

Histogram of gradients of Time-Frequency Representations for Audio scene detection

Alain Rakotomamonjy, Gilles Gasso

▶ **To cite this version:**

Alain Rakotomamonjy, Gilles Gasso. Histogram of gradients of Time-Frequency Representations for Audio scene detection. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2015, 23 (1), pp.142-153. 10.1109/TASLP.2014.2375575 . hal-00951990v6

HAL Id: hal-00951990
https://hal.science/hal-00951990v6

Submitted on 3 Aug 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Histogram of Gradients of Time-Frequency Representations for Audio Scene Detection

A. Rakotomamonjy, G. Gasso

***Abstract*—This paper addresses the problem of audio scenes classification and contributes to the state of the art by proposing a novel feature. We build this feature by considering histogram of gradients (HOG) of an audio scene time-frequency representation. Contrarily to classical audio features like MFCC, we make the hypothesis that histograms of gradients are able to encode some relevant informations in a time-frequency representation: namely, the local direction of variation (in time and frequency) of the signal spectral power. In addition, in order to gain more invariance and robustness, histograms of gradients are locally pooled. We have evaluated the relevance of the novel feature by comparing its performances with state-of-the-art competitors, on several datasets, including a novel one that we provide, as part of our contribution. This dataset, that we make publicly available, involves 19 classes and contains about 1500 minutes of audio scene recordings. We thus believe that it may be the next standard dataset for evaluating audio scene classification algorithms. Our comparison results clearly show that the HOG-based features outperform its competitors.**

***Index Terms*—Histogram of gradients; Time-Frequency Representation; audio scene; MFCC; support vector machines**

I. Introduction

The problem of recognizing acoustic environments is known as the problem of audio scene classification and it is one of the most difficult task in the general context of computational auditory scene analysis (CASA) [1]. This classification task is of primary importance in the domain of machine listening since it is strongly related to the context in which the acquisition device capturing the audio scene lives. Typically, in order to get some context awareness, a machine, say a smart-phone or any mobile electronic device, should be able to predict the environment in which it currently resides. The main goal is to help the machine adapting itself to the context of the user (for instance by automatically turning off the ring tone in some situations). Such awareness can be brought through vision or audio scene analysis. While most of the efforts have focused on vision, there is now a growing interest of environment recognition based on audio modality.

Audio scene classification is a very complex problem since a recording related to a given location can be potentially composed of a very large amount of single sound events while only few of these events provide some information on the scene of the recording. More specifically, an audio scene is associated to a recording taken at a given location and this location is expected to generate some acoustic events that make it distinguishable from other audio scenes. These discriminative acoustic events may be produced by different phenomena and they may have a very large variability.

Hence, the recent works on audio scene classification have devoted much efforts on designing methods and algorithms for automatically extracting audio features that capture the specificities of these events. The natural hope is that the designed features are still able to capture the discriminative power of a given audio event.

For instance, following its success in speech recognition, one of the most prominent features that has been considered for audio scene recognition are mel-frequency cepstral coefficients (MFCC) [2], [3], [4], [5]. These features are typically used in conjunction with different machine learning techniques in order to capture the variations that help in discriminating scenes. For instance, [2] consider a Gaussian Mixture Model for estimating the distribution of the MFCC coefficients, while [6] have proposed a sparse feature learning approach for capturing relevant MFCC coefficients.

In addition to MFCC, several kinds of features have also been evaluated for solving this problem of audio scene recognition. [7] proposed an ensemble of time-frequency features obtained from a matching pursuit decomposition of the audio signal. Recently, [8] have considered a large set of features, including spectral, energy-based and voicing-related features. Another family of relevant features can be obtained from MFCC by considering recursive quantitative analyzing (RQA) as introduced by [9]. RQA has the advantage to allow the

This work has been partly supported by the French ANR (12-BS02-004).

AR is with LITIS EA 4108, Université de Rouen, France. alain.rakoto@insa-rouen.fr
GG is with LITIS EA 4108, INSA de Rouen, France. gilles.gasso@insa-rouen.fr

analysis of recurrent behaviour in the MFCC coefficients over time. According to the recent D-Case challenge on audio scene recognition [10], combining MFCC features with RQA features extracted from MFCC yield to an highly efficient set of features.

Another trend aims at building higher-level features from the time-frequency representations of the audio scene. In this context, [11] have investigated methods for automatically extracting spatio-temporal patches that are discriminative of the audio scene. Typically, these patches are obtained through a non-negative matrix factorization of a time-frequency representation. [12] have followed similar ideas but instead of considering NMF they employed a probabilistic model denoted as probabilistic latent component analysis. Likewise, other works propose features, like texture-based features that are directly computed from time-frequency representations [13], [14].

In this paper, we follow this trend and propose a novel feature for automatic recognition of audio scene. The main originality of the proposed feature is the use of histogram of gradients (HOG) on time-frequency representations. These HOG features have been genuinely introduced for human detection in images [15] but we strongly believe that their properties make them highly valuable for extracting relevant features based on time-frequency representation (TFR). Indeed, while MFCC can also be considered as features extracted from a time-frequency (TF) representation, they essentially capture non-linear information on the power spectrum of the signal. Instead, histogram of gradients of a TF representation provides information on the spectral power direction of variation. For instance, if an audio scene has been obtained in a bus that it is accelerating or decelerating, we expect the *chirp* effect present in the TFR to be captured and better discriminated by the histogram of gradients representation, than by MFCC. This property will be empirically illustrated in the sequel and it provides rationale that for audio scene recognition the novel feature we present is strongly relevant.

Algorithms for audio scene recognition have to be validated and evaluated on some datasets. In order to make comparisons of different designs of features including signal processing set-ups or different learning techniques possible, these datasets should be publicly available. The recent D-Case challenge [10] is an excellent initiative of this kind although its number of examples is limited (100 for the publicly available examples). Hence, another contribution we present is a new dataset for audio scene recognition. It is based on about 1500 minutes of recording on different locations (up to 19). This dataset is publicly available and we expect that it will become one of the standard benchmarks for audio scene recognition.

The paper is organized as follows: we first describe the pipeline we propose for extracting our novel HOG-based feature. Then, as one of our contribution is also to introduce a novel benchmark dataset, we carefully detail all the datasets we considered for evaluating our feature and its competitors as well as the experimental protocol we employed for the comparisons. Extensive experimental analyses have been carried out and they show that our HOG-based feature achieves state-of-the-art performances on all datasets. As we advocate result reproducibility, all the codes used for this work will be made publicly available on the author's website.

II. From signal to histogram of gradients features

This section describes the feature extraction pipeline we propose for analyzing audio scene signals. We first provide the big picture before detailing each part of the flowchart.

A. The global feature extraction scheme

The features we propose for recognizing audio scenes are based on some specific information extracted from a time-frequency representation (TFR) of the signal. After the TFR image has been computed, it is processed so as to attenuate some spurious noises that may hinder relevant information related to high-energy time-frequency structures. Afterwards, the resulting processed time-frequency representation image is used as input of our histogram of gradients feature extraction. In a nutshell, the idea of histogram of gradients is to locally analyze the direction of energy's variation in the time-frequency representation. As detailed in the sequel, the local HOG informations over the whole TF image are combined in order to generate the final feature vector. The dimension of this vector depends on the number of bin in the (local) histogram and on how all the local histograms are pooled together in order to form the final feature vector.

The block diagram of this feature extraction scheme is illustrated in Figure 1. Each block of this diagram is discussed in the next paragraphs.

B. From signals to TFR images

Because of their non-stationary nature, sounds are typically represented on a short-time power frequency representation, which idea is to capture the power spectrum of the signal on a varying short local window. A large part of the literature on sound recognitizion problems use such time-frequency representations of sound.

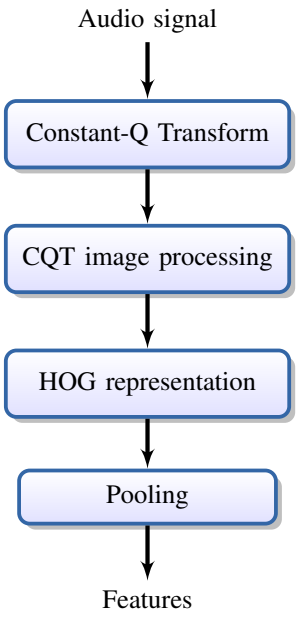


Fig. 1. Block Diagram of the HOG-based feature extraction

Depending on the task at hand, they either consider wavelet-based [16], [17], MFCC-based [18] or short-time Fourier-based representations [19]. In this work, we do not depart from this widely adopted framework and choose to represent the signal according to a constant-Q transform [20]. Contrarily to a short-time Fourier transform, this transform provides a frequency analysis on a log-scale which makes it more adapted to sound and music representations [20].

Once this TFR has been computed, we now have an image that can be processed as such. In order to obtain a processing system independent of the signal length and sampling frequency, as well as the CQT parameters, we have chosen to resize all TFRs to a 512×512 image. This resizing is performed on the CQT matrix by means of a bicubic interpolation. Hence, the image we obtain is not exactly equivalent to a CQT with a total of 512 frequency bins but it preserves the time-frequency structures of the audio scene as one can see in Figure 2.

Then, depending on the TFR images, some image processing tools can be used so as to enhance relevant time-frequency structures. In our work, because we have few prior knowledge on the signal noise, we have just applied a mean filtering so as to smooth the TFR image. Our goal with this smoothing is to reduce strong local variations in the image that will tend to produce high gradients, which may be not relevant for audio scene recognition. The size of the average kernel for mean filtering can be considered then as an hyperparameter of the feature extraction scheme and its influence will be investigated in the experimental analysis.

C. Histogram of gradients

Histogram of gradients have been originally introduced by [15] for human detection in images. Our main objective in this feature extraction stage is to capture the shape of some time-frequency structures in the hope that such structures are relevant for characterizing an audio scene. From works in computer vision [15], [21], [22], it is now well acknowledged that local shape information can be described through gradient intensity and orientations. Histograms of gradients basically provide information about the occurrence of gradient orientations in a localized region of the images. Hence, they are able to characterize shapes in that regions.

Two main approaches have been proposed for computing HOG in images [15], [21] and they are both based on the following steps:

1) compute the gradient of the TF image
2) compute angles of all pixel gradients
3) split images into non-overlapping cells
4) count the occurrence of gradient orientations in a given cell
5) eventually normalize each cell histogram according to histogram norm of neighboring cells.

Variants on this theme are essentially based on whether the gradient orientations are bidirectional or not, whether the magnitude of the gradient is taken into account in the counting and on how normalization factors are computed within block of neighboring cells. For this work, we have used the implementation in the *VLFeat* toolbox [23]. *VLFeat* is an open-source computer vision toolbox that includes the major functionalities of the most popular computer vision and pattern recognition algorithms. In particular, it implements several histogram-based feature extraction algorithms including SIFT [24] and HOG. For more details, we refer the reader to [15] and [21] and to the *VLFeat* Hog tutorial[1].

As an illustration, we depict in Figure 2 the 512×512 mean filtered image of a linear chirp's CQT transform, as well as the resulting histogram of gradients we obtain for each 32×32 cell with 8 gradient orientations. From the right panel, we remark that the HOG representation properly captures the directions of power spectrum's variation along the high-energy chirp signal. However, we can also note that several spurious cells depict non-zero histogram of gradients (top and right bottom parts

[1] http://www.vlfeat.org/overview/hog.html

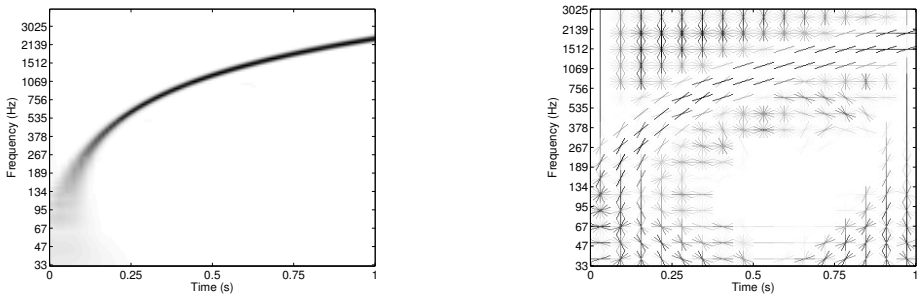


Fig. 2. Example of HOG for a toy linear chirp. (left) 512×512 image of the CQ transform of the signal. (right) histogram of gradients representation of the signal. Note that for a sake of interpretation, each cell in the plots represents the occurrence of edge orientation in the cell and the darker the orientation is, the more present the orientation is. We thus note that along the chirp, the HOG representation correctly captures the direction of energy variation.

of the image). They are essentially due to presence of small variations of gradient in low-energy time-frequency structure in the CQ transform, inducing non-zero gradients. However, these noisy cells can be easily recognized as having an almost flat histogram, denoting thus the presence of multiple orientations of gradient in the cells. Note that in our feature extraction scheme, we have not considered any pre-processing and post-processing strategies for handling these spurious cells, they are taken into account as they are into the HOG features.

After having computed the histogram of gradients in the images, we are left with a representation composed of histograms in all cells. If we concatenate all these histograms for yielding the final feature vector, we obtain a vector whose dimension is large (number of cells $\times$ number of orientations in the histogram). In the example in Figure 2, cells are sized 32×32 pixels, this results in vector of dimension $16^2 \times 8 = 2048$, 16^2 being the total number of cells. Of course, this dimensionality may further increase if we choose to reduce cell's size or increase the number of orientations in the histogram computation. Depending on the number of audio scene examples, it thus may be beneficial to reduce the dimensionality of the problem for instance by pooling the histograms of gradients.

D. Time-Frequency Histogram Pooling

Pooling consists in combining the responses of a feature extraction algorithm computed at nearby locations. The underlying idea is to summarize local features into another feature (of lower dimensionality) that is expected to keep relevant information over the neighborhood. The pooling helps in getting a more robust information. This technique is a step commonly considered with success in modern visual recognition algorithms [25]. In our case, pooling histograms over neighboring cells aims at building new histograms that capture information on time-frequency structures which may be larger than a cell or that have been slightly translated in time or frequency. In this work, we will investigate several forms of time-frequency region pooling (see Figure 3), while the pooling operation will be kept fixed as an averaging operator. We will consider the following poolings:

- Marginalized pooling over time: for this pooling, we average all histograms along the time axis of the TFR representation. This results in a feature vector which has lost all temporal information.
- Marginalized pooling over frequency: in this case, the averaging is performed over the frequency axis. Hence, all frequency informations of the HOG are now merged into a single one.
- Block-size pooling: pooling is performed on nearby cells with the size of the neighorhood being user-defined.

The vector resulting from the concatenation of the pooled histograms forms now the feature vector that will be used for learning the audio scene classifier.

E. Discussions

Now that we have explained how the HOG on time-frequency representation feature is obtained, we want to discuss some properties of these HOG features and their advantages over features like MFCC for audio scene characterization.

Our initial objective was to design features that is able to characterize some time-frequency structures that occur in a time-frequency representation. By construction, since we bin the orientations when counting a given gradient, the histogram of gradients is invariant to rotation if this rotation is smaller than the bin size. In our case, rotation would correspond to a small rotation of a time-frequency structure leading then to a change

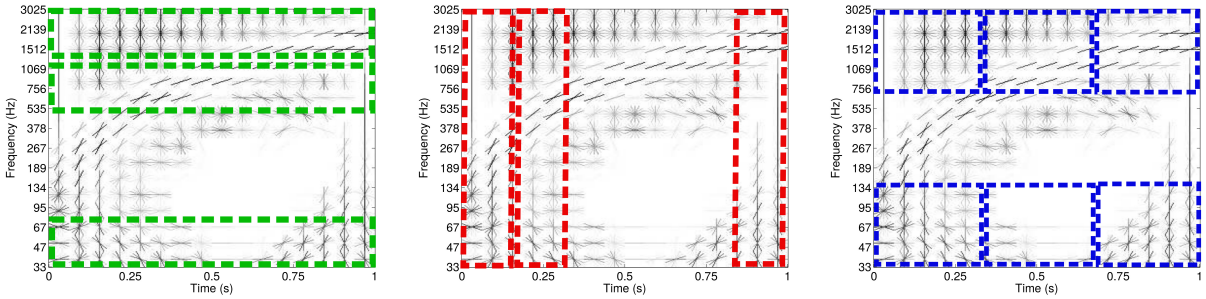


Fig. 3. Illustrating the different types of pooling we investigate based on the HOG issued from Figure 2. (Left) time pooling. (Middle) frequency pooling. (right) block-sized pooling. The green, red and blue boxes provide an example on the regions on which local histograms are averaged. The x-axis and y-axis respectively denote the time and frequency axes. Best viewed in color.

of orientation of its gradients. Such a situation may occur for instance in audio scenes capturing a moving object, like a bus or a car. Indeed, variations of a bus's acceleration induce variations of steepness in the time-frequency structure related to the sound of that bus. Owing to the binning of the gradient orientation, the HOG feature will be invariant to these variations. Furthermore, as we build an histogram from a cell of pixels and then average them over a larger region, our pooled histogram of gradient is invariant to translation over that region of pooling.

Compared to classical features like MFCC used for audio applications, HOG-based features present several benefits. For instance, they are, by construction, invariant to small time and frequency translations. But most interestingly, they bring information that are not provided by other power-spectrum based features, namely local direction of variation of power spectrum. As an illustration of this point, we will compare the features obtained, by MFCC and the HOG-based approach on two linear chirps, one with increasing frequency and the other one with a decreasing frequency, but both covering the same frequency range. Our experimental results will show that bag of sole MFCC will fail in fully capturing the discriminative information brought by these signals at the contrary of the features we propose.

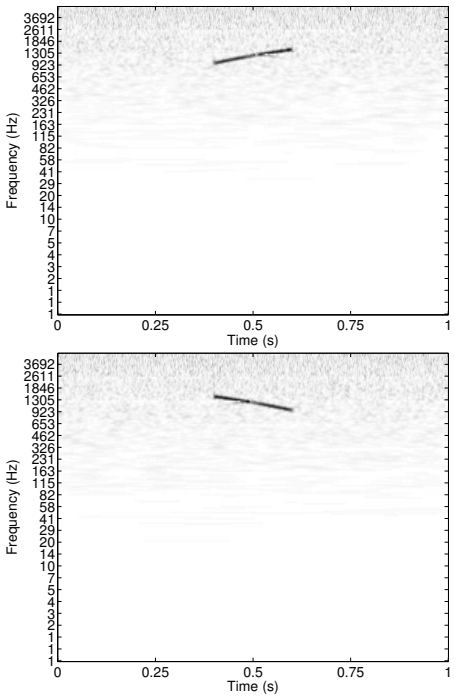


Fig. 4. Examples of CQT representation of the two localized linear chirps used in the toy dataset. These images are obtained after a 512×512 resizing and a 15×15 average filtering.

III. Data and Classifiers

We provide in this section some details about the datasets we have considered for evaluating the feature we propose. Description of the classifier we used as well as the experimental protocol are also given.

A. Toy dataset

For evaluating our features, we have created a toy problem which highlights the ability or the failure of studied features (including HOG and competitors) in capturing power spectrum's direction of variation. As such, we have created a binary classification problem where signals from each class are composed of a localized linear chirp, respectively of increasing and decreasing frequency, defined as

$$s(t) = \Pi_{[t_1,t_2]}(t) \cos\left(2\pi(at+b)t\right) + n(t)$$

with $t \in [0,1]$ and n is a centered Gaussian noise of standard deviation 0.4, $a = 1200$, $b = 0$ for the class

$y = +1$, $a = -1200$, $b = 2400$ for the class $y = -1$ and $\Pi_{[t_1,t_2]}(t)$ a function which value is 1 when $t_1 \leq t \leq t_2$ and 0 otherwise. We have set $t_1 = 0.4$ and $t_2 = 0.6$. Figure 4 depicts the CQ transform of representative samples of both classes. One can notice the localized spectral contents of the chirps.

B. D-case challenge dataset

For the purpose of a challenge, a dataset providing environmental sound recordings has been recently released by [10]. Each example in the dataset consists of a 30-second audio scene, which has been captured at one of the 10 following locations : *bus, busy street, office, open air market, park, quiet street, restaurant, supermarket, tube, tubestation*. Recording has occurred at a rate of 44.1 kHz and the number of examples available is 100 with 10 examples per class. Note that, the challenge's organisers have only made available the development dataset[2]

C. East Anglia (EA) dataset

This dataset[3] has been collected in the early 2000 by Ma et al. [26] at the East Anglia University. It provides environmental sounds coming from 10 different locations: *bar, beach, bus, car, football match, laundrette, lecture, office, railstation and street.* The length of each recording is 4 minutes and it has been recorded at a frequency of 22100 Hz. Similarly to the *D-case* dataset, we have split the recording in 30-second audio scene examples. Hence, we have only 8 examples per class for this dataset.

D. Litis Rouen dataset

This dataset we make publicly available[4] goes beyond the above ones in terms of volume and number of locations. Recordings have been performed using a Galaxy S3 smartphone equipped with Android by means of the *Hi-Q MP3 recorder* application. While such an equipment may be considered as poor, we believe that the resulting recordings would be similar to those obtained for real applications where cheap and ubiquitous microphones are more likely to be used. The sampling frequency we have used is 44100 Hz and the recording is saved as a MP3 file with a bitrate of 64 kbps. When transformed into raw audio signals, they have been downsampled to 22050 Hz. Overall, about 1500 minutes of audio scene have been recorded. They took place from December 2012 to June 2014. For a given class, recordings occured at several days of that period. The dataset is composed of 19 classes and audio scenes forming a given class have been recorded at different locations. Note that in order to reduce temporal dependencies in our dataset, recordings usually last 1 minute but in some locations, their durations can reach up to 10 minutes. Again in order to be consistent with the *D-case* challenge, each example is composed of a 30-second audio scene. The 30-sec examples have been obtained by splitting a given signal into 30-second segments without overlapping. A summary of the dataset is given in Table I and Figure 5 presents some samples of CQT for 2 different audio scenes. The plots in this figure show typical characteristics of an audio scene of the class. For instance, in the *bus*'s CQT, we can note the low-frequency line related to the bus's acceleration and deceleration. In the *kid game hall* scene, we see some high-frequency structures induced by kid screams.

TABLE I
SUMMARY OF THE LITIS ROUEN AUDIO SCENE DATASET

Classes	#examples
plane	23
busy street	143
bus	192
cafe	120
car	243
train station hall	269
kid game hall	145
market	276
metro-paris	139
metro-rouen	249
biliard pool hall	155
quiet street	90
student hall	88
restaurant	133
pedestrian street	122
shop	203
train	164
high-speed train	147
tube station	125
	3026

E. Competing features, classifier and protocols

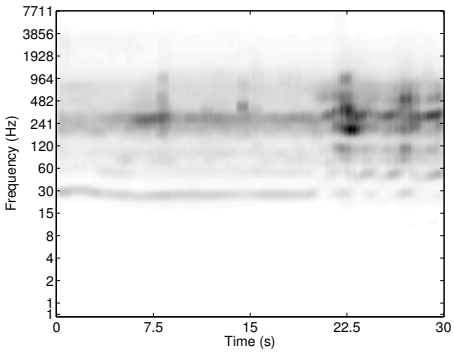
In order to evaluate how well the HOG-based feature we propose performs, we have compared its performance to those of other features. As a sake of comparison, we have considered the following ones:

- Bag of MFCC: these features are obtained by computing the MFCC features on windowed part of the signals and then in concatenating them all [2]. The setting for the MFCC computations are typical. We have extracted MFCC features from

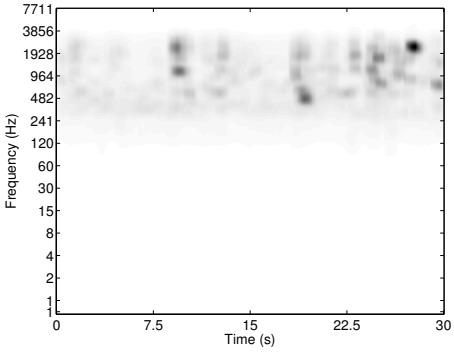
[2]http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/29

[3]available at http://lemur.cmp.uea.ac.uk/Research/noise_db/

[4]available on the first author's website at : https://sites.google.com/site/alainrakotomamonjy/home/audio-scene



(bus)



(kid game hall)

Fig. 5. Examples of CQT of audio signals from 2 different scenes of the Rouen's dataset after image mean filtering with an average kernel of size 15×15.

each audio scene by means of sliding windows of size 25 ms with hops of 10 ms. For each window, 13 cepstra over 40 bands have been computed. The toolbox we have employed is the *rastamat* one with the dithering option on [27]. The lower and upper frequencies of the spectral analysis are respectively set to 1 and 10000Hz. For the toy dataset, the upper frequency is set to the Nyquist frequency. For obtaining the final features, we average the obtained MFCC over batch of 40 windows with overlap of 20 windows and concatenated together all the MFCC averages and standard deviations.

- Bag of MFCC-D-DD: these features are the same as above but in addition to the MFCC averages and standard deviations over a batch of 40 windows, we also concatenate to the feature vector, the average of first-order and second-order differences of the MFCC over these windows. These features should provide a better description of the dynamic behaviour of the signal.
- Frame/majority-based MFCC-D-DD : For this approach, we use the typical solution which consists in computing MFCCs and its first-order and second-order MFCC derivatives for every 10ms-spaced frames of length 25ms. Each frame is described with a feature vector of size 39 and is assigned the label of the signal. Classifiers are trained on all frames and at prediction time, a given 30-second signal is labelled according to the majority vote obtained over the frames it is composed by.
- Texture-based TFR analysis: we have also implemented the features extracted from time-frequency representations as proposed in [13]. These features exploit specific repeating spectral patterns in the TFR. Provided M blocks (or filters) of different sizes in time-frequency plane, feature generation proceeds by locally matching each of these blocks with the CQT representation and retaining the highest local degree of similarity of the block with the CQT image. This best local similarity serves as feature, leading to an M-dimensional feature vector. In agreement with [13], the used local filters are randomly sampled over the CQT representations of the training set. For a strongly textured TFR, the retrieved filters are likely to be representative of the repeated patterns. However, for less-structured audio scene, the ability of this approach to retrieve discriminative features may be spoiled. In addition, these features calculation is computationally demanding as it requires 2D convolutions of the full TFR for the matching. In our experiments, we have considered $M = 20$ blocks of different sizes.
- Recurrence plot analysis: these features are those introduced by [9], and they have achieved the best performance on the test set of the D-case audio scene challenge [10]. These are the features that we consider as the state-of-the-art. The idea is to extract from MFCC features, other characteristics that provide informations about recurrence over time of some specific MFCC patterns. Interestingly, the final features proposed by [9] are obtained through averaging over time of all time-localized MFCC and recurrence plot features. Hence, their features are of very low-dimensional and do not provide any time-related information. MFCC features have been computed as above. Then, for all the MFCCs obtained over batch of 40 windows with overlap of 20, 11 RQA features have been computed. Afterwards, MFCC features and RQA features are all averaged over time and MFCC averages, standard deviations and RQA averages are concatenated to form a 37-dimensional features.

  Note that we have considered an higher upper

frequency of the spectral analysis instead of the 900Hz used by [9]. Indeed, we believe that their choice was optimal for one dataset at the expense of genericity.

For our HOG feature, we have set the following parameters. The CQT transform is computed, by means of the [28]'s toolbox, on the same frequency range as the MFCC features and with 8 bins per octave. All other parameters have been kept as default as proposed in [28]. The time-frequency representation is then transformed into a 512×512 image. The cells for the histogram computation are of size 8×8 and we have chosen 8 orientations. Note that it is also possible to consider the signed direction of gradient (leading thus to a histogram of size 16). As described above, histograms are normalized according to some norms, 4 normalization factors are computed by the *vlfeat* toolbox [29] and we have considered the possibility of using them as complementary features.

In order to compare our HOG-based feature to its competitors, we fed them to the same classifier and evaluated the resulting performance. The classifier we have considered is an SVM classifier with either a linear kernel or a Gaussian one. All problems except the toy one are multiclass classification problems. Hence, we have used a one-against-one scheme for dealing with this situation.

For all the experiments, we have provided averaged results where the averaging occurs over 20 different splits of the dataset into a training set and a test set. For all datasets except the toy ones, 80% of the examples have been used for training. For the toy problem, we considered only 40 training examples among the 200 available. Note that all features have been normalized so as to have zero mean and unit variance on the training set. The test set has also been normalized accordingly. All the parameters of the SVM are tuned according to a validation scheme. The C parameter is selected among 10 values logarithmically scaled between 0.001 and 100 while the parameter σ of the Gaussian kernel $e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$ is chosen among $[1, 5, 10, 20, 50, 100]$. Model selection is performed by resampling 5 times, the training set into a learning and validation set of equal size. The best hyperparameters are considered as those maximizing averaged performances on the validation set.

As an evaluation criterion, we have considered the mean average precision, defined as

$$MAP = \frac{1}{N} \sum_{i=1}^{N} \frac{TP(i)}{\#N(i)}$$

where $TP(i)$ and $\#N(i)$ are respectively the number of examples of class i correctly classified and the total number of examples classified in class i.

IV. Experimental results

We have run several experiments aiming at showing the benefits of our HOG-based feature compared to the state-of-the-art, as well as at analyzing the influence of the different parameters of the HOG feature extraction pipeline.

A. Comparison with classical features

Our first result compares HOG features to some classical ones, the bag of MFCC, the bag of MFCC-D-DD as described above as well as the features based on MFCC, Texture, Recurrence quantitative analysis and the MFCC-D-DD frame-based classifier. We have considered two sets of signed HOG features. The first set, denoted as *HOG-full*, concatenates all histograms obtained from all the cells, resulting in a feature of very high-dimensionality. The second set of HOG features is obtained by averaging all the histograms over the time and over the frequency and by concatenating the two averaged histograms, denoted as *HOG-marginalized*. Note that we have also used MFCC and MFCC-RQA features obtained with an upper frequency of 900 Hz as in the paper of [9]. Results obtained according to the above-described protocol and the feature extraction parameters are depicted in Table II. A Wilcoxon signed-rank test with a p-value of 0.005 have been computed between the best competitor and the other approaches.

For the toy dataset, the best performance is obtained by the Texture and MFCC-D-DD based features with a perfect classification. These features seem to be able to capture the discriminative part of the signal either by capturing the MFCC variations or by finding the good matching patch. We also explain this good performance by the fact that the two classes have high-energy well-localized time-frequency structures. Except for these features, our HOG-based feature performs significantly better than its competitor RQA. MFCC is still able to discriminate the two chirps but less powerfully than the HOG feature, which natively captures the power spectrum variation. Interestingly, the best performing HOG feature is the one which considers all the histograms, despite the very-high dimensionality and the low number of training examples (40). A rationale for this, is that the discriminative parts of the signal are very well-localized, thus the HOG features obtained from the cells covering this region are strongly discriminative. Of course, removing other spurious HOG may have further

TABLE II

COMPARING PERFORMANCES OF DIFFERENT FEATURES ON THE DIFFERENT DATASETS. BOLD RESULTS DEPICT BEST PERFORMANCES FOR EACH DATASET AS WELL AS RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANTLY DIFFERENT ACCORDING TO A WILCOXON SIGNRANK TEST WITH A P-VALUE = 0.005. TEXTURE DENOTES THE FEATURES OBTAINED FROM THE WORK OF YU ET AL. [13] WHILE MFCC-D-DD IS RELATED TO THE MFCC AND DERIVATIVES FEATURES. MFCC, MFCC-RQA, MFCC-900 AND MFCC-RQA-900 RESPECTIVELY DENOTE THE MFCC FEATURES, THE MFCC AND RQA FEATURES AT CUT-OFF FREQUENCY OF 10 KHZ, THE MFCC AND THE MFCC AND RQA FEATURES WITH UPPER FREQUENCY SET AT 900 HZ. THE HOG (FULL) AND (MARGINALIZED) ARE RELATED TO THE HOG FEATURES WHICH ARE RESPECTIVELY OBTAINED BY CONCATENATING THE HISTOGRAMS FROM ALL CELLS AND BY CONCATENATING THE TWO MARGINALIZED HOG FEATURES.

			Datasets			
features	dim	classifier	Toy	EA	D-Case	Rouen
Texture	20	linear	**1.00 ± 0.00**	0.57 ± 0.13	0.23 ± 0.10	-
Texture	20	gaussian	**1.00 ± 0.01**	0.49 ± 0.13	0.19 ± 0.09	-
mfcc-d-dd	7800	linear	**1.00 ± 0.00**	**0.98 ± 0.04**	0.51 ± 0.13	0.66 ± 0.02
mfcc-d-dd	7800	gaussian	**1.00 ± 0.01**	0.97 ± 0.05	0.49 ± 0.15	0.69 ± 0.02
mfcc-d-dd	39	frame/majority	0.47 ± 0.04	0.95 ± 0.05	0.32 ± 0.12	0.36 ± 0.04
mfcc	3900	linear	0.78 ± 0.04	**1.00 ± 0.01**	0.53 ± 0.12	0.67 ± 0.01
mfcc	3900	gaussian	0.77 ± 0.03	**0.99 ± 0.03**	0.52 ± 0.10	0.76 ± 0.02
mfcc-900	3900	linear	0.50 ± 0.04	0.91 ± 0.07	0.53 ± 0.11	0.60 ± 0.02
mfcc-900	3900	gaussian	0.50 ± 0.04	0.86 ± 0.09	0.56 ± 0.12	0.66 ± 0.02
mfcc+RQA	37	linear	0.51 ± 0.02	0.95 ± 0.08	0.54 ± 0.09	0.78 ± 0.01
mfcc+RQA	37	gaussian	0.51 ± 0.04	0.96 ± 0.08	0.52 ± 0.10	**0.86 ± 0.02**
mfcc+RQA-900	37	linear	0.50 ± 0.04	0.93 ± 0.06	**0.68 ± 0.10**	0.72 ± 0.02
mfcc+RQA-900	37	gaussian	0.49 ± 0.04	0.93 ± 0.07	**0.65 ± 0.11**	0.80 ± 0.01
Hog-full	65536	linear	0.97 ± 0.02	**0.99 ± 0.02**	0.64 ± 0.09	0.84 ± 0.01
Hog-marginalized	2048	linear	0.94 ± 0.02	**0.97 ± 0.06**	**0.75 ± 0.10**	0.86 ± 0.01
Hog-marginalized	2048	gaussian	0.91 ± 0.04	0.95 ± 0.06	**0.71 ± 0.12**	**0.87 ± 0.01**

increased performance. We can also remark that the use of the frame/majority-based classifier with MFCC-D-DD leads to poor results. This can be easily explained by the fact that there are few discriminative frames in the audio scene and this induces errors in the majority vote scheme. We thus believe that such an approach is more appropriate to audio classification problems where discriminative features are persistent over time like in music genre classification. We will show in the sequel that when using appropriate cell size, our HOG feature will also obtain perfect classification.

The *East Anglia*'s dataset seems to be fairly easy and all features except the Texture-based and frame/majority approaches perform well with a slight advantage to MFCC.

For the *D-case* challenge, the Texture, MFCC, MFCC-D-DD and MFCC-RQA features perform poorly, with performances around 50% and even around 20% for the Texture. However, with a more adapted upper-frequency of the spectral analysis, performances of the MFCC-RQA reach 68% of mean average precision. Marginalized HOG features perform significantly better than competitors with a gain in performance of about 19% when the range of frequency 1-10000 Hz is considered. This gain drops to 7% but is still consequent when the range of frequency 1-900 Hz is used for MFCC-RQA. However, we can note that using the full HOG representation induces a slight loss of performances and that it seems valuable to consider some HOG pooling. We also want to highlight that the performance we report for MFCC-RQA-900 is slightly lower than those given in [9] and this is due to the fact our results are averages over 20 splits instead of a 5-fold cross-validation precision.

For the *Rouen*'s dataset we have introduced, the marginalized HOG features perform on par with MFCC-RQA and significantly better than all other competitors. Using a cut-off frequency of 900 Hz induces a larger loss of performance for these MFCC-RQA features. Also note that we cannot provide results using the Texture-based features as after one week, the feature computation was still running proving that approach is intractable even for medium-scale datasets. Again, MFCC-D-DD features with the frame/majority-based classifier fail to capture the discriminative information of the audio scene.

More interestingly, we highlight that the marginalized HOG feature is robust across the different datasets, especially when used in conjunction with a linear kernel, even though its extraction parameters have not been tuned. This is a very promising result concerning the generalization capability of these features. We will see in the sequel that by properly tuning the HOG parameters we are able to statistically significantly perform better than the MFCC-RQA feature on the Rouen's dataset.

B. Analyzing HOG feature parameters

In the first part of this experiment, we have investigated the influence of two parameters of the HOG features on the global classification performance. We have used the marginalized HOG features, as in the previous experiment, in conjunction with a linear kernel. The HOG features are composed of either signed, unsigned or both histograms eventually completed with the 4 normalization factors. Hence for each cell, the size of the feature ranges from 8 to $28 = 8 \times 3 + 4$. Since we have 64 rows and columns of 8×8 cells in the image, this results in feature vector of size ranging from 1024 to 3584.

The results we obtain for the different datasets are presented in Table III. We can note that the parameters we are evaluating clearly influence performances. Depending on the datasets, the variation of performances is in between 2% (for *East Anglia*) to 6% (for *D-case* and the toy dataset). The most consistent feature seems to be the ones for which histograms are computed with both signed and unsigned gradient orientations and the normalization factors are not included.

In the second part of the experiment, we have analyzed the effect of the average kernel size of the mean filtering, the number of orientations in the HOG as well as the size of the cell.The results are reported in Table IV. We note that these parameters have different influences given the dataset. However, it seems that the choices of a cell size of 8 and 8 orientations for the HOG computation are a good default choice. The parameter with most influence seems to be the filtering kernel size. Depending on the datasets, the best kernel size varies from 1, which corresponds to no filtering to 15. According to this result, we thus suggest practictioner to tune this parameter value according to its own dataset.

C. On the effect of pooling

We have analyzed the effects of pooling on the performances of the HOG features. Indeed, it is well known from the computer vision literature that pooling plays an important role when it comes to pattern recognition [25] and we believe that a proper choice of pooling can also improve performances in our audio scene classification problem. In the experiment, we varied the size of the average pooling in the time and frequency axis. Here, the HOG features is obtained using both signed and unsigned histograms and without the normalization factors. The mean filtering size, the number of cells and the number of orientations have been set to the values that maximize performances according to Table IV. Again, linear kernel is used in the SVM. Note that the form of HOG pooling investigated in this experiment does not necessarily perform better than those used in previous ones.

Results for different sizes of pooling are presented in Table V. Note that the first and last rows correspond respectively to the results obtained for the red and green pooling in Figure 3. Other rows are related to more general pooling form as in the blue pooling in Figure 3. A striking result can first be highlighted, regarding the importance of carefully selecting the pooling form : variation of performance between the worst and the best pooling form is at least 30% for all real datasets.

Worst performance is achieved by pooling over frequency, which means that we average all the obtained histograms over the frequency, losing all informations about spectral contents. This finding is rather intuitive as we believe that the audio scenes can be mostly discriminated by their spectral contents and the local variations of their spectral contents.

At the other end, best performances are obtained by pooling over time, especially when we consider the real datasets. This result is also interesting in the sense that best performances are achieved while no time information are kept in the features as they are totally translation-invariant. We make the hypothesis that this occurs because most audio scenes can be distinguished according to some global analysis (enhanced by pooling over time) of some recurrent patterns without the needs to look at some short-time single events, although these events may carry discriminative information. This rationale is also corroborated by the fact that MFCC-RQA features that are averaged over time performs reasonably well on the real datasets. While this approach works pretty well for the audio scene classes we have considered, we believe that features able to leverage on short-time events will be needed for fine-grained audio-scene classification.

Regarding other pooling forms, we can note a clear trend of improving performances as the pooling over frequency is decreased and the one over time increases.

D. More insights on the Rouen's dataset

As one of our main contribution in this paper is to introduce a novel audio scene dataset, we discuss in the sequel our findings regarding this dataset.

In Table V, we have shown that mean average precision, obtained as an average over 20 trials, is 0.9170. Table VI presents the normalized sum of all confusion matrices obtained from these 20 training/test splits. They have been obtained using the best performing HOG feature : namely the one with signed and unsigned orientations, without normalization factors of the histograms

TABLE III

ANALYZING THE EFFECTS OF HOG FEATURE PARAMETERS. TWO DIFFERENT PARAMETERS HAVE BEEN EVALUATED: THE SIGN OF GRADIENT ORIENTATIONS IN THE HISTOGRAM COMPUTATIONS (SIGNED, UNSIGNED AND BOTH) AND THE INCLUSION (WITH OR WITHOUT) OF THE NORMALIZATION FACTORS. BOLD RESULTS DEPICT BEST PERFORMANCES FOR EACH DATASET AS WELL AS RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANTLY DIFFERENT ACCORDING TO A WILCOXON SIGNRANK TEST WITH A P-VALUE = 0.005.

			Datasets			
sign	factors	dim	Toy	EA	D-case	Rouen
signed	w/o	2048	0.94 ± 0.02	**0.97 ± 0.06**	**0.75 ± 0.10**	0.86 ± 0.01
signed	with	2560	0.91 ± 0.02	**0.97 ± 0.08**	**0.73 ± 0.11**	**0.86 ± 0.02**
unsigned	w/o	1024	**0.97 ± 0.01**	**0.97 ± 0.07**	**0.70 ± 0.11**	0.83 ± 0.02
unsigned	with	1536	0.93 ± 0.02	**0.97 ± 0.08**	0.67 ± 0.11	0.84 ± 0.02
both	w/o	3072	**0.96 ± 0.01**	**0.98 ± 0.04**	**0.75 ± 0.12**	**0.86 ± 0.01**
both	with	3584	0.95 ± 0.02	**0.98 ± 0.06**	**0.75 ± 0.11**	**0.86 ± 0.01**

TABLE IV

ANALYZING THE EFFECTS OF HOG FEATURE PARAMETERS AND THE IMAGE FILTERING PARAMETER. THREE DIFFERENT PARAMETERS HAVE BEEN EVALUATED: THE SIZE OF THE CELL FOR NORMALIZATION, THE NUMBER OF ORIENTATIONS IN THE HOG COMPUTATION AND THE AVERAGE IMAGE FILTERING SIZE. BOLD RESULTS DEPICT BEST PERFORMANCES FOR EACH DATASET AS WELL AS RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANTLY DIFFERENT ACCORDING TO A WILCOXON SIGNRANK TEST WITH A P-VALUE = 0.005.

				Datasets			
filter size	Nb. Orient	cell size	dim	Toy	EA	D-case	Rouen
15	8	2	12288	**1.00 ± 0.00**	**0.97 ± 0.05**	0.66 ± 0.13	0.84 ± 0.01
15	8	4	6144	0.98 ± 0.01	**0.98 ± 0.05**	**0.73 ± 0.12**	0.85 ± 0.01
15	8	8	3072	0.96 ± 0.01	**0.98 ± 0.04**	**0.75 ± 0.12**	0.86 ± 0.01
15	8	16	1536	0.98 ± 0.01	**0.96 ± 0.07**	**0.73 ± 0.15**	0.86 ± 0.02
15	8	32	768	**1.00 ± 0.00**	**0.98 ± 0.04**	0.68 ± 0.12	0.83 ± 0.02
1	8	8	3072	0.97 ± 0.01	**0.97 ± 0.05**	**0.65 ± 0.12**	**0.89 ± 0.01**
5	8	8	3072	0.96 ± 0.01	**0.99 ± 0.01**	**0.69 ± 0.14**	**0.88 ± 0.01**
15	8	8	3072	0.96 ± 0.01	**0.98 ± 0.04**	**0.75 ± 0.12**	0.86 ± 0.01
30	8	8	3072	0.99 ± 0.01	**0.96 ± 0.07**	**0.69 ± 0.12**	0.82 ± 0.02
15	4	8	1536	0.79 ± 0.03	**0.98 ± 0.05**	**0.74 ± 0.12**	0.86 ± 0.01
15	8	8	3072	0.96 ± 0.01	**0.98 ± 0.04**	**0.75 ± 0.12**	0.86 ± 0.01
15	16	8	6144	0.98 ± 0.01	**0.97 ± 0.07**	**0.76 ± 0.10**	0.86 ± 0.02

and fully pooled over time, the HOG being obtained with cell size of 8, 8 bins of the orientations and no average filtering. The average precision obtained from this matrix is 0.915 and it is different to the mean average precision over the 20 runs as presented in Table V.

From Table VI, we can first note a group of conveyances *plane*, *bus*, *car*, *train* and *high-speed train* that are precisely recognized (with precisions above 97%). Figure 6 shows examples of CQT and HOG representation of some audio scenes related to these classes. Each class has specific signature with time-frequency structures around 30 Hz, 60 Hz and 240 Hz for *bus*, *car* and *plane*. Two other classes of conveyance *metro-rouen* and *metro-paris* are sometimes mislabelled showing that our HOG feature is not able to totally capture the fine-grained discriminative features between these two classes (if there is any).

A group of audio scenes composed of speeches (eventually loud ones) with specific short-time events (kid's scream and impacting balls) that occur all along the scenes is also precisely recognized with precision above 96%. These classes are *kid game hall* and *billiard pool hall*. Figure 7 shows examples of CQT and HOG representation of these audio scenes. In these plots, the short time-frequency structures appearing around 2000 Hz correspond to shouts and impacting balls.

We can also notice a group of scenes that seems difficult to distinguish as seen in Fig. 8: the ones in which some people are walking, namely *pedestrian street*, *market*, *train station hall*, *quiet street* and *shop*. These confusions are easily understandable. For instance, the main difference between a quiet street and pedestrian street would be the number of people walking in the scene. Such a difference is hardly taken into account by our HOG feature.

Despite the good skills exhibited by the proposed HOG features, the latter remarks show that there are still rooms for improvement, by addressing the issues raised by classes that are mixed. We believe that these classes show the needs for features (that can still be based on HOG) capturing discriminative short-time events that come in complement to our "global" features.

TABLE V

ANALYZING THE EFFECTS OF POOLING. THE NUMBER UNDER THE *Freq* AND *Time* LABELS DEPICTS THE NUMBER OF HISTOGRAMS ON THE FREQUENCY AND TIME AXES AFTER POOLING. FOR INSTANCE, THE FIRST ROW PRESENTS THE RESULT OF POOLING WHERE ALL HISTOGRAMS HAVE BEEN AVERAGED OVER THE FREQUENCY AXES. THIS POOLING CORRESPONDS TO THE RED POOLING IN FIGURE 3. BOLD RESULTS DEPICT BEST PERFORMANCES FOR EACH DATASET AS WELL AS RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANTLY DIFFERENT ACCORDING TO A WILCOXON SIGNRANK TEST WITH A P-VALUE = 0.005.

			Datasets			
Freq	Time	dim	Toy	EA	D-case	Rouen
1	64	1536	**0.99 ± 0.01**	0.69 ± 0.10	0.43 ± 0.11	0.41 ± 0.02
2	32	1536	0.98 ± 0.02	0.87 ± 0.07	0.54 ± 0.15	0.56 ± 0.02
4	16	1536	0.96 ± 0.02	0.92 ± 0.08	0.56 ± 0.13	0.68 ± 0.02
8	8	1536	0.93 ± 0.02	0.97 ± 0.05	0.64 ± 0.14	0.78 ± 0.01
16	4	1536	**0.99 ± 0.01**	**0.97 ± 0.06**	**0.70 ± 0.12**	0.85 ± 0.01
32	2	1536	0.98 ± 0.01	**0.99 ± 0.04**	**0.75 ± 0.12**	0.88 ± 0.01
64	1	1536	**0.99 ± 0.01**	**1.00 ± 0.01**	**0.73 ± 0.10**	**0.92 ± 0.01**

TABLE VI

NORMALIZED SUM OF ALL CONFUSION MATRICES OBTAINED OVER THE 20 TRAINING/TEST SPLITS. THE NORMALIZATION OCCURS BY COLUMNS SO THAT THE DIAGONAL TERMS REPRESENT THE PRECISION OBTAINED FOR A GIVEN CLASS. THE OBTAINED AVERAGE PRECISION IS 0.915. THE HOG FEATURES WE USED ARE THE BEST PERFORMING ONES ACCORDING TO ABOVE EXPERIMENTS. ROWS DEPICT THE REAL CLASS OF THE AUDIO SCENE WHILE COLUMNS ARE RELATED TO THE PREDICTED ONE.

	plane	busy street	bus	cafe	car	train station hall	kid game hall	market	metro-paris	metro-rouen	billiard pool hall	quiet street	student hall	restaurant	pedestrian street	shop	train	high-speed train	tubestation
plane	97.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
busy street	0.0	83.2	0.0	2.1	0.0	0.1	1.3	0.9	1.0	2.4	0.0	6.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0
bus	0.0	0.0	98.6	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
cafe	0.0	1.0	0.0	82.9	0.0	0.2	0.0	0.6	0.0	0.0	0.0	1.9	0.0	0.2	11.3	0.6	0.0	0.0	0.0
car	0.0	0.0	0.1	0.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
train station hall	0.0	0.2	0.0	0.2	0.0	92.9	1.2	1.7	0.2	0.0	0.0	0.0	0.0	0.2	2.2	0.0	0.0	0.0	0.0
kid game hall	0.0	0.0	0.0	0.0	0.0	0.0	96.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
market	0.0	0.3	0.0	1.4	0.0	1.3	0.0	90.5	0.0	0.0	0.0	0.6	0.0	0.7	4.6	1.4	0.0	0.0	0.2
metro-paris	0.0	2.8	0.5	0.2	0.0	0.2	0.0	0.0	86.7	7.2	0.0	1.3	0.0	0.0	0.2	0.6	0.0	0.0	0.0
metro-rouen	3.0	2.2	0.3	0.6	0.0	0.0	0.7	0.0	10.0	88.3	0.0	0.3	0.0	0.0	0.0	0.0	0.6	0.0	0.0
billiard pool hall	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	100.0	0.0	0.0	0.0	0.7	0.2	0.0	0.0	0.0
quiet street	0.0	7.5	0.0	3.5	0.0	0.0	0.0	0.0	0.6	0.2	0.0	75.9	0.3	0.0	3.9	4.3	0.0	0.0	0.4
student hall	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.5	2.1	1.7	1.3	0.0	0.0	0.0
restaurant	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.8	0.0	0.0	0.0	0.0	0.9	92.4	0.0	0.6	0.0	0.0	0.0
pedestrian street	0.0	1.2	0.0	7.2	0.0	2.8	0.0	2.5	1.0	0.0	0.0	5.1	0.0	1.1	70.9	2.9	0.0	0.0	0.2
shop	0.0	0.2	0.0	1.2	0.0	0.6	0.2	2.5	0.0	0.3	0.0	5.1	0.3	3.2	3.3	86.4	0.0	0.0	0.2
train	0.0	0.0	0.5	0.0	0.1	0.6	0.0	0.0	0.2	1.1	0.0	0.0	0.0	0.0	0.0	0.0	99.1	0.0	0.0
high-speed train	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.3	100.0	0.0
tubestation	0.0	1.5	0.0	0.4	0.0	0.8	0.0	0.4	0.0	0.0	0.0	3.5	0.0	0.2	1.1	1.6	0.0	0.0	98.9

V. CONCLUSION

The problem of classifying audio scene is currently a hot topic in the computational auditory scene analysis domain. For this specific problem, we have introduced in this paper a novel feature that seems to be very promising at capturing relevant discriminative informations. The main block of the feature we proposed has been initially proposed in the computer vision domain, namely histogram of gradients.

Our novel feature has been obtained by computing histogram of gradients of a constant Q-transform followed by an appropriate pooling. We have experimentally proved that these histograms of gradients were useful for capturing specific characteristics present in a time-frequency representation that classical features such as MFCC can not encode namely the local variation of power spectrum. Then, our experimental results on real datasets clearly show that our features achieve state-of-the-art classification performances on several datasets.

While our HOG-based feature is globally efficient, the overall pipeline for audio scene classification still lacks in discriminating some difficult classes. In order to further improve the scheme, some efforts are still needed. Our future researches focus on improving discriminative ability of HOG-based feature by working on the pooling strategy. The supervised learning paradigm may also be improved by taking into account an hierarchical taxonomy of the classes. We plan to investigate this taxonomy by learning it directly from the data.

REFERENCES

[1] D. Wang and G. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms and applications*. Wiley-Interscience, 2006.

[2] J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frame approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

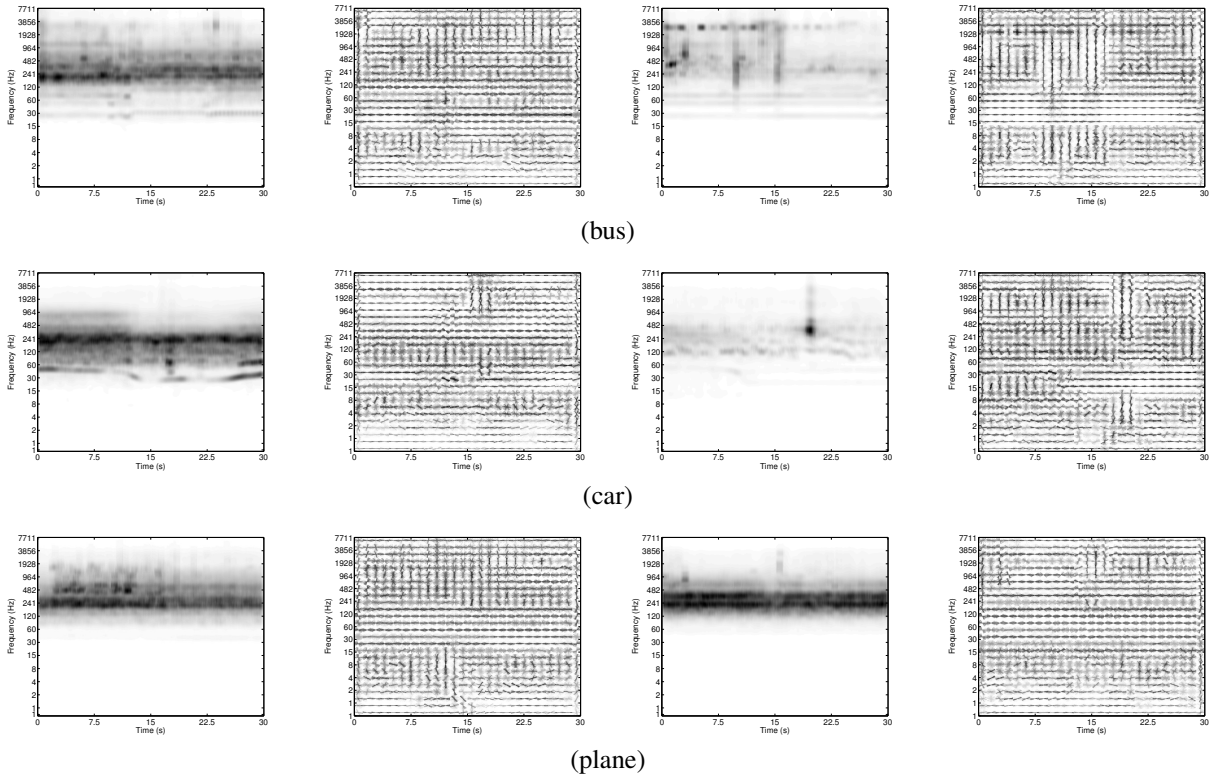


Fig. 6. Examples of CQT representation and associated HOG representations for transportation devices (top) bus. (middle) car. (bottom) plane.

[3] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing*, vol. 3, 2006.

[4] B. Cauchi, "Non-negative matrix factorization applied to auditory scenes classification," Master's thesis, Master ATIAM, Université Pierre et Marie Curie, 2011.

[5] P. Hu, W. Liu, and W. Jiang, "Combining frame and segment based models for environmental sound classification," in *Proceedings of 13th Annual Conference of the International Speech Communication Association*, 2012.

[6] K. Lee, Z. Hyung, and J. Nam, "Acoustic scene classification using sparse feature learning and event based pooling," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[7] S. Chu, S. Narayan, and C. J. Kuo, "Environment sound recognition with time-frequency audio features," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[8] J. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[9] G. Roma, W. Nogueira, and P. Herrera, "Recurrence quantification analysis features for environmental sound recognition," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[10] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, and M. Lagrange, "Detection and classification of acoustic scenes and events: an ieee aasp challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[11] C. Cotton and D. Ellis, "Spectral vs spectro-temporal features for acoustic event detection," in *IEEE Workshop on applications of Signal processing to audio and acoustics*, 2011, pp. 69–72.

[12] E. Benetos, M. Lagrange, and S. Dixon, "Characterisation of acoustic scenes using a temporally-constrained shift-invariant model," in *Proceedings of the fifteenth International Conference on Digital Audio effects*, 2012.

[13] G. Yu and J. Slotine, "Audio classification from time-frequency texture," in *in Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing*, 2009.

[14] J. Dennis, H. Tran, and E. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 367–377, 2013.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[16] J. Neumann, C. Schnorr, and G. Steidl, "Efficient wavelet adaptation for hybrid wavelet-large margin classifiers," *Pattern Recognition*, vol. 38, no. 11, pp. 1815–1830, 2005.

[17] S.-Y. Lung, "Feature extracted from wavelet decomposition using biorthogonal riesz basis for text-independent speaker recognition," *Pattern recognition*, vol. 41, no. 10, pp. 3068–3070, 2008.

[18] J. Cheng, Y. Sun, and L. Ji, "A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines," *Pattern reco*, vol. 43, no. 11, pp. 3846–3852, 2010.

[19] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern recognition letters*, vol. 24, no. 15, pp. 2895–2907, 2003.

[20] J. Brown, "Calculation of a constant Q spectral transform,"

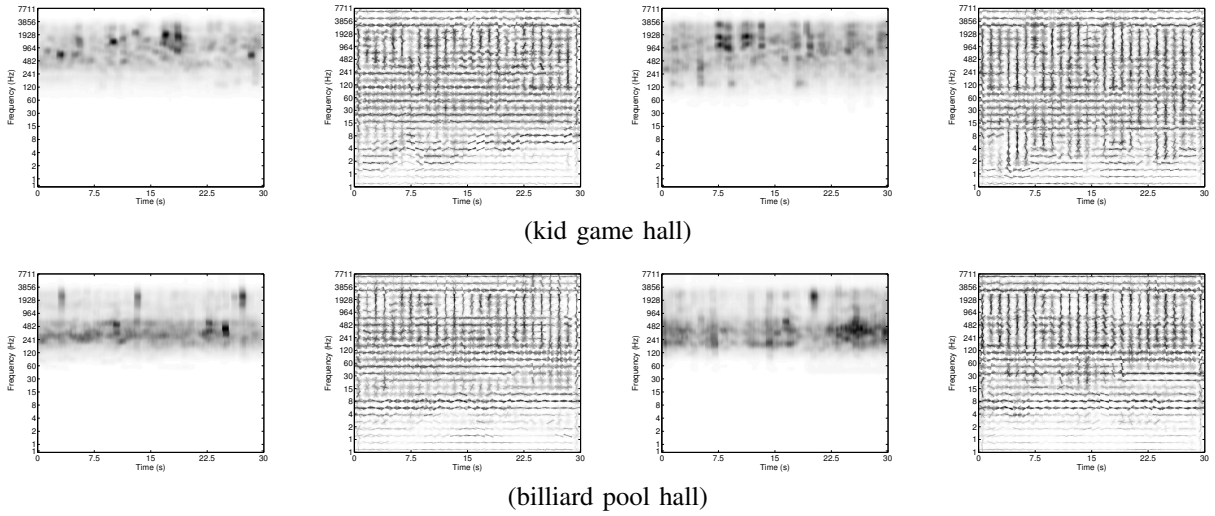


Fig. 7. Examples of CQT representation and associated HOG representations for audio scene with babble noises and short-time events (top) kid game hall (bottom) pool hall.

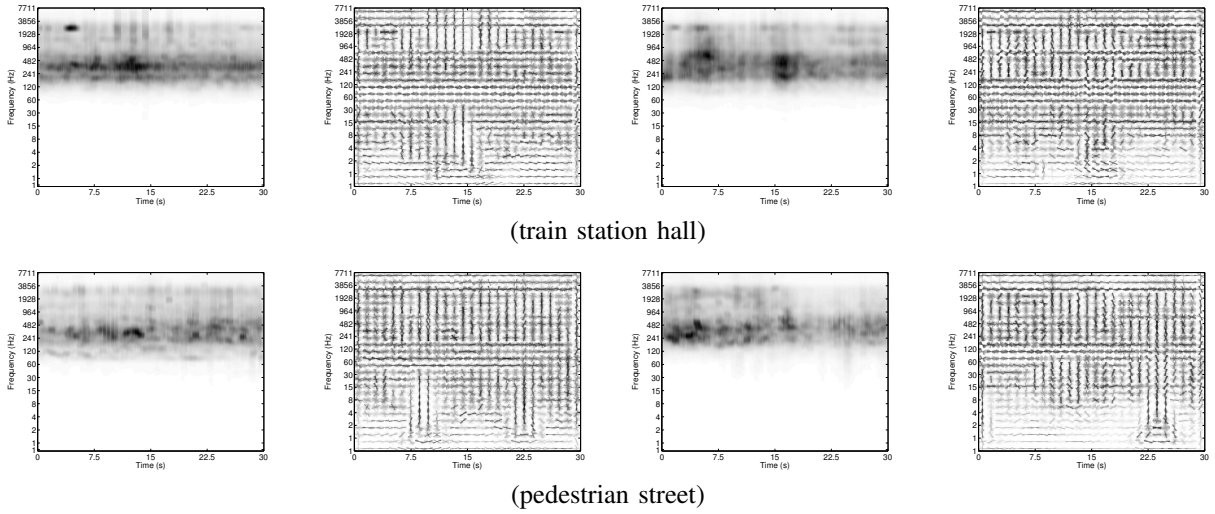


Fig. 8. Examples of CQT representation and associated HOG representations for audio scene with several people walking (top) train station hall (bottom) pedestrian street.

Journal of Acoustic Society of America, vol. 89, no. 1, pp. 425–434, 1991.

[21] P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[22] R. Minetto, N. Thome, M. Cord, N. Leite, and J. Stolfi, "T-HOG: An effective gradient-based descriptor for single line text regions," *Pattern recognition*, vol. 46, no. 3, pp. 1078–1090, 2013.

[23] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in vision algorithms," in *Proceedings of the International Conference on Machine Learning*, 2010.

[26] L. Ma, D. Smith, and B. Milner, "Context awareness using environmental noise classification," in *Proceedings of Eurospeech*, 2003, pp. 2237–2240.

[27] D. Ellis, "PLP and RASTA and MFCC and inversion in matlab," online web resource : available at http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat, 2005.

[28] C. Schoerkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proceedings of the Sound and Music Computing Conference*, 2010.

[29] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.