HISTOGRAM OF GRADIENTS OF TIME-FREQUENCY REPRESENTATIONS FOR AUDIO SCENE CLASSIFICATION

A. Rakotomamonjy[*]

Normandie Université
UR LITIS,
76800 Saint Etienne du Rouvray France
alain.rakoto@insa-rouen.fr

G. Gasso

Normandie Université
IR LITIS
76800 Saint Etienne du Rouvray France
gilles.gasso@insa-rouen.fr

ABSTRACT

This abstract presents our entry to the *Detection and Classification of Acoustic Scenes* challenge. The approach we propose for classifying acoustic scenes is based on transforming the audio signal into a time-frequency representation and then in extracting relevant features about shapes and evolutions of time-frequency structures. These features are based on histogram of gradients that are subsequently fed to a multi-class linear support vector machines.

Index Terms— Constant Q transform, Histogram of gradient, Support vector machines.

1. INTRODUCTION

In order to get some context awareness, a machine, say a smartphone or any portable electronic device, should be able to predict the environment in which it currently resides. Such awareness can be brought throught vision or audio scene analysis. The latter is a very complex problem since a scene related to a given place can be composed of potentially an infinite amount of single sound events while only few of these events provide some information on the location of the recording.

While the literature on this problem of acoustic scene recognition is rather large and several approaches have been proposed [1, 2, 3], a timely challenge has been organised in order to evaluate, with less bias as possible, the different algorithms available for solving this problem. This abstract describes the algorithm we have proposed for this challenge. Basically, we solve the problem by transforming the audio scene into a time-frequency image and by computing discriminative features that are commonly considered in the computer vision community for recognizing objects. These features are then fed to a multi-class linear support vector machines for learning a decision function.

2. DATA AND FEATURE EXTRACTION

The data we have to deal with are 30s audio scenes acquired in different places. Our objective is to learn from some labeled examples of audio scene the place where they have been acquired : this task is thus a task of classifying acoustic scenes. 10 examples par classes are available and they are 10 different classes.

The approach we have developed for solving this task is to cast the problem as a machine learning problem where each of the labeled acoustic scene is considered as a single object. Hence, as in many machine learning task, the most difficult problem is to design some features that are able to grasp the specificities of each acoustic scene class while preserving discriminative power.

The main novelty of our approach is provided by this feature extraction stage since the classification learning stage will be performed by a standard multi-class linear SVM in a *one-against-one* framework. The main idea of the feature extraction is to represent this 30s acoustic scene through a time-frequency representation, in our case a constant Q transform, to consider this TF representation as an image and then to consider typical image classification feature extraction as our acoustic scene features.

For transforming the acoustic scene signal into an 512×512 image, we have applied the following steps.

- the stereo signal is averaged over the two channels and a constant Q tranform of this resulting signal is computed over the frequency f_{min} to f_{max}, N_b bins per octave and a Q factor of 1. This transform has been computed owing to the Matlab toolbox described in [4].
- The resulting time-frequency representation is transformed into a 512×512 image by means of a cubic interpolation followed by an $N_a \times N_a$ averaging filtering in order to smooth all abrupt variations. One of the reason of why we have considered this image transformation is the large number of temporal samples on which the constant Q tranform has been computed. The resulting image can now be considered as a low temporal resolution of the original signal and its time-frequency representation, but we believe that this resolution is sufficient enough for discriminating acousting scenes.

Figure 1 shows an example of a constant Q tranform as a 512×512 image. We can note that the image representation seems to preserve all the time-frequency structures present in the original signal. Now the question is : how can be extract some features from this image that is able to provide information on these TF structures?

For this purpose, our feature extraction relies on a technique that has been proved to be very efficient for object detection especially for person or pedestrian detection, namely histograms of oriented gradient (HOG) [5, 6]. The idea behind HOG is that an object

[*]This work has been partially funded by the French ANR agency through the grant ANR Blanc-12 GRETA.

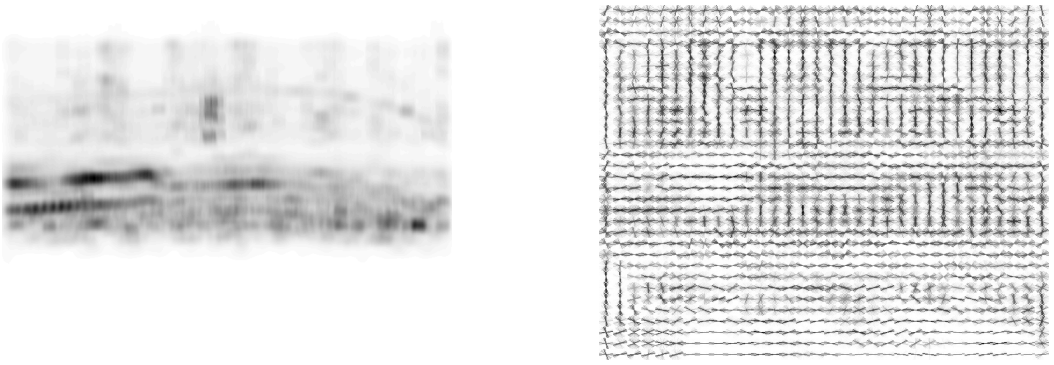

Figure 1: Example of a constant Q transform of an acoustic scene and its histogram of gradient representation.

in an image can be accurately described according to the distribution of its oriented gradients or its edge directions. Hence, such a method describes an image by a set of local histograms. These histograms count occurences of gradient orientation in a local part of the image, defined as cell of the image.

A cell could be defined as a spatial region like a square with a predefined size in pixels. For each cell, we then compute the histogram of gradients by accumulating votes into bins for each orientation. Votes could be weighted by the magnitude of a gradient, so that histogram takes into account the importance of gradient at a given point. When all histograms have been computed for each cell, we can build the descriptor vector of an image concatenating all histograms in a single vector. Figure 1 depicts all the local histogram of oriented edges of the related audio scene.

In our case, we are not interested in objects but in some specific time-frequency signatures in the constant Q representation. We believe that the discriminative power of that representation is brought by the structures and shapes of some of these signatures. That is the main rationale behind the use of histogram of oriented gradients.

3. PROTOCOL AND RESULTS

We have evaluated this feature extractor on the development training examples provided by challenge organisers and we have optimized its parameters according to the performance on this dataset.

The protocol is the following : we have built by random split, a training and testing set which sizes are respectively 80% and 20% of the full dataset. We learn all the binary classifiers of a *one-against-one* strategy [7] after having selected by cross-validation the optimal hyperparameter which weigths the misclassified examples of a linear SVM. Results, using the best parameters of the feature extractor (f_{min}, f_{max}, N_b and N_a) we have found, have been reported in Table 1. From this results, we can note that some classes are rather easy to classify whereas some others are hardly distinguishable. For instance, we can see that the *bus* class is an easy class while the *tube* is frequently confused with *bus* and *tubestation*.

	Prediction									
	1	2	3	4	5	6	7	8	9	10
1	19	0	0	0	0	0	0	0	1	0
2	0	20	0	0	0	0	0	0	0	0
3	0	0	18	0	0	1	0	1	0	0
4	0	0	0	16	0	0	1	3	0	0
5	0	0	5	0	15	0	0	0	0	0
6	0	0	0	1	5	14	0	0	0	0
7	0	0	0	3	0	0	12	5	0	0
8	0	0	0	4	0	1	0	15	0	0
9	4	0	0	2	0	0	0	0	8	6
10	0	0	0	2	0	0	1	1	3	13

Table 1: Sum of the confusion matrices over 10 trials giving an average performance of 75% of accuracy. Predicted classes are given in column. Class number follows the ordering provided in the challenge dataset : 1 = *bus*, ..., 10 = *tubestation*.

4. REFERENCES

[1] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," in *European Signal Processing Conference*, 2011.

[2] C. V. Cotton and D. P. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 69–72.

[3] E. Benetos, M. Lagrange, and S. Dixon, "Characterisation of acoustic scenes using a temporally-constrained shift-invariant model," in *Conference on Digital Audio Effects Conference (DAFx-12)*, vol. 17, 2012, p. 21.

[4] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[5] 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.

[6] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in *Intelligent Vehicles Symposium, 2006 IEEE*. IEEE, 2006, pp. 206–212.

[7] K. Duan and S. Keerthi, "Which Is the Best Multiclass SVM Method? An Empirical Study," in *Multiple Classifier Systems*, 2005, pp. 278–285. [Online]. Available: http://www.keerthis.com/multiclass.html