HOG AND SUBBAND POWER DISTRIBUTION IMAGE FEATURES FOR ACOUSTIC SCENE CLASSIFICATION

Victor Bisot, Slim Essid, Gaël Richard

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, 37-39 rue Dareau, 75014 Paris, France

<firstname>.<lastname>@telecom-paristech.fr

ABSTRACT

Acoustic scene classification is a difficult problem mostly due to the high density of events concurrently occurring in audio scenes. In order to capture the occurrences of these events we propose to use the Subband Power Distribution (SPD) as a feature. We extract it by computing the histogram of amplitude values in each frequency band of a spectrogram image. The SPD allows us to model the density of events in each frequency band. Our method is evaluated on a large acoustic scene dataset using support vector machines. We outperform the previous methods when using the SPD in conjunction with the histogram of gradients. To reach further improvement, we also consider the use of an approximation of the earth mover's distance kernel to compare histograms in a more suitable way. Using the so-called Sinkhorn kernel improves the results on most of the feature configurations. Best performances reach a 92.8% F1 score.

Index Terms— Acoustic scene classification, subband power distribution image, Sinkhorn distance, support vector machine

1. INTRODUCTION

The main objective of acoustic scene classification (ASC) is to identify the acoustic environment in which the sound was recorded directly from the audio signal. The interest for ASC has been increasing in the last few years and is becoming an important challenge in the machine listening community. Despite the somewhat limited performances of current ASC methods, they already have numerous applications in real life such as robotic navigation [1] or forensics [2]. As many context aware devices only use visual information to adapt to their current location, complementary information can be given by analysing the surrounding audio environment.

Due to the large variety of sound events possibly occurring in an audio scene, characterising an acoustical environment as a whole is known to be a difficult problem. Many early works in ASC have tried to use various methods from speech recognition or event classification methods. Moreover, the specificity and complexity of general acoustic scenes cannot be well described by general purpose methods. Indeed, it is now widely recognised that specific methods need to be developed for ASC.

As mentioned above, early works in ASC were heavily inspired by speech recognition systems, for instance features like Mel Frequency Cepstral Coefficients (MFCC) [3] have been widely explored, they are often used as a baseline system for classifying audio scenes. Several other conventional features have also been tested such as low level spectral features (zero-crossing rate, spectral centroid, spectral roll-off) [4], linear predictive coefficients [5] or auditory filter features such as Gammatones [6]. Some other works focused more on designing new features capable of describing the scene as a whole. This leads to more complex features such as expansion coefficients based on a decomposition over a Gabor dictionary [7] or even minimum statistics of a spectrogram to describe the acoustical background of a scene [8]. Many of these features are extracted locally frame by frame which naturally leads to an effort on finding a proper temporal modelling. The temporal information has often been taken into account by using various statistical functions or by analysing the features recurrent behaviours using recursive quantitative analysis (RQA) [9]. In some cases features are extracted from the time frequency representation of the full audio excerpt, for instance features such as the histogram of gradients (HOG) based on constant Q-transform of the complete signal [10].

In this paper we also follow the trend of using features based on a long-term time-frequency representation. Our work significantly improves the state of the art results on a large ASC data set by combining different spectrogram image features and using an adapted kernel for the classification. Specifically, we propose to use the Subband Power Distribution (SPD) as a feature for ASC. The SPD has previously been used to compute features for acoustic event classification [11], it represents the distribution over time of the time-frequency pixels in a spectrogram image in each frequency band. In the ASC context, by computing a histogram of the power spectrum amplitudes in each frequency band we intend to capture the density and the energy of events in a given band for each acoustic scene. The use of the SPD complements the previous proposal of applying Histogram of

This work was partly funded by the European Union under the FP7-LASIE project (grant 607480)

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

Gradients to acoustic scene classification [10]. The HOG features capture the directions of the variations in a spectrogram image. In order to improve the classification we use the HOG features and the SPD simultaneously to characterise both the variations and the content of the scenes power spectrum. We also explore the use of perceptual loudness as an alternative time frequency representation for the extraction of the SPD features. Finally, since a support vector machine (SVM) is used to classify the audio scene classes, we focus on finding a more suitable distance between the features. In order to have a kernel adapted to the classification of histograms, we choose to compute it based on the Sinkhorn distance which is an approximation of the earth mover's distance (EMD) [12].

The rest of the paper is organised as follows. Section 2 describes the SPD feature extraction. Section 3 details the use of the earth mover's distance to design the new kernel. Section 4 describes the data set and our experiments, before Section 5 concludes the work.

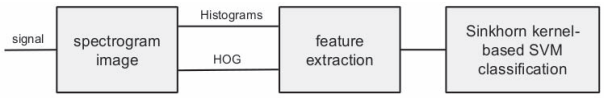

Fig. 1: Flowchart of the proposed acoustic scene classification method

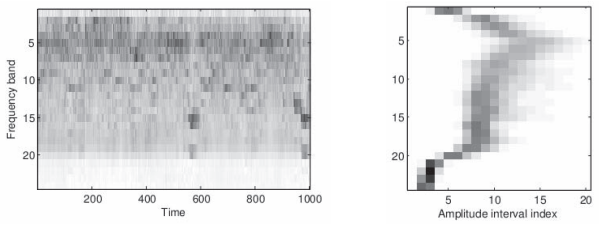

Fig. 2: (Left) Loudness spectrogram of a "kidgame" sample (right) SPD image

2. FEATURE EXTRACTION

2.1. Modelling the event density

The main difficulty in the acoustic scene classification field is the high quantity of information an audio scene contains. In only a few seconds of audio recorded in an urban environment one can find an important number of different sound sources that each contribute to the acoustic signature of the scene. The first supposition made is that all these sound sources correspond to events (such as a car horn) that are characteristic of certain environments (such as a street). One can also suppose that these events have a rather constant spectral distribution and that having a way of identifying how often these spectral distributions happen in a given example would help characterising the different environments. In order to capture the occurrences of these events, or at least of the repeating spectral content, we propose to use the Subband Power distribution image. The SPD image approximates the distribution of the spectrogram amplitudes in each subband using histograms. The SPD will allow us to approximate for a given scene: in what frequency bands the sounds events are, how often they occur and finally how loud they are. Moreover, for scenes that contain constant background sounds across the whole example (such as a car engine), the SPD will also be able to capture this information by having a high value in the bin corresponding to the background sound's amplitude interval.

While we expect the SPD features will provide crucial information for characterising the scenes we suppose they may not be sufficient. Even if the HOG features already give the best results on a few ASC data sets [10], they capture different aspects of the spectrogram image. Because they model the directions of the variations in the time frequency image, having a way of describing the content of the time frequency representation before looking at its variations can aid the classification. The SPD features are not meant to outperform the HOG features alone but rather give complementary information by the concatenation of the two different features. Using the SPD in conjunction with the HOG will prove to achieve a better characterisation of the spectrogram images and will consequently lead to improved classification results.

2.2. The SPD extraction

The SPD as well as the HOG are extracted from a spectrogram image. A constant Q-transform (CQT) is used in [10] to compute the HOG features. The CQT has log-scaled frequency bands which usually provides an appropriate representation for analysing sounds. Instead, we propose to use the perceptual loudness time frequency representation in order to better mimic the human auditory system. The Loudness allows us to have a more similar frequency scale to the human auditory system to model the human understanding of sounds. The loudness coefficients correspond to the energy in each Bark band normalized by its overall sum. Because the choice of using the SPD was motivated by the human comprehension of audio scenes, we believe that using a spectrogram based on the Bark bands will improve the description of the acoustic scenes. Figure 2 shows an example of a SPD image extracted from a loudness spectrogram.

The extraction of the SPD features is similar to the procedure for extracting the HOG as the two descriptors are meant to be combined for the classification. In order to get the SPD we start by computing a spectrogram image of each full audio example in the data set. The SPD features are extracted by computing a histogram of the pixel values in each frequency band of the spectrogram image. We split the pixel value range of the image into a fixed number of amplitude intervals and simply count the number of pixels that are in each amplitude interval for every frequency band. We finally obtain as many histograms as frequency bands initially in the spectrogram,

the concatenation of all the histograms (one per band) will form the feature used a for the classification.

3. USING THE SINKHORN KERNEL FOR SVM CLASSIFICATION

3.1. Changing the kernel

The experiments we run to evaluate the spectrogram features use a support vector machine for the classification. It is common to change the feature space by using a SVM with a non linear kernel when dealing with complex data. For instance, the Gaussian kernel $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ is a widely used kernel in many applications including ASC methods. Instead, we propose to use the earth mover's distance (EMD) [12]. The EMD kernel is known to compare histograms and distributions in a more suitable way than other classic distances. Actually, to avoid the overwhelming complexity of EMD algorithms we use an approximation of the EMD called the Sinkhorn distance [12].

3.2. A distance for histograms

The earth mover's distance (EMD) is a formulation of the optimal transport distances widely used in computer vision. The EMD and the optimal transport distances have proven to be a powerful geometry to compare probability distributions. By fixing a cost function representing the cost of moving information from a histogram bin to another, one can define a new distance between features as the solution of an optimal transport problem. The principal advantage of using the EMD for our application is that we can incorporate prior knowledge about our features by means of the cost function. In fact one can adjust the cost of moving information from an amplitude interval at a fixed frequency to another one. The importance of the frequency position or the amplitude range can be tuned in order to obtain a better discrimination of the classes than with the Gaussian kernel. If we consider M amplitude intervals and N frequency bands the SPD feature H can be written as:

$$H = (h_{f_1 a_1}, ..., h_{f_1 a_M}, ..., h_{f_N a_1}, ..., h_{f_N a_M}) \; ; \qquad (1)$$

where f_i is the index of the corresponding frequency band and a_j the index of the amplitude intervals. We then propose to use the following cost function

$$c(h_{f_1 a_1}, h_{f_1 a_1}) = |f_k - f_i|^p + |a_l - a_j|^q \; ; \qquad (2)$$

where p and q are positive parameters that can adjust the impact of the frequency and the amplitudes when comparing histogram bins. This adjustment can be useful if we feel the data used for classification is better characterized by the presence of information in a frequency band or by the general amplitude distribution of the spectral content in the time frequency image. In our case, the data set described in the following section contains many different acoustic environments each very dense in audio events. The cost function will be kept rather general to allow it to penalise distant frequencies and distant amplitudes.

Finally, we need to compute the cost matrix C necessary to solve the optimal transport problem, the matrix C contains the pairwise costs for each histogram bin couple. When using the feature concatenation we do not want to allow any transfers from the HOG to the SPD. To do so, the cost between a bin from the HOG and a bin from the SPD feature is set to an arbitrarily high value. In this case, the bin couples coming from the same feature type still follow the cost function (2).

3.3. The Sinkhorn kernel for classification

The major downside of using the earth mover's distance is its complexity, even the best implementation is not meant to be used with histogram dimensions over a few hundreds. To avoid such a restriction, we exploit a recent work on optimal transport that offers huge improvement in computation time by adding an entropic constraint to the problem controlled by a regularisation parameter λ [12]. The optimum obtained is also a distance called the Sinkhorn distance which corresponds to an upper bound to the earth mover's distance. Using this faster computation of optimal transport, one is able to solve the optimal transport on the proposed features in a reasonable time. Giving the algorithm the cost matrix C and a regularisation parameter λ for the entropic constraint, the Sinkhorn distances between each feature vector are obtained. The Sinkhorn distances are used to approximate the EMD kernel for the support vector machine classification. Finally, the kernel function k used to define the Sinkhorn kernel can be written as:

$$k(x, y) = e^{-\frac{S(x,y)}{\sigma^2}} . \qquad (3)$$

Where $S(x, y)$ is the Sinkhorn distance between the two feature vectors x and y.

4. EXPERIMENTAL EVALUATION

4.1. The Dataset

We evaluate the spectrogram features and the Sinkhorn kernel on the LITIS Rouen data set for acoustic scene classification [10]. To our knowledge it is by far the largest publicly available data set for ASC. This data set contains 1500 minutes of urban audio scenes recorded with a smartphone, split into 3026 examples of 30 seconds without overlapping and forming 19 different classes. Each class corresponds to a specific location such as in a train station, in an airplane or at the market. Since this data set has been released recently, the only published results were obtained using the HOG features compared to various methods based on MFCC.

	Precision	Recall	F1 Score	Accuracy
Gaussian Kernel with CQT				
[10]	91.7	-	-	-
HOG	91.2	90.2	90.5	91.2
SPD	90.8	89.2	89.7	90.2
HOG+SPD	93.3	92.5	**92.8**	**93.4**
Gaussian Kernel with Loudness				
HOG	90.4	88.4	89.1	89.4
SPD	88.5	87.2	87.5	87.5
HOG+SPD	92.4	91.5	**91.7**	**92.0**
Sinkhorn Kernel with CQT				
HOG	91.4	90.3	90.7	91.3
SPD	88.7	86.9	87.4	88.6
HOG+SPD	92.3	90.6	**91.4**	**92.3**
Sinkhorn Kernel with Loudness				
HOG	92.2	92.0	92.0	92.0
SPD	90.1	89.0	89.2	89.6
HOG+SPD	93.2	92.4	**92.6**	**93.0**

Table 1: Summary of the different experiments comparing the features, the time frequency representation and the kernel for the SVM. The results all have a 0.1 standard deviation.

4.2. Classification protocol

All the experiments use the same training-testing splits suggested by the creators of the LITIS data set to ensure comparable results. All the features and distances we use will follow the same classification scheme using a support vector machine. The Sinkhorn kernel for the SVM is compared to the Gaussian kernel. The results are averaged over 20 train-test splits of the data. In each split 80% of the examples are kept for training and the rest are for testing. In order to estimate the best regularisation parameter and the best σ for the Gaussian kernel we perform a grid search on these parameters for each cross-validation iteration. To do so, we split evenly the training set 5 times into a learning and validation set and we keep the parameters giving the best average result on the 5 validation subsets.

4.3. Comparing the addition of SPD features

In the first part of our evaluation we look into the benefits of using the concatenation of HOG and the SPD as the features for the classification. The HOG features are extracted from a time frequency representation image of the whole scene resized to 512×512 by bicubic interpolation. Signed orientations as well as a frequency pooling are used to compute the HOG. More details about the HOG possible settings are given in [10]. To compute the SPD we do not re-size the time frequency representation as the samples in the data set are all of equal length. The best results have been found using 20 bins

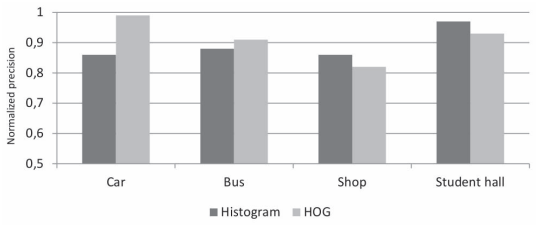

Fig. 3: F1 score on 4 scene classes obtained with the loudness and the Sinkhorn kernel

corresponding to 20 linearly spaced amplitude intervals in the pixel value range. All the features have been preprocessed in order to have zero mean and unit variance.

Table 1 summarises the results obtained and shows that the performances depend on the time frequency representation and on the kernel used for classification. Most importantly, the best F1 score on all the proposed settings is obtained with the feature concatenation. This supports the assumption that the SPD features generally give complementary information to the classifier as they do not describe the same phenomena as the HOG. The best result obtained so far is a 92.8% F1 score using the concatenation of the two features. Previous state of the art on this data set was of 91.7% precision using only the HOG features with different settings while a 93.3% precision is reached with the feature concatenation. The difference between the best precision results obtained with concatenation and the previous best precision score is statistically significant at 5%. In order to understand the differences between the two features we show the F1 score obtained on a few of the classes in Figure 3. HOG features have good performances on classes such as bus or car (and all other public transports). These locations often do not have a located frequency signature because of the presence of acceleration sounds in the examples. Because the HOG are designed to capture evolutions in the spectrogram, they are expected to work well on classes containing acceleration sounds. On the other hand, the SPD features outperform the HOG on classes such as shop or student hall. Both of these classes have mostly a stable frequency signature over time due the high density of events occurring during the recorded examples.

4.4. Using the Sinkhorn kernel

The EMD kernel for the classification has been computed using the light-speed implementation of the Sinkhorn distance using the cost function described in (2). The cost matrix has been divided by its median value and the regularisation parameter λ is set to 11 as it consistently gives better results. We tested different values of the parameters p and q in (2) for each feature set, the results in Table 1 are obtained with the best parameters we found so far. First, we see that the Sinkhorn distance gives similar results to the Gaussian kernel on the concatenated features. The way the concatenation is

taken into account in the cost function could explain the lack of improvement. Having two different kernels corresponding to the HOG and the SPD could lead to better results using multiple kernel learning but would bring even more parameters to tune.

Using the Sinkhorn kernel improves the F1 score for four out of the six feature configurations tested in Table 1 but it does not help improving the overall best result on the data set. The results are not yet worth the increased complexity compared to the Gaussian kernel. Although we could possibly increase the performance by focusing more on the parameter tuning it would still be hard to use on much larger data sets despite using the lightspeed implementation.

4.5. The Loudness spectrogram instead of the CQT

The previous experiments were tested with two different time frequency representations, the constant Q-transform, initially used to compute the HOG and the loudness power spectrum as discussed in Section 2. The CQT is extracted using a frequency range from 1Hz to 10kHz using 8 bins per octave. The perceptual loudness power spectrum has 24 frequency bands and is extracted using the YAAFE implementation [13]. The results show that using the loudness slightly improves the F1 scores compared to the CQT for all the features tested with the Sinkhorn distance kernel. On the other hand, it does not help the Gaussian kernel SVM exept for the HOG feature. The loudness power spectrum also leads to features with lower dimension (because of the reduced amount of frequency bans) which helps lowering the computation time of the Sinkhorn distance algorithm and of the classification in general.

5. CONCLUSION

In this paper we proposed to use the Subband Power Distribution image as a feature for acoustic scene classification leading to a more robust representation of the scene when used jointly with the HOG features. An experiment run on the largest available data set proved that adding the SPD to the HOG significantly improves the state of the art results. We also discussed the interest of using the earth mover's distance to provide a more suitable distance between our features. The Sinkhorn kernel does not offer a consistant increase in performance compared to the Gaussian kernel. Finally we looked into using a loudness spectrogram instead of a constant Q-transform in order to model the human auditory system. The loudness does provide slightly better results when used with the Sinkhorn kernel.

REFERENCES

[1] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 885–888.

[2] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected mpeg-7 audio features and mel-frequency cepstral coefficients," in *Fifth International Conference on Digital Telecommunications (ICDT)*, 2010.

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

[4] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[5] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[6] X. Valero and F. Alías, "Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

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

[8] S. Deng, J. Han, C. Zhang, T. Zheng, and G. Zheng, "Robust minimum statistics project coefficients feature for acoustic environment recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 8232–8236.

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

[10] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene detection," Tech. Rep., HAL, 2014.

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

[12] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, 2013, pp. 2292–2300.

[13] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software.," in *ISMIR*, 2010, pp. 441–446.