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

Alain Rakotomamonjy, Gilles Gasso

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

Alain Rakotomamonjy[a,*], Gilles Gasso[b]

[a] *LITIS EA4108, Normandie Université, UR, Avenue de l'université, 76800 Saint Etienne du Rouvray*
[b] *LITIS EA4108, Normandie Université, IR, Avenue de l'université, 76800 Saint Etienne du Rouvray*

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 time-frequency representation of an audio scene. Contrarily to classical audio features like MFCC, we make the hypothesis that histogram 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, histogram 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 900 minutes of audio scene recording. We thus believe that it may be the next standard dataset for evaluating audio scene classification algorithms. Our comparison results clearly show that our HOG-based features outperform its competitors
Keywords: histogram of gradient; Time-Frequency Representation; audio scene; MFCC; support vector machines.

1. 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) [26]. 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

*Corresponding author: Tel.: +33232955334;
Email address: alain.rakoto@insa-rouen.fr (Alain Rakotomamonjy)

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 can be assimilated to an audio recording specific to 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) [1, 18, 5, 15]. 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, [1] consider a Gaussian Mixture Model for estimating the distribution of the MFCC coefficients, while [16] 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, [13] 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 [22]. The advantage of RQA is that it allows the analysis of recurrent behaviour in the MFCC coefficients over time. According to the recent D-Case challenge on audio scene recognition [14], combining MFCC features in conjunction to 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, [8] 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. [2] have followed similar idea but instead of considering NMF they employed a probabilistic model denoted as probabilistic latent component analysis.

In this paper, we 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 [10] 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 this 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 [14] is an excellent initiative of this kind although its number of examples is limited (100 for the publicly available examples). Hence, another of our contribution is to present a new dataset for audio scene recognition. It is based on about 900 minutes of recording on different locations (up to 19). This dataset will be made publicly available and we expect that it becomes one of the classical benchmark of this kind of recognition task.

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.

2. From signal to histogram of gradients features

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

2.1. 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 preprocessed so as to attenuate some spurious noises that may hinder some relevant information related to high-energy time-frequency structures. Afterwards, the resulting pre-processed time-frequency representation image is used as input of our histogram of gradient feature extraction. In a nutshell, the idea of histogram of gradient is to locally analyze the direction of energy's variation in the time-frequency representation. As described in more details in what follows, 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.

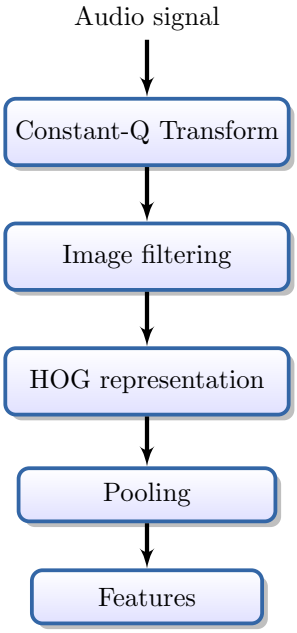


Figure 1: Block Diagram of the HOG-based feature extraction

The block diagram of this feature extraction scheme is illustrated in Figure 1.

2.2. 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 a time-frequency representations of sound. Depending on the task at hand, they either consider wavelet-based [21, 17], MFCC-based [6] or short-time fourier-based representations [9]. In this work, we do not depart from this widely adopted framework and choose to represent the signal according to a constant-Q transform [4]. 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 [4]. Example of a representation of a linear chirp signal is given in Figure 2.

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. Then, depending on the noise present in the TFR images, some denoising process can be applied. In our work, we have kept things simple and only used an average filtering. Our goal when smoothing the image, is to reduce strong local variations in the image that will tend to produce high gradient variation (which may be not relevant for audio scene recognition).

2.3. Histogram of gradients

Histogram of gradients have been originally introduced by [10] 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 [10, 12, 20], 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 [10, 12] and they are both based on the following steps:

1. compute the gradient of the TF image
2. compute angle 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 orientation are bidirectional or not, whether the magnitude of the gradient are 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 [25], and for more details, we refer the reader to [10] and [12].

As an illustration, we depict in Figure 2 the 512×512 average filtered image of a linear chirp's CQT transform, as well as the resulting histogram of gradient 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. They are essentially due to presence of small variations of gradient in low-energy time-frequency structure in the CQT 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.

After having computed the histogram of gradient in the images, we are thus left with a representation composed of an histogram in all the cells. If we concatenate all these histograms for yielding our 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

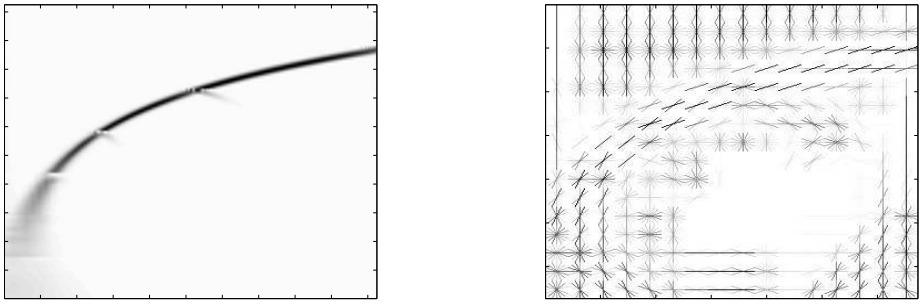


Figure 2: Example of HOG for a toy linear chirp. (left) 512×512 image of the CQ transform of the signal. (right) histogram of gradient representation of the signal. Note that for a sake of interpretation, each cell in the plots represents the occurrence of edge orientation in the cell and the darker 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.

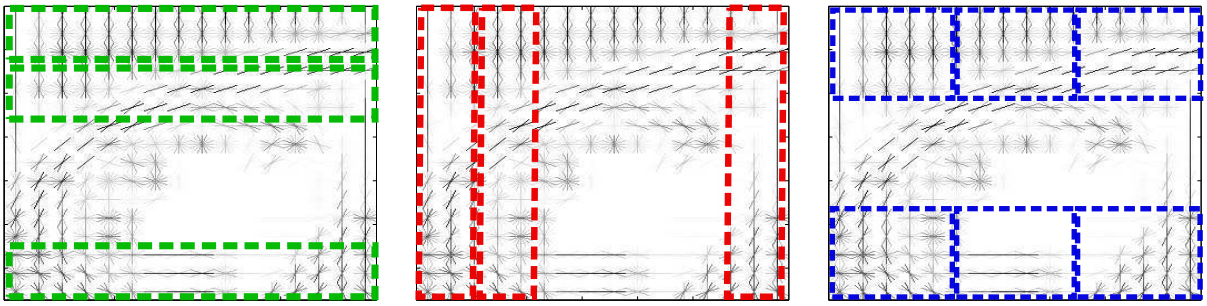


Figure 3: Illustrating the different types of pooling we investigate. (left) time pooling.(Middle) frequency pooling. (right) block-sized pooling. The green, red and blue boxes provide an example on the regions on which local histograms are averaged. The x-axis and y-axis respectively denote the time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

on the number of audio scene examples, it thus may be beneficial to reduce the dimensionality of the problem for instance by pooling.

2.4. 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 neighbourhood. The pooling may result in a more smoothed and robust information. This technique is a step commonly considered with success in modern visual recognition algorithms [3]. 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 all the pooled histograms forms now the feature vector that will be used for learning the audio scene classifier.

2.5. Discusssions

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. Still by construction, 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 MFCC will fail in fully capturing the discriminative information brought by these signals at the contrary of the features we propose.

3. 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.

3.1. Toy dataset

For evaluating our features, we have created a toy problem which highlights the ability of our HOG-based feature in capturing power spectrum's direction of variation. As such, we have created a binary classification

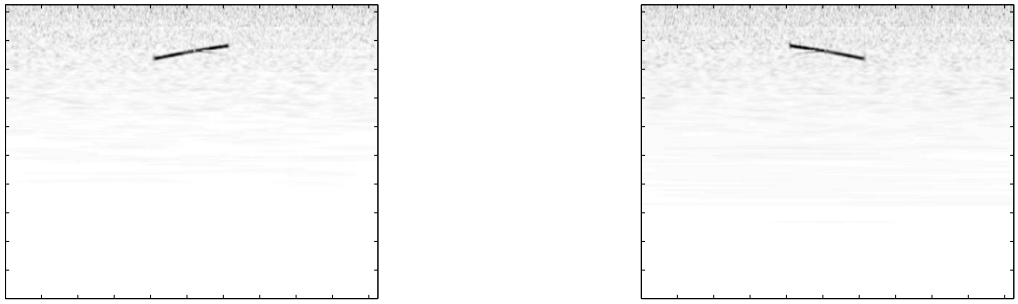


Figure 4: Examples of CQT representation of the two localized linear chirps used in the toy dataset. These images are obtained after a 512×512 resizing and a 15×15 average filtering

.

problem where signals from each class are composed of a localized linear chirp, respectively of increasing and decreasing frequency, defined as

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

where n is a centered Gaussian noise of standard deviation 0.4 and $\Pi_{[t_1,t_2]}(t)$ a function which value is 1 when $t_1 \leq t \leq t_2$ and 0 otherwise. For the first class, we have set $a = 1200$ and $b = 0$ while for the second class we have $a = -1200$ and $b = 2400$. These signals have been sampled at $10kHz$ on the interval $[0, 1]$, with $t_1 = 0.4$ and $t_2 = 0.6$. We have created 100 examples per classes.

3.2. D-case challenge dataset

For the purpose of a challenge, a dataset providing environmental sound recordings has been recently released by [14]. 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 $44.1kHz$ 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[1]

3.3. East Anglia (EA) dataset

This dataset[2] has been collected in the early 2000 by Ma et al. [19] 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

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

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

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.

3.4. Litis Rouen dataset

This dataset we make publicly available[3] goes beyond the above ones in terms of volume and number of locations. Recordings have been performed using a Galaxy S3 smart-phone 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 microphone is more likely to be used. The sampling frequency we have used is 22050 Hz and the recording is saved as a MP3 file with a bitrate of 64 kbps. Overall, about 900 minutes of audio scene have been recorded. They took place from December 2012 to November 2013. 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. A summary of the dataset is given in Table 1 and Figure 5 presents some samples of CQT for 3 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. For the *train station hall*, we remark the presence of some time-frequency structures associated with babble noises and sounds of people walking.

3.5. 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 [1]. The setting for the MFCC computations are typical. We have extracted MFCC features from 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* with the dithering option on [11]. 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 windows of size 40ms and hops of 20ms and concatenated together all the MFCC averages and standard deviations.

[3] the website for downloading the data will be accessible as soon as this companion paper will be available online

Classes	#examples
plane	23
busy street	92
bus	65
cafe	46
car	37
train station hall	175
kid game hall	56
market	51
metro-paris	46
metro-rouen	99
billiard pool hall	100
quiet street	26
student hall	44
restaurant	534
pedestrian street	59
shop	52
train	66
high-speed train	53
tubestation	169
	1793

Table 1: Summary of the Litis Rouen audio scene dataset

- Recurrence plot analysis: these features are those introduced by [22], and they have achieved the best performance on the test set of the D-case audio scene challenge [14]. 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 [22] 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 each sliding window of size 40ms and hops of 20ms, 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 than the 900Hz used

by [22]. However, we believe that their choice is 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 [23]'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 [23]. 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 [24] 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 a 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

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

where $TP(i)$ and $\#C(i)$ are respectively the number of examples of class i correctly classified and the total number of examples in class i. This is equilavent to the correct recognition rate when classes are even (this is the case for all datasets except the Rouen's ones).

4. 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.

4.1. Comparison with classical features

Our first result compares HOG features to some classical ones, the bag of MFCC as described above as well as the features based on MFCC and Recurrence quantitative analysis. We have considered two sets of signed HOG features, one which concatenates all histograms obtained from all the cells, denoted as *HOG-full*. This results, as reported, 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 [22]. Results obtained according to the above-described protocol and the feature extraction parameters are depicted in Table 2. A Wilcoxon signed-rank test with a p-value of 0.05 have been computed between the best HOG feature and its best competitor in order to evaluate whether differences are statistically significant.

As expected, for the toy dataset, our HOG-based feature performs significantly better than its competitor. MFCC is still able to discriminate the two chirps but less powerfully than the HOG feature, which natively capture the power spectrum variation. Interestingly, the best performing 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 increased performance.

The *East Anglia*'s dataset seems to be fairly easy and all features perform good with a slight advantage to MFCC.

For the *D-case* challenge, the MFCC and MFCC-RQA features perform poorly, with performances around 55%. 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 performs significantly better than competitors with a gain in performance of about 17% when the range of frequency (1-10 KHz) is considered. This gain drops to 5% 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 want to highlight that the performance we report for MFCC-RQA-900 is slightly lower than those given in [22] and this is due to the fact our results is obtained as an average of 20 splits instead of a 5-fold cross-validation.

For the *Rouen*'s dataset we have introduced, the marginalized HOG features perform again significantly better than all competitors and the MFCC-RQA features yield about $3-4\%$ loss of performances compared to the HOG. Note that using a cut-off frequency of 900 Hz induces a larger loss of performance for this MFCC-RQA features. More interestingly, we highlight that the marginalized HOG feature is robust across

			Datasets			
features	dim	kernel	Toy	EA	D-case	Rouen
mfcc	3900	linear	0.76 ± 0.04	**0.99 ± 0.02**	0.55 ± 0.09	0.71 ± 0.02
mfcc	3900	gaussian	0.75 ± 0.04	**0.99 ± 0.02**	0.54 ± 0.08	0.74 ± 0.02
mfcc-900	3900	linear	0.54 ± 0.04	0.97 ± 0.03	0.55 ± 0.08	0.66 ± 0.02
mfcc-900	3900	gaussian	0.51 ± 0.04	0.97 ± 0.05	0.54 ± 0.12	0.68 ± 0.02
mfcc-RQA	37	linear	0.50 ± 0.04	0.95 ± 0.04	0.56 ± 0.07	0.76 ± 0.02
mfcc-RQA	37	gaussian	0.51 ± 0.04	0.94 ± 0.04	0.54 ± 0.08	0.82 ± 0.03
mfcc-RQA-900	37	linear	0.49 ± 0.04	0.95 ± 0.04	0.68 ± 0.09	0.74 ± 0.02
mfcc-RQA-900	37	gaussian	0.49 ± 0.03	0.96 ± 0.03	0.65 ± 0.08	0.79 ± 0.02
Hog-full	65536	linear	**0.93 ± 0.02**	0.95 ± 0.06	0.65 ± 0.06	0.75 ± 0.02
Hog-marginalized	2048	linear	0.92 ± 0.03	**0.97 ± 0.05**	**0.73 ± 0.09**	0.85 ± 0.02
Hog-marginalized	2048	gaussian	0.89 ± 0.06	0.94 ± 0.06	0.69 ± 0.11	**0.86 ± 0.02**

Table 2: 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 = 0.05. 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.

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

4.2. Analyzing HOG feature parameters

In this next 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 3. 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*). The most consistent feature seems to be the ones for which histograms are computed with both signed and unsigned gradient orientations and the normalization

			Datasets			
sign	factors	dim	Toy	EA	D-case	Rouen
signed	w/o	2048	0.92 ± 0.03	**0.97 ± 0.05**	**0.73 ± 0.09**	0.85 ± 0.02
signed	with	2560	0.90 ± 0.03	**0.97 ± 0.04**	**0.72 ± 0.09**	**0.86 ± 0.02**
unsigned	w/o	1024	**0.96 ± 0.02**	0.96 ± 0.05	0.68 ± 0.09	0.84 ± 0.02
unsigned	with	1536	0.92 ± 0.02	0.95 ± 0.06	0.67 ± 0.09	0.85 ± 0.02
both	w/o	3072	**0.96 ± 0.02**	**0.97 ± 0.04**	**0.73 ± 0.10**	**0.86 ± 0.02**
both	with	3584	0.94 ± 0.02	**0.97 ± 0.04**	**0.73 ± 0.09**	**0.87 ± 0.02**

Table 3: 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 = 0.05.

factors are not included.

4.3. 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 [3] 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. Again, linear kernel is used in the SVM.

Results for different sizes are presented in Table 4. 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. From this point of view, this finding is thus 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 : they are totally translation-invariant. We make the hypothesis that this occurs because the audio scenes we are trying to classify are “textured” audio scene. By “textured”, we

			Datasets			
Freq	Time	dim	Toy	EA	D-case	Rouen
1	64	1536	0.89 ± 0.02	0.63 ± 0.09	0.43 ± 0.11	0.42 ± 0.04
2	32	1536	0.87 ± 0.03	0.77 ± 0.07	0.57 ± 0.10	0.57 ± 0.02
4	16	1536	0.93 ± 0.02	0.87 ± 0.07	0.52 ± 0.08	0.67 ± 0.02
8	8	1536	**0.97 ± 0.02**	0.93 ± 0.05	0.62 ± 0.07	0.76 ± 0.03
16	4	1536	**0.97 ± 0.02**	0.96 ± 0.05	0.69 ± 0.09	0.83 ± 0.03
32	2	1536	**0.97 ± 0.02**	**0.97 ± 0.05**	**0.72 ± 0.09**	0.86 ± 0.02
64	1	1536	0.92 ± 0.03	**0.98 ± 0.04**	**0.74 ± 0.08**	**0.88 ± 0.02**

Table 4: Analyzing the effects of pooling. The number under the *Freq* and *Time* labels depict 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 = 0.05.

mean that most of them can be distinguished according to some global analysis of some recurrent patterns without the needs to look at some short-time single events. This rationale we provide is also corroborated by the fact that MFCC-RQA features that are averaged over time performs reasonably well on the real datasets.

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.

4.4. 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.

Table 5 presents the sum of all confusion matrices obtained over the 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 histograms and fully pooled over time.

We can first note that 6 among the 19 classes are always correctly predicted. These classes are : *plane*, *bus*, *car*, *high-speed train*, *kid game hall* and *billiard pool hall*. Interestingly, they form 2 groups with distinct characteristics. The first group is a group of transportation devices, in which each item has a distinct audio signature. The second group is a group of locations composed of speeches (eventually loud ones) with specific short-time events (kid's scream and impacting balls) that occur all along the scenes. Thus, it seems that our HOG feature is also able to capture these types of discriminating features.

There are several groups of audio scenes that are frequently confused. The two most difficult classes to discriminate are the *metro-rouen* and the *tubestation* ones. Again, this seems natural as the *tubestation*

	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	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
busy street	0	332	0	0	0	0	0	7	1	8	0	3	0	0	2	0	0	0	7
bus	0	0	260	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cafe	0	0	0	161	0	1	0	5	0	0	0	3	0	1	4	5	0	0	0
car	0	0	0	0	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0
train station hall	0	6	0	0	0	656	0	0	0	0	0	0	0	2	0	0	0	0	36
kid game hall	0	0	0	0	0	0	220	0	0	0	0	0	0	0	0	0	0	0	0
market	0	0	0	0	0	5	0	183	0	0	0	0	0	9	0	3	0	0	0
metro-paris	0	1	0	0	0	2	0	1	162	12	0	0	0	0	0	1	0	0	1
metro-rouen	0	9	0	0	0	1	0	3	10	224	0	1	0	0	0	0	0	0	152
billiard pool hall	0	0	0	0	0	0	0	0	0	0	400	0	0	0	0	0	0	0	0
quiet street	0	18	0	10	0	0	0	0	0	0	0	48	0	0	22	0	0	0	2
student hall	0	0	0	0	0	0	0	0	0	0	1	0	162	17	0	0	0	0	0
restaurant	0	0	0	0	0	0	0	0	0	0	0	0	2	2130	3	5	0	0	0
pedestrian street	0	1	0	11	0	7	0	4	0	4	4	4	0	0	197	6	0	0	2
shop	0	6	0	8	0	10	0	0	0	0	0	0	0	6	10	160	0	0	0
train	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	254	0	4
high-speed train	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	220	0
tubestation	0	8	0	0	0	37	0	1	9	138	0	1	0	0	7	4	0	0	475

Table 5: Sum of all confusion matrices obtained over the 20 training/test splits. 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.

recordings are frequently composed of arriving and departing *metro-rouen* and the main difference between the scenes is that the former one is recorded while inside the tube and the latter is obtained while the tube is arriving, departing or when none of these situations is active.

We can also remark another group of audio scenes that seems to be difficult to distinguish: the ones composed of walking people, namely *pedestrian street*, *quiet street* and *train station hall*.

While the features we propose seems to capture the discriminative features of all audio scenes, there are still rooms for improvement, by addressing the issues raised by these classes that are highly 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.

5. 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, the local variation of power spectrum. Then, our experimental results on real datasets clearly shown 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 dicriminative 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 take into account this taxonomy by learning it directly from the data.

References

[1] Aucouturier, J., Defreville, B., Pachet, F., 2007. 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 122, 881–891.

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

[3] Boureau, Y.L., Ponce, J., LeCun, Y., 2010. A theoretical analysis of feature pooling in vision algorithms, in: Proceedings of the International Conference on Machine Learning.

[4] Brown, J., 1991. Calculation of a constant Q spectral transform. Journal of Acoustic Society of America 89, 425–434.

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

[6] Cheng, J., Sun, Y., Ji, L., 2010. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. Pattern reco 43, 3846–3852.

[7] Chu, S., Narayan, S., Kuo, C.J., 2009. Environment sound recognition with time-frequency audio features. IEEE Trans. on Audio, Speech and Language Processing 17, 1142–1158.

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

[9] Cowling, M., Sitte, R., 2003. Comparison of techniques for environmental sound recognition. Pattern recognition letters 24, 2895–2907.

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

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

[12] Felzenszwalb, P., Grishick, R., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part based models. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 1627–1645.

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

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

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

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

[17] Lung, S.Y., 2008. Feature extracted from wavelet decomposition using biorthogonal riesz basis for text-independent speaker recognition. Pattern recognition 41, 3068–3070.

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

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

[20] Minetto, R., Thome, N., Cord, M., Leite, N., Stolfi, J., 2013. T-HOG: An effective gradient-based descriptor for single line text regions. Pattern recognition 46, 1078–1090.

[21] Neumann, J., Schnorr, C., Steidl, G., 2005. Efficient wavelet adaptation for hybrid wavelet-large margin classifiers. Pattern Recognition 38, 1815–1830.

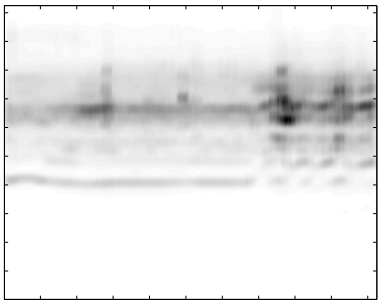
[22] Roma, G., Nogueira, W., Herrera, P., 2013. Recurrence quantification analysis features for environmental sound recognition, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

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

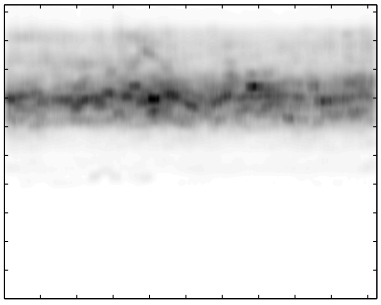
[24] Vedaldi, A., Fulkerson, B., 2008. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

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

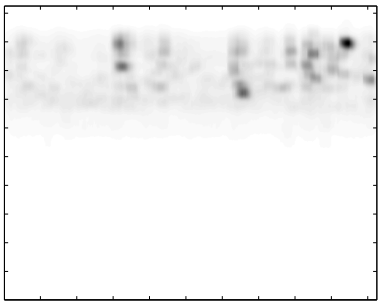
[26] Wang, D., Brown, G. (Eds.), 2006. Computational auditory scene analysis: Principles, algorithms and applications. Wiley-Interscience.



(bus)



(train station hall)



(kid game hall)

Figure 5: Examples of CQT of audio signals from 3 different scenes of the Rouen's dataset. x-axis represents time while y-axis denotes increasing frequencies.