-1D descriptors, which are directly computed from the 1D audio signal. Second, to achieve a strong characterization, we utilize a feature collaboration technique by combining the spectrogram-based 2D LBP descriptors (original LBP, variance LBP, and extended LBP) and audio features (MFCC, GFCC, ZCR, CHROMA, CQT, and STE) to successfully classify environmental sound. Third, we demonstrate that the proposed method offers the benefit of running

Manuscript received December 7, 2019; revised May 17, 2020 and September 13, 2020; accepted October 19, 2020. Date of publication November 4, 2020; date of current version November 18, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chi-Chun Lee. *(Corresponding author: Ohini Toffa.)*

The authors are with the vision lab of the Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal, Faculté des Arts et des Sciences, Montréal H3C 3J7, QC, Canada (e-mail: ohinfa@yahoo.fr; mignotte@iro.umontreal.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3035275.

Digital Object Identifier 10.1109/TMM.2020.3035275

faster than a deep neural network model that runs on a low-end GPU. The proposed method achieves interesting performance compared to the other methods.

The rest of this paper is organized as follows: Section 2 describes previous works on the ESC task. Section 3 details our method, which consists of 1D and 2D LBP descriptors as well as a combination with audio descriptors. Section 4 provides and analyses the results of our experiments, and Section 5 concludes the paper.

II. Related Work

ESC usually consists of the extraction of manually designed features that are then processed by a conventional classifier such as support vector machines (SVM), random forest or k-nearest neighbor (kNN) [1], [2], [4], [5], [7], [11], [27]–[29]. A good survey work [8] separates the methods into stationary (MFCC, STE, ZCR, MPEG-7, Gabor filters, Chroma, ZCR, STE,and linear prediction coefficients) and nonstationary (or wavelet-based) approaches such as continuous wavelet transform (CWT), fast wavelet transform (FWT), and Gaussian mixture models (GMM).

Most of the features listed in the previous paragraph are audio features. Kobayashi et *al.* [26] in 2014 were the first authors to use an image descriptor LBP to tackle a sound classification task. They enhanced the discriminative power of the LBP features with L_2-Hellinger normalization and obtained good classification results using linear SVM on RWCP, a sound-event dataset. Their work was followed in 2017 by Ren et *al.*, [1], who applied a multichannel LBP to the gammatone spectrogram for robot hearing and demonstrated good performance on two sound-event datasets: RWCP and NTU-SEC.

With the success of deep learning in the field of image classification, many works have replaced the conventional classifiers with a CNN that is able to better learn the time-frequency features using weight-sharing and pooling. In this context, Huzaifah [30] compared CQT, CWT and short-time Fourier transform (STFT) on CNNs. Sharma et *al.* [31] implemented a deep CNN of multiple features channels composed of MFCC, GFCC, CQT and chromagram. On the other hand, the CNN is sometimes directly applied to the signal with end-to-end training [32], [33], or to its spectrogram without preliminary feature extraction [15]. The authors of SoundNet [34] trained their network by transferring discriminative knowledge from visual recognition networks into sound networks. Boddapati et *al.* [16] achieved good classification using transfer learning with image recognition networks GoogleLeNet and AlexNet. Also worth mentioning are the temporal attention mechanism [35], restricted Boltzmann machine (RBM) [36], very deep network [37], between-class technique [38], etc., to name only those among a long list of deep learning methods that have tackled ESC with success. While the CNN methods perform better than classical and conventional classifiers, they face issues of data scarcity or a lack of diversity in the datasets. These issues are usually resolved by data augmentation techniques such as time stretching, pitch shifting, and background noise and dynamic range compression [15], [39]. Despite the good accuracy achieved by CNN methods, they are time and resource consuming. For that reason, classical machine learning methods coupled with handcrafted features represents a good trade-off between accuracy and speed in the presence of a relatively small quantity of data.

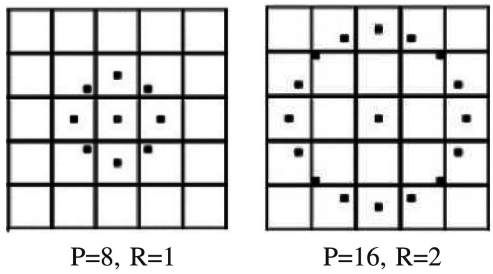


Fig. 1. Neighborhood of P pixels and radius R.

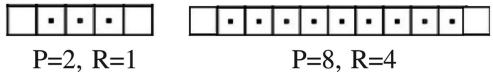


Fig. 2. LBP1D neighborhood of P pixels and radius R.

III. Proposed Method

It is common in audio or image classification tasks to aggregate multiple features in order to achieve higher accuracy [8], [31]. This is due to the complementary structure of many features. The particularity of our model is to use two different types of features: image and audio. In this section, we will describe the image and audio features, the result of each feature when applied to audio from the ESC-50 dataset (see Figs. 3 and 4) and the features collaboration technique.

A. Image Features

In this section, we present features that are commonly used in the image recognition and classification domain and their adaptation for audio classification.

1) Lbp/Var: LBP, as proposed by Ojala et *al.* [17]–[19], characterizes an image by a group of local patterns or microtexture. A local pattern is formed by encoding the difference in gray level between the pixel in the center and its neighbors, considering only the sign. The resulting binary codes of M-bits are concatenated to a decimal number. The histogram of different local patterns is used as a texture descriptor. Initially designed for a 3×3 pixel neighborhood, the LBP operator was quickly extended to a circular neighborhood of P pixels and radius R (see Fig. 1).

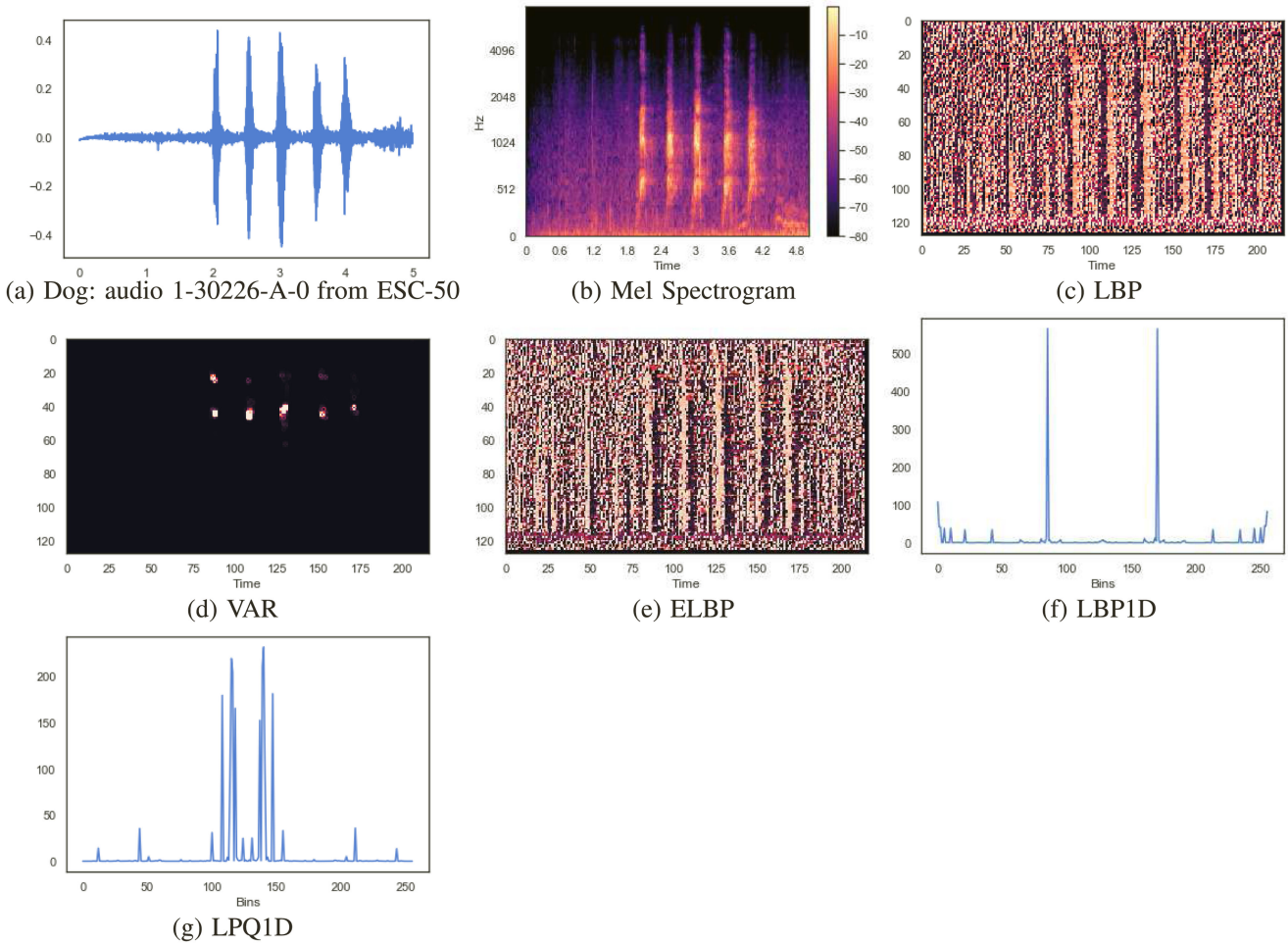
Given a center pixel c with gray level g_c, the LBP of the pixel is computed as follows:

$$LBP_c = \sum_{p-0}^{P-1} s(g_p - g_c)2^p \tag{1}$$

where

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{2}$$

The authors in [19] identified a rotation invariant version of the LBP, but it is not useful in the context of a sound spectrogram. The LBP operator, as defined in the previous equation, is not affected by any monotonic transformation of the grayscale. It is



(a) Dog: audio 1-30226-A-0 from ESC-50 (b) Mel Spectrogram (c) LBP

(d) VAR (e) ELBP (f) LBP1D

(g) LPQ1D

Fig. 3. Image Features.

a grayscale-invariant measure that unfortunately discards contrast. To detect the contrast, the authors in [19] proposed the VAR operator:

$$VAR_c = \frac{1}{P}\sum_{p-0}^{P-1}(g_p - \mu)^2 \quad \text{where} \quad \mu = \frac{1}{P}\sum_{p-0}^{P-1} g_p \tag{3}$$

LBP and VAR are two complementary measures, and their combination or joint distribution is a very powerful measure of local image texture. The results of LBP and VAR on ESC-50 audio are provided in Figs. 3(c) and 3(d), respectively.

2) Extended LBP: Because of the popularity and simplicity of implementation of LBP, a large number of methods have been developed over the years to improve its performance: *opponent color* LBP (OCLBP) [40], [41], *multiscale color* LBP (MS-CLBP) [25], *completed* LBP (CLBP) [42], and *discriminative completed* LBP (disCLBP) [22], to name the best-known methods. The extended LBP [43], one of the most robust among these methods, extends the LBP with four complementary descriptors: the central pixel intensity (CI), the neighbor intensity (NI), the radial difference (RD) and the angular difference (AD). Their formulation is as follows:

$$CI - LBP_c = s(g_c - \mu_I) \tag{4}$$

relative to μ_I, the mean of the image is I.

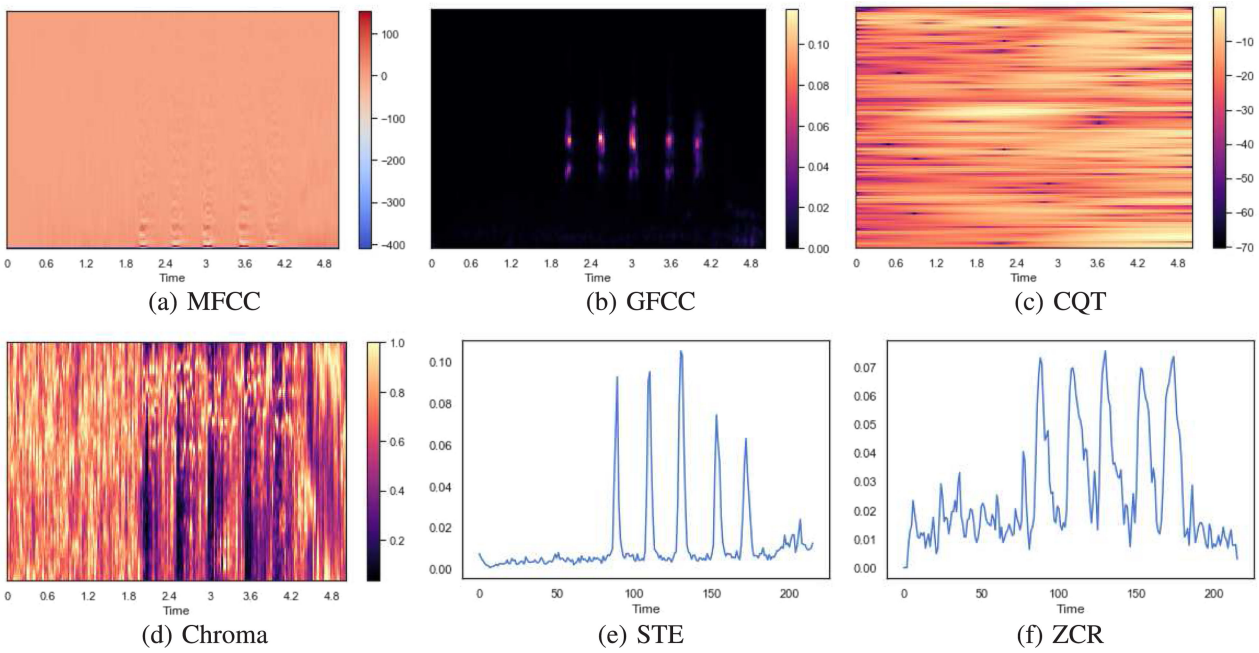
$$NI - LBP_c = \sum_{p-0}^{P-1} s(g_p - \mu)2^p \quad \text{where} \quad \mu = \frac{1}{P}\sum_{p-0}^{P-1} g_p \tag{5}$$

$$RD - LBP_c = \sum_{p-0}^{P-1} s(\Delta^{Rad})2^p \tag{6}$$

$$AD - LBP_c = \sum_{p-0}^{P-1} s(\Delta^{Ang})2^p \tag{7}$$

The AD-LBP is not used because it is too weak [43] and inadequate to provide a reliable and meaningful description of texture images. On the other hand, NI-RD and NI-RD-CI are strong descriptors. See the results of ELBP on ESC-50 audio in Fig. 3(e).

3) LBP-1D/LPQ-1D: Unlike an image signal where the neighborhood is a circle covering an angle of $360^c irc$, the audio signal has only two neighborhood angles: 0° and 180°. This allows us to define an audio texture in the time domain of a signal as a joint distribution of the pixel and its P (> 1 and even) neighbor points located in a horizontal radius R (Fig. 2). We can then apply the equation 1 to the 1D signal. For example, a



(a) MFCC (b) GFCC (c) CQT

(d) Chroma (e) STE (f) ZCR

Fig. 4. Audio Features.

neighborhood of $P = 2$ and $R = 1$ presents 4 categories of patterns that cover the distinct characteristics of a signal, namely, growth, decay, minimum and maximum.

The LPQ [44], the frequency version of the LBP, consists of quantizing, in an eight-dimensional space, the phases of the four low-frequency coefficients of the STFT. The authors [44] showed that the low-frequency phase components are ideally invariant to centrally symmetric blur and that the LPQ is more efficient than the LBP in the presence of noise. The LPQ1D simply involves applying the same reasoning to the 1D signal. See the results of the histograms of LBP1D and LPQ1D on ESC-50 audio in Figs. 3(f) and 3(g), respectively.

B. Audio Features

For comparison reasons and to achieve image and audio feature collaboration, we consider the following descriptors that are usually used for audio classification: MFCC, GFCC, STE, ZCR, CQT, and chromagram [7], [8], [30], [31].

1) MFCC: The mel-frequency cepstral coefficients were developed to resemble the human auditory system and have been successfully used in music modeling, speech recognition and audio classification [13], [27]. This work employs a short-term spectral-based feature computed from the mel spectrogram of the sound using a defined number of filters. In our experiment, we used 128 bands. See Fig. 4(a) for an example of MFCC.

2) GFCC: The gammatone filterbank attempts to approximate the human auditory system as the MFCC, using gamma distributions and sinusoidal tones [10], [11]. GFCC is known for its strong capability to represent impulsive signal classes such as transient sounds and offers complementary use with MFCC [8]. In our experiment, we used 128 bands for the GFCC; an example is displayed in Fig. 4(b).

3) CQT: The constant-Q transform (CQT) is a technique that transforms a time-domain signal $x(n)$ into the time-frequency domain so that the center frequencies of the frequency bins are geometrically spaced and their Q-factors are all equal [9], [30]. CQT exhibits good results in the audio classification task [30]. In our experiment, we used 128 bins; see Fig. 4(c) for an example.

4) Chromagram: The chroma is used to characterize the sound by decomposing the signal into a number of pitch class profiles [12]. It captures the harmonic and melodic characteristics of the music. In our experiment, we used 128 pitch class profiles; see Fig. 4(d) for an example.

5) STE: The short-time energy of the signal is the energy of the signal computed over a window of time. It enables a convenient representation of the amplitude variation over time. The STE also permits detection of the periodicity and the silence of a signal and characterizes the harmony of music [7]. It has the particularity of being able to be computed using time space or frequency space. See Fig. 4(e) for an example of STE.

6) ZCR: Unlike most of the previous audio features that are frequency-based, the zero-crossing rate is a time-based feature. It permits determination of whether successive samples have different signs. The rate at which zero-crossing occurs is a simple measure of the frequency content of a signal. It permits separation of a vocal sound from a nonvocal sound [7]. An example of ZCR is shown in Fig. 4(f).

C. Features Collaboration

The features collaboration technique consists of the exploitation of multiple features in order to construct a strong discriminator. Each feature is usually designed to characterize one of the many temporal and spectral content properties of an audio signal, including the pitch, frequency, energy, loudness, timbre, amplitude, etc. While the ZCR, for example, can easily discriminate between a vocal and a nonvocal sound, the chromagram can only capture the harmony and melody of music. It very

early became obvious for many studies to aggregate, concatenate or fuse multiple features in order to obtain a majority of distinguishable and complementary features that can classify all categories of environmental sounds. Muhammad *et al.* [45] aggregated MPEG-7 audio low-level descriptors together with conventional MFCC and demonstrated a significant improvement in the recognition performance of the proposed system over MFCC or full MPEG-7. In [8], the authors showed that combining MFCC with GFCC is very successful since they complement each other. In [27], PicZak concatenated ZCR and MFCC to successfully classify the ESC-50 dataset. Sharma *et al.* [31] achieved state-of-the-art performance by combining MFCC, GFCC, chromagram and CQT in a multichannel neural network coupled with an attention network. In our case, we will concatenate the image and audio features in order to cover more characteristics available in all categories of environmental sound, and we demonstrate in the next section that such a combination achieves better performance than using a sole type of feature, namely, image or audio.

IV. Experimental Results

A. Datasets

The ESC-50 dataset proposed by Piczak [27] in 2015 consists of 2000 labeled environmental sounds split equally among 50 classes (40 records per class). Each record has a duration of 5 seconds, with a sampling rate of 44.1 kHz. The 50 classes are divided into 5 categories (10 per category): animal sounds, natural soundscapes and water sounds, human (nonspeech) sounds, interior/domestic sounds, and exterior/urban sounds. The ESC-50 dataset is popularly used in the ESC task [15], [16], [27], [30]–[32], [34], [35]. For rapid testing, Piczak also proposed an ESC-10 subset of 400 records and 10 classes, with good separability between classes.

B. Setup

In our experiments, we segment the audio into multiple windows of 2048 samples, each with an overlapping 1024 samples. A wideband (2048) segmentation has been proved to offer minor advantages over narrowband (1024) [30] and permits having descriptors with reduced size in time/frequency dimensions (e.g., 216 for a signal of 5 seconds). A 3×3 neighborhood ($P = 8$ and $R = 1$) is used for the 2D image features. For LBP1D and LPQ1D, we use $P = 8$ and $R = 4$. In both 1D and 2D, a descriptor of 256 bins is obtained. The histograms of the image descriptors are computed for each window and then concatenated to form a final descriptor. The audio features are 128 bins, except for the STE and ZCR, which are both 1 bin. The implementation is performed using the librosa package[1] and the gammatone library.[2]

Each descriptor ends with a size of $nbins \times 216$ for an audio signal of 5 seconds, which is too high. One way to reduce this size is to use multidimensional scaling (MDS), principal component analysis (PCA), independent component analysis (ICA) or any other dimension reduction technique. The drawback of those methods is time consumption, so we prefer to compute a simple mean and a standard deviation along the time axis in order to reduce the size of the descriptor to $nbins \times 2$. We then process the descriptors using conventional machine learning algorithms SVM, random forest and kNN with a 5-fold cross-validation regime. The extraction of all of the descriptors requires 20 seconds per audio segment on an Intel(R) Core(TM) i5-7440HQ 2.80 GHz CPU without any GPU acceleration. Each algorithm requires only a few seconds to run the 5-fold validation and display an average accuracy result. The results are presented and commented upon in the next sections of the paper.

[1]librosa: v0.7.1 library by B. McFee *et al.*, doi: http://dx.doi.org/10.5281/zenodo.12714, Accessed: Aug. 5, 2015.

[2]gammatone: https://github.com/detly/gammatone.git

TABLE I
ESC-10: Results of Classification With One Feature

Features	kNN	RF	SVM
LBP1D	50.7	61.5	58
LPQ1D	53	66.4	65.5
LBP	**65.9**	**73.5**	**79.7**
VAR	52	42.2	56.7
ELBP	**65**	**73.5**	**76.5**
STE	49.7	44.2	46.7
ZCR	44.7	38.2	22.9
MFCC	**64**	**72.2**	**77**
GFCC	61.7	**77**	30.7
CQT	29	29.2	20.2
CHROMA	53.7	61.2	53.7

C. One Feature

In this section, we present the accuracy results of each descriptor on both the ESC-10 and ESC-50 datasets. Presented in bold are the best three accuracy values per algorithm.

The results on ESC-10 are presented in Table I. The best two results using the kNN method are obtained for LBP (65.9%) and ELBP (65%). The MFCC, the first audio descriptor to perform well, shows a result of 64% at the third position. Using random forest, the GFCC presents the best accuracy of 77%, followed by LBP and ELBP, both at 73.5%, and MFCC at 72.2%. With the SVM method, the LBP (79.7%) is in the first position, followed by the MFCC (77%) and the ELBP (76.5%). Note that the accuracy of 79.7% for the LBP is the best classification score obtained here and so far. The LBP remains robust, regardless of the machine learning method, and then performs better than the rest of the descriptors. The ELBP performance is not very far from that of LBP, which is normal since the implementations of these two operators are only slightly different. MFCC is next, followed by the GFCC, which has unfortunately vanished with the SVM. The 1D descriptors LBP1D and LPQ1D, which we introduced in Section III-A3, turn out to be weak descriptors. This result is somehow expected because they only exploit the 1D variation of the signal, while the other features are based on the spectrogram, which is a more complete time/frequency representation of the signal. However, LBP1D and LPQ1D did perform better than other classic audio features such as STE, ZCR and CQT. LPQ1D outperforms LBP1D since it is more robust to noise [44]. Although the result of VAR is weak, we will not devote much attention to it because it is normally used conjointly with the LBP. We will study its impact in the next section.

Table II shows the accuracy results on ESC-50. The accuracy is reduced by approximately 20% for all features because the

TABLE II
ESC-50: RESULTS OF CLASSIFICATION WITH ONE FEATURE

Features	kNN	RF	SVM
LBP1D	17.6	23.6	26.7
LPQ1D	17.6	27.2	29.5
LBP	**35.7**	**45.9**	**54.5**
VAR	13.3	13.4	15.4
ELBP	**33.5**	**43.7**	**53.7**
STE	10.1	9.2	9.2
ZCR	9.4	8.1	6.1
MFCC	22.6	**43.7**	**46.2**
GFCC	**26.3**	**42.6**	10.3
CQT	5.1	7.3	5.8
CHROMA	16.2	22.5	17.4

number of classes has increased. We can observe that the LBP and ELBP remain the best descriptors regardless of the machine learning algorithm. They are followed by MFCC, GFCC, LPQ1D, LBP1D and the rest of the descriptors. The overall best result of 54.5% on the ESC-50 dataset is obtained for a single LBP descriptor using SVM.

Those results outperform the top values of 72.7 on ESC-10 and 44.3 on ESC-50 that were originally obtained by Piczak [27] on his datasets using MFCC-ZCR with random forest and SVM.

D. Multiple Features

We showed in the previous section that LBP and ELBP features that are image descriptors offer better performance on the ESC-10 and ESC-50 datasets than the strongest commonly used classic audio descriptors such as MFCC or GFCC. Though the obtained results are very interesting, the LBP-based descriptors remain handcrafted features and are not able to fully extract all of the patterns that can fully characterize a signal. To correct this drawback, one of the solutions is to combine multiple features that will complementarily qualify the signal, which is called feature collaboration.

We start by combining the image descriptors together, then the audio descriptors together, and finally the image descriptors with the audio descriptors.

The results of the feature combination on ESC-10 are presented in Table III. The best accuracy is obtained for the kNN algorithm by the aggregation of the four strongest features LBP-ELBP-MFCC-GFCC (68%), followed by LBP-VAR-ELBP-MFCC-GFCC (67.7%), LBP-VAR (67.5%) and LBP-VAR-ELBP (67.5%). We can notice the natural complementarity between LBP (strong descriptor) and VAR (weak descriptor), as demonstrated by the authors in [19], which permits us to obtain a good result near the best obtained with 4 strong descriptors. The LBP-VAR even presents the best result of 84.9% on SVM. The combination of LBP-VAR with ELBP does not make a great difference since both algorithms are similar. However, associating the LBP-based image descriptors with the 1D descriptors and the audio descriptors, LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC, with random forest provides the global best result of **88.2%**.

The results on ESC-50 presented in Table IV show the same behaviors of the features as those observed on ESC-10. The combination of LBP, VAR and ELB provides the best accuracies with the kNN and SVM. The topmost result of **64.6%** is obtained using random forest and the mix of the three types of descriptors LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA. Note that this result is similar to that obtained using a convolutional network in [15].

TABLE III
ESC-10: RESULTS OF CLASSIFICATION WITH MULTIPLE FEATURES

Features	kNN	RF	SVM
LBP-VAR	**67.5**	77.2	**84.9**
LBP-VAR-ELBP	**67.5**	78.2	**84.2**
LBP-VAR-ELBP-LBP1D	52.5	80	72.7
LBP-VAR-ELBP-LBP1D-LPQ1D	57.2	82.4	68.7
MFCC-GFCC	64	83.5	77
MFCC-GFCC-CHROMA	64.2	84.5	77.5
MFCC-GFCC-CHROMA-STE	64.2	84.2	77.9
MFCC-GFCC-CHROMA-STE-ZCR	64.2	84.5	77.9
MFCC-GFCC-CHROMA-STE-ZCR-CQT	62.2	85	79.2
LBP-ELBP-MFCC-GFCC	**68**	86.9	**82.5**
LBP-VAR-ELBP-MFCC-GFCC	**67.7**	86.2	81.7
LBP-VAR-ELBP-LBP1D-MFCC-GFCC	64.5	87.2	81.2
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC	64.2	**88.2**	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA	64.2	86.9	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE	64.2	**87.5**	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR	64.2	**87.5**	76.7
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR-CQT	65	**87**	78.4

TABLE IV
ESC-50: RESULTS OF CLASSIFICATION WITH MULTIPLE FEATURES

Features	kNN	RF	SVM
LBP-VAR	**33.3**	47.4	**54.5**
LBP-VAR-ELBP	**34.3**	49.6	**55.8**
LBP-VAR-LBP1D-ELBP	19	54.1	44.6
LBP-VAR-LBP1D-LPQ1D-ELBP	20.9	55.6	39.6
MFCC-GFCC	22.6	51.9	46.2
MFCC-GFCC-CHROMA	22.6	53.2	46.6
MFCC-GFCC-CHROMA-STE	22.7	53.6	46.6
MFCC-GFCC-CHROMA-STE-ZCR	22.7	52.9	46.6
MFCC-GFCC-CHROMA-STE-ZCR-CQT	20.7	54.1	43.6
LBP-ELBP-MFCC-GFCC	23.4	62.3	54.3
LBP-VAR-ELBP-MFCC-GFCC	23.5	61.4	**55**
LBP-VAR-ELBP-LBP1D-MFCC-GFCC	25	63.4	51.6
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC	24.7	63.2	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA	24.7	**64.6**	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE	24.7	**64.1**	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR	24.7	63.4	46.8
LBP-VAR-ELBP-LBP1D-LPQ1D-MFCC-GFCC-CHROMA-STE-ZCR-CQT	**25.6**	**64.3**	47.2

E. Analysis

Note that on both datasets, associating only image descriptors or audio descriptors does not yield the best accuracy. On ESC-50, for example, the accuracy values achieved with such combinations are all under 60%. However, as soon as we mix an image descriptor with an audio descriptor, breaking the 60% mark becomes possible. The audio and image descriptors can therefore be considered complementary.

We also observe that the random forest algorithm is more sensitive to the mixture of features. This is because SVM and kNN

TABLE V
BEST PROPOSED MODEL SCORE AND COMPARISON WITH THE STATE OF THE ART

Algo	ESC-10	ESC-50
Human [27]	95.7	81.3
Attention Network [35]	94.2	84
EnvNet [32]	86.8	66.4
EnvNet2 [38]	88.8	81.6
PiczakCNN [15]	90.2	64.5
AlexNet [16]	86	65
GoogleNet [16]	86	73
Piczak kNN [27]	66.7	32.2
Piczak RF [27]	72.7	44.3
Piczak SVM [27]	67.5	39.6
CNN+CQT+CWT+MEL [30]	-	56.4
SoundNet [34]	92.2	74.2
Our best Method	**88.5**	**64.5**

TABLE VI
SPECTROGRAM STREAM OF THE MULTISTREAM WITH ATTENTION NETWORK [35]

Layer	Filter Size	No. of Filters
Conv	3x3	128
Conv	3x3	128
MaxPooling	4x3	-
Conv	3x3	128
Conv	3x3	256
MaxPooling	4x4	-
Conv	3x3	256
Conv	3x3	512
MaxPooling	2x2	-
Conv	3x3	512
MaxPooling	2x2	-
Conv	3x3	1024
MaxPooling	2x2	-
Conv	3x3	1024
Conv	3x3	2048
MaxPooling	2x2	-
Dense	-	4096
Dense	-	4096

lose their power when the feature vector size is increased. The weak descriptors such as STE, ZCR and CQT perform poorly. They exert minor impacts, or sometimes negative impacts, when combined with the other descriptors.

Our best model performs well compared with the state of the art, as shown in Table V. It outperforms any model based on conventional machine learning and offers similar performance to some deep learning methods. Most of the research in the ESC field is now oriented to deep learning methods, but our model represents a good alternative in the presence of limited computing power or a lack of data. It requires only a few hours to extract the descriptors and run a model, while a few days or weeks would be required for a deep learning method.

To prove this fact, we train on a low-end GPU the stream that processes the spectrogram in the three-stream network available in [35], which is the leader in Table V. Such a network, presented in Table VI, is very deep and is composed of 18 levels of 2D convolution, max pooling and dense layers for a total of 88,143,882 trainable parameters. The input size is 512x384. Although it is not visible in Table VI, each convolution layer is followed by a batch normalization and a ReLU. We use a data augmentation technique [39] (time stretch, time shift, noise, pitch shift) on ESC-10 that increases the number of samples from 400 to 2,000. We train the network on the ES-10 dataset on an NVIDIA GeForce 930MX with 2 GB RAM, a low-end GPU, which has specifications that are comparable with low-power AI systems available on the market: NVIDIA Jetson Nano (Quad Cortex A57 @ 1.43 GHz, 4 GB RAM) and Google Edge TPU (Quad Cortex-A53, Cortex-M4F @ 2 GB RAM). The implementation is performed with the TensorFlow [46] library with a minimal batch size of 1, but the *resource exhausted error* is triggered because of the limited GPU memory of 2 GB. Since the test is only for performance measurement and not accuracy, we divide by 8 the number of filters at each level in order to obtain a reduced network of 2,700,922 trainable parameters. It takes 16 minutes to train the 5-fold ESC-10 for 1 epoch, *i.e.*, 26.6 hours for 100 epochs. In the paper [35], three networks were used, with an additional fourth network for the attention, and were trained for more than 100 epochs. Many weeks will be required to train such a system on a low-end GPU with sufficient memory, and months to train ESC-50. If we eliminate the data preparation time in both cases, this time is enormous compared with the 20 seconds required to run SVM or random forest on the same dataset. This fact shows that although low-power AI systems are available for the lowest price on the market, training a network that is slightly deep is not affordable for everyone. Those GPUs are not designed for training, but rather for inference and supporting limited transfer learning.

V. CONCLUSION

In this paper, we presented a new ESC method that exploits LBP, a 2D texture classification descriptor. The LBP method was applied to the signal in one dimension as well as on the sound texture represented by the spectrogram. The results showed that LBP features outperform the audio descriptors and are more efficient on the two datasets. We showed that the combination of the audio descriptors with the LBP achieves state-of-the-art results using a simple machine learning classifier such as SVM or random forest. It also performs well compared with some CNN-based methods. This approach is faster than deep learning and represents a good alternative when there is data scarcity or minimal computing power. Our method has many advantages but is not the leader of the state-of-the-art methods. It can be improved by using CNN with a multichannel descriptor consisting of a mixture of LBP and audio features. This improvement represents the topic of our future research.

REFERENCES

[1] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, "Sound-event classification using robust texture features for robot hearing," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 447–458, Mar. 2017.
[2] M. J. Kim and H. Kim, "Audio-based objectionable content detection using discriminative transforms of time-frequency dynamics," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1390–1400, Oct. 2012.
[3] M. A. Sehili, D. Istrate, B. Dorizzi, and J. Boudy, "Daily sound recognition using a combination of gmm and svm for home automation," in *Proc. 20th Eur. Signal Process. Conf.*, Aug. 2012, pp. 1673–1677.
[4] J.-C. Wang, C. Lin, B.-W. Chen, and M.-K. Tsai, "Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation," *IEEE Trans. Automat. Sci. Eng.*, vol. 11, no. 2, pp. 607–613, Apr. 2014.
[5] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," *Sensors*, vol. 18, no. 6, Jun. 2018.

[6] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with timefrequency audio features," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.
[7] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, May 2001.
[8] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–9.
[9] C. Schrkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *Proc. 7th Sound Music Comput.*, 2010.
[10] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Comput. Tech. Rep.*, vol. 35, 1993.
[11] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.
[12] R. N. Shepard, "Circularity in judgments of relative pitch," *J. Acoustical Soc. Amer.*, vol. 36, no. 12, pp. 2346–2353, 1964.
[13] B. Logan, "Mel frequency cepstral coefficients for music modeling," *ISMIR*, vol. 270, p. 111, Oct. 2000.
[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[15] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process.*. ScienceDirect, 2015, pp. 1–6.
[16] V. Boddapatia, A. Petefb, J. Rasmussonb, and L. Lundberga, "Classifying environmental sounds using image recognition networks," in *Proc. Int. Conf. Knowl. Based Intell. Inf. Eng. Syst.*, Sep. 2017, pp. 6–8.
[17] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," *Proc. Int. Conf. Pattern Recognit.*, 1994, vol. 1, pp. 582–585.
[18] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, p. 5159, 1996.
[19] T. O. M. Pietikäinen and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[20] P. Král, A. Vrba, and L. Lenc, "Enhanced local binary patterns for automatic face recognition," in *Proc. Int. Conf. Artif. Intell. Soft Comput.*, vol. 11509. Lecture Notes in Computer Science, 2019, pp. 27–36.
[21] D. K. Iakovidis, E. G. Keramidas, and D. Maroulis, "Fuzzy local binary patterns for ultrasound texture characterization," *Image Anal. Recognit. (Lecture Notes in Comput. Sci.)*, vol. 5112, pp. 750–759, 2008.
[22] Y. Guo, G. Zhao, and M. Pietikäinen, "Discriminative features for texture description," *Pattern Recognit.*, vol. 45, no. 10, pp. 3834–3843, Oct. 2012.
[23] Z. Lei and S. Z. Li, "Fast multi-scale local phase quantization histogram for face recognition," *Proc. Int. Conf. Pattern Recognit.*, 2012, vol. 33, no. 13, pp. 1761–1767.
[24] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," *Proc. Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[25] A. Joshi and A. K. Gangwar, "Color local phase quantization (CLPQ)- a new face representation approach using color texture cues," in *Proc. Int. Conf. Biometrics*, May 2015, pp. 177–184.
[26] T. Kobayashi and J. Ye, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 3052–3056.
[27] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. 23rd Annu. ACM Conf. Multimedia* . ACM Press, Oct. 2015, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390
[28] S. Chandrakala and S. Jayalakshmi, "Generative model-driven representation learning in a hybrid framework for environmental audio scene and sound event recognition," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 3–14, Jan. 2020.
[29] J.-C. Wang, J.-F. Wang, K. W. He, and C.-S. Hsu, "Environmental sound classification using hybrid svm/knn classifier and mpeg-7 audio low-level descriptor," in *Proc. IEEE Int. Joint Conf. Neural Netw. Proc.*, Jul. 2006, pp. 1731–1735.
[30] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," Jun. 2017, *arXiv:1706.07156*.
[31] J. Sharma, O.-C. Granmo, and M. Goodwin, "Environment sound classification using multiple feature channels and deep convolutional neural networks," in *Proc. Interspeech 2020*, Aug. 2019, pp. 1186–1190.
[32] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2721–2725.
[33] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1 d convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019.
[34] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. 30th Conf. Neural Inf. Process. Syst.*, 2016.
[35] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," in *Proc. Interspeech*, 2019, pp. 3604–3608.
[36] H. Sailor, D. Agrawal, and H. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification)," in *Proc. INTERSPEECH*, Aug. 2017.
[37] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 421–425.
[38] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," in *Proc. Int. Conf. Learn. Representations*, Feb. 2018.
[39] 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.
[40] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, "Computer vision using local binary patterns," *Springer*, 2011.
[41] T. Mäenpää, and M. Pieti;inen, "Computer vision using local binary patternsclassification with color and texture: Jointly or separately," *Pattern Recognit.*, vol. 37, pp. 1629–1640, 2004.
[42] Z. Guo, L. Zhang, and D. Zhangr, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010.
[43] L. Liu, L. Zhao, Y. Long, G. Kuang, and P. Fieguth, "Extended local binary patterns for texture classification," *Image Vis. Comput.*, vol. 30, no. 2, pp. 86–99, Feb. 2012.
[44] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Proc. Int. Conf. Image Signal Process.*, 2008, pp. 236–243.
[45] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and mel-frequency cepstral coefficients," in *Proc. Digit. Telecommun. 5th Int. Conf.*, Jun. 2010, pp. 11–16.
[46] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Des. Implementation (OSDI 16), USENIX Assoc.*, 2016, pp. 265–283.

Ohini Kafui Toffa received the M.Sc. degree in computer science from the Department of Computer Science, University of Sherbrooke, Sherbrooke, QC, Canada, in 2009. He is currently working toward the Ph.D. degree in vision laboratory with the University of Montreal (DIRO), and a Supervisor of software development with Intrado Montreal. His research interests include multimodal methods, applied mathematics and machine learning, in sound, image, and video classification.

Max Mignotte received the DEA (Postgraduate degree) in digital signal, image and speech processing from INPG University, Grenoble, France, in 1993, and the Ph.D. degree in electronics and computer engineering from the University of Bretagne Occidentale (UBO), and the Digital Signal Laboratory (GTS) of the French Naval Academy, France, in 1998. He was an INRIA Postdoctoral Fellow with the University of Montreal (DIRO), (Quebec), Canada. from 1998 to 1999. He is currently a Professor with DIRO, Computer Vision & Geometric Modeling Lab, University of Montreal. He is also a member of LIO (Laboratoire de recherche en imagerie et orthopedie, Centre de recherche du CHUM, Hopital Notre-Dame) and a Researcher with CHUM. His current research interests include statistical methods, bayesian inference and energy-based models (especially encoding non-local pairwise pixel interactions) for solving diverse large-scale high-dimensional ill-posed inverse problems in imaging.