ENHANCED LBP TEXTURE FEATURES FROM TIME FREQUENCY REPRESENTATIONS FOR ACOUSTIC SCENE CLASSIFICATION

Shamsiah Abidin[a], Roberto Togneri[a], Ferdous Sohel[b]

[a]School of Electrical, Electronics and Computer Engineering, The University of Western Australia, Perth 6009, WA, Australia
[b]School of Engineering and Information Technology, Murdoch University, Perth 6150, WA, Australia

ABSTRACT

This paper introduces the use of local binary patterns (LBP) extracted from a time-frequency representation (TFR) for acoustic scene classification. As LBP provides a description of the global TFR texture we propose a novel zoning mechanism that provides a simple solution to extract spectrally relevant local features which better characterize the audio TFRs. To further improve the classification performance, we perform feature and score level fusion of the proposed LBP (with zoning) with histogram of gradients (HOG) of the TFR images. Our technique demonstrates an improved performance by achieving a classification accuracy of 95.2% using a fusion of time-frequency derived features.

Index Terms— acoustic scene, local binary patterns, feature extraction, time-frequency analysis, fusion

1. INTRODUCTION

For the past two decades, there has been much interest in research related to acoustic scene classification (ASC) due to its significance in various emerging applications such as automatic audio surveillance, robotics sensing, multimedia content analysis and machine listening. The objective of acoustic scene classification is to recognize the environment in which an audio stream has been produced. The importance of machine listening has been highlighted by Stowell et al. [1] and suggests that intelligent machine listening should have the capability to automatically recognize the scene based on the acoustic information.

Early research on ASC widely used speech perceptual features based on the human auditory system. The pioneering work on acoustics scene classification by Sawhney and Maes used power spectral density and Gammatone filterbank analysis which mimic the response of the human cochlea [2]. In the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) for acoustic scene classification [1][3], the Mel-frequency cepstral coefficient (MFCC) was a popular choice for feature extraction. Other acoustic features such as zero crossing rate, frequency-band energy, spectral centroid, Mel-scaled filter-bank coefficients and Gabor representation have also been explored for ASC [4][5][6]. In addition to acoustic features, Heittola et al. [7] suggested the use of event histograms for scene classification. These features are used in conjunction with different machine learning methods such as support vector machines (SVM), Gaussian mixture models (GMM) and random forests to capture and classify the temporal and spectral variations that constitute a scene [3]. The best reported work to date is that of Bisot et al. [8] who used unsupervised learning with non-negative matrix factorization. Their reported F1-score was 95.6% using the kernel extension for Principal Component Analysis (PCA), which is the best classification result on the LITIS Rouen dataset.

As an alternative approach, Rakotomamonjy and Gasso [9] applied techniques from image processing to acoustic scene analysis. Audio signals were first converted to a TFR from which HOG features were extracted [9]. Dealing with the non-stationarity of acoustic signals is a known issue in acoustic scene classification. Rakotomamonjy and Gasso addressed this by pooling HOG features to obtain histograms that are able to capture and preserve the information of the time-frequency structures which better characterize an audio TFR.

In this work we investigate LBP as features we can extract from the TFR image. The LBP is highly discriminative for texture patterns and computationally efficient [10]. LBP provides a global representation, however for scene classification tasks, a localized view is more important and to address this we propose a novel zoning mechanism for LBP. By zoning the TFR into non-overlapping uniform slices we extract local information which better highlight distinctive spectral regions of the scene. These local LBP histograms in each zone are extracted and concatenated to form an enhanced LBP feature vector. As the LBP features with zoning provides complementary information to HOG features with pooling we present the results from a simple feature and score level fusion of the LBP and HOG features. We then extend this to a multiresolution LBP analysis which when fused with HOG

features provides one of the best performing results on the LITIS Rouen dataset.

The remainder of the paper is organized as follows: Section 2 explains the proposed framework, Section 3 describes the experimental setup, result and analysis and we provided our conclusions in Section 4.

2. THE PROPOSED FRAMEWORK

Our proposed framework uses a time-frequency representation of the acoustic signals for feature extraction. Figure 1 illustrates the conceptual framework for the proposed system. The audio signals are converted to a log scale time-frequency representation using the Constant-Q Transform (CQT). Considering the CQT time-frequency representation as a texture pattern, we have chosen the LBP operator for feature extraction. In this work, we provide details of our novel TFR zoning mechanism for LBP features. The LBP histograms in each zone are extracted and concatenated to form an enhanced LBP feature vector. A multiresolution LBP texture analysis has been introduced into the proposed zoning mechanism to provide fine-grained analysis of the TFR texture patterns. The multiresolution LBP features are then fused with HOG features.

2.1. Time Frequency Representation (TFR)

The time frequency representations are computed using the CQT. The CQT transforms a time-domain acoustic signal into the time-frequency domain and provides a log scale frequency resolution that is approximately similar to auditory perception providing a finer resolution at lower frequencies. The CQT transform is computed by means of the Constant-Q transform [11] toolbox. The CQT TFR is resized to 512 × 512 using bicubic interpolation which based on [9] to obtain a uniform size of the TFR independent of the signal length, sampling frequency and CQT parameters. Table 1 shows the CQT parameters used to compute the TFR.

Table 1: CQT parameter

Q factor, $Q = 1$	
Sampling frequency, *Fs*	22050 Hz
Bins per Octave, *B*	8
Maximum frequency, *Fmax*	10000 Hz

2.2. LBP features

LBP is a state-of-the-art feature extraction method for analyzing image textures [10] due to its robustness to grayscale variation and computational simplicity. The CQT TFR patterns represent image textures suitable for applying LBP feature extraction to acoustic scene analysis.

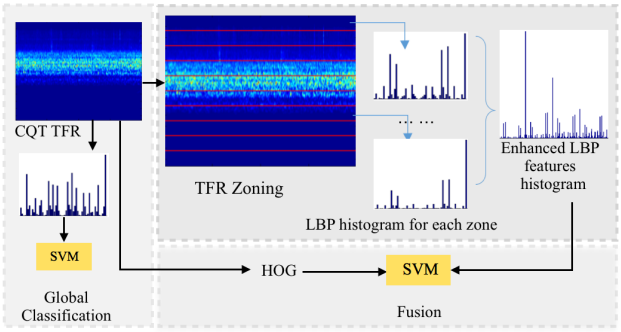

Figure 1: Conceptual framework of the proposed system

As given by Eq. (1):

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_{p,R} - g_c)2^p \ , \qquad (1)$$

where:

$$s(x) = \begin{cases} 1, if\ x \geq 0 \\ 0, otherwise \end{cases}$$

LBP considers each pixel of an image and is calculated by comparing the grayscale value of each central pixel, g_c, with the grayscale value of its P neighboring pixels, $g_{p,R}$ for $p = 0,1, \ldots P-1$, where R, is the radial distance between the central pixel and each neighboring pixel. The LBP feature provides a P-bit binary encoding of the neighboring pixels whose grayscale value g_p is larger than the central pixel, g_c. The results are returned as the local binary pattern histogram as shown by Figure 1.

LBP extensions such as multiresolution analysis can further enhance the LBP operator performance [12] to provide fine-grained analysis. We present a straightforward method for combining operators of different spatial resolutions. In this multiresolution analysis, two resolutions are considered. We performed preliminary experiments by varying the values of P and R. We chose the two best performing common (P, R) pair values for a uniform LBP of (8,1) and (12,2). These uniform LBP feature analyses for each scene yields 59 dimensions for $\mathrm{LBP}_{8,1}$ and 135 dimension for $\mathrm{LBP}_{12,2}$.

2.3. Zoning

From our analysis of the CQT TFR, we have noticed that the texture patterns are not uniform. Therefore, the use of a 'global' LBP may not sufficiently represent the local non-uniformity. For LBP it is important to capture and properly preserve local features. Analyzing the TFR as a global feature will average out the distinctive regions of the TFR. The underlying idea of our proposed solution is to identify local features pertinent to the relevant spectral information. In Figure 2 we notice that at low frequencies the textures are quite similar but they change as the frequency increases.

We also observe that, for example, the high-speed train and café scenes, the textures at frequencies < 30 Hz are indistinguishable. However, the textures between 30 Hz and 460 Hz for the high-speed train (a), (b) are similar and quite different from the distinctive textures between 60 Hz to 7711Hz for café (a), (b).

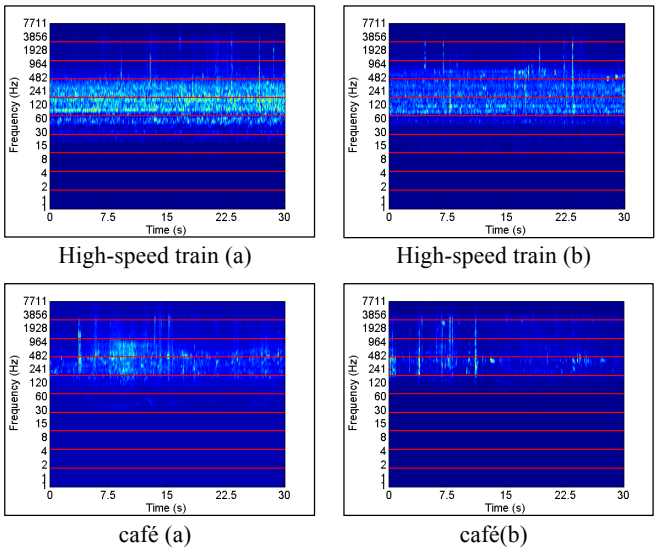

Figure 2: CQT representation for acoustic scenes with linear zoning

To address this, we propose a zoning technique to obtain the local information of a given pattern by dividing the TFR into a number of zones, *n,* to be analyzed locally as shown in Figure 3. The zoning size, z is dependent on the TFR height, h and the number of zones, n and is given by $z = h/n$. The number of zones, *n,* was varied empirically to get the best result. In our work we found $n = 10$ and $n = 15$ provide the optimal number of zones for the (8,1) and (12,2) resolutions respectively.

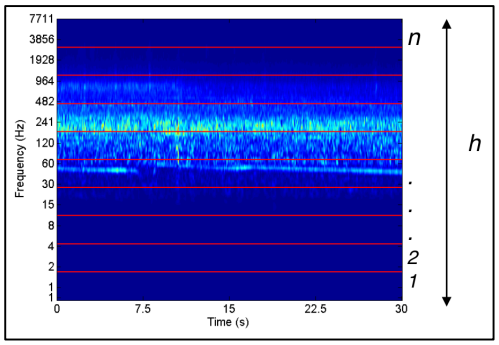

Figure 3: A Linear zoning mechanism is used to extract local information

Let Xn represent the CQT pixels in the nth zone which is a $z \times 512$ image, where $h = 512$ and $z = 512/n$. We carry out an independent LBP on Xn by using Eq. (1) followed by concatenation to form an enhanced LBP feature vector as below:

$$\mathrm{LBP}_{\mathrm{enhanced}} = [\mathrm{LBP}_{X1}\ \mathrm{LBP}_{X2} \cdots \mathrm{LBP}_{Xn}] \qquad (2)$$

From Eq. (2) the number of dimensions was increased by a factor of n.

Our proposed zoning is similar to the pooling of HOG used by [9] and was first introduced using LBP features for music genre classification in [13]. In that work a Mel-spaced zoning mechanism was used with short-time Fourier transform features and only provided a 3% improvement in performance.

2.4. Fusion

Fusion is a common method for improving recognition rates when features are complementary. The LBP and HOG features emphasize different characteristics of the TFR. LBP deals with pixel intensity and texture, and provides the uniform pattern representation of the CQT. Whereas, HOG deals with the distribution of gradients at different orientations making it more suitable to deal with random textures. We fused the multiresolution LBP and HOG features to complement the different textures of the CQT pattern by both feature level fusion and score level fusion. The HOG features have been obtained using the best performing features from [9]. For feature level fusion, we fuse the LBP and HOG features by concatenating the HOG directly with the LBP. For score level fusion, the decision is based on the sum rule [14]. An advantage of score level fusion is that it does not increase the dimensionality of the features as feature level fusion, but does require two separate classification systems.

3. EXPERIMENTS

3.1. Experimental setup

Our experiments were evaluated on the LITIS Rouen dataset [9]. This is the largest well-known publicly available dataset for ASC. The dataset contains 3026 audio files of 19 scene classes with 30 seconds duration recordings for each scene example. For classification, the SVM classifier is implemented using the LIBSVM toolkit [21]. The "one-against-one" approach for multiclass classification was used with a linear kernel. The dataset was divided into 80% of training scenes examples and 20% of test scenes over 20 averaged random trials. We use accuracy and precision as the evaluation metrics. The evaluation is calculated for each class i, where $\{i = 1, 2, \ldots, k\}$ and k is the number of classes. The accuracy, A, is calculated as the number of total correct scenes, c, divided by the total number of test scenes, N. For the precision metric, the calculations are based on the confusion matrix: the rows provide the instances of the actual class and the columns are the instances of the predicted class. The precision, P_i, averaged over all classes, is defined as the number of correctly predicted scenes, c_i,

divided by the sum of correctly predicted scenes, c_i and the number of false positives, fp in the class. Accuracy and Precision are defined as below:

$$\text{Accuracy, A} = \frac{c}{N} = \frac{\sum c_i}{N} \quad ; \qquad \text{Precision, } P_i = \frac{c_i}{(c_i + fp)}$$

3.2. Experimental result and analysis

Several experiments were conducted to show the benefits of applying LBP features and the influence of the proposed zoning mechanism. As the baseline, we first evaluated the accuracy of LBP features of the global TFR without applying the zoning mechanism. Both the accuracy and precision are only 72% as it only considers one global feature and therefore some time-frequency information is lost. Then we apply zoning in order to summarize LBP local features while preserving relevant frequency information over the specified time. Table 2 presents the results obtained after empirically testing different numbers of TFR zoning and LBP parameters. From Table 2 it can be seen that the proposed zoning of the LBP features provided a significant improvement from 72% to 91.5% in accuracy and precision. Therefore, from the results obtained it can be observed that the zoning mechanism is instrumental in extracting more discriminative spectral detail for ASC.

As different LBP resolutions provide complementary information, we carried out a multiresolution analysis by concatenating the $LBP_{8,1}$ features with the $LBP_{12,2}$ features. The multiresolution LBP performance in Table 3 yielded a 1% improvement in accuracy compared to the single resolution LBP. This is because the original LBP features are calculated in a local 3×3 neighborhood and cannot capture large scale structures [10]. By combining 2 LBP operators, 2 different LBP codes are assigned to each pixel in the TFR. Joint distribution of these codes resulted in more accurate information of the acoustic scene. Our proposed method shows the result of better precision (92.7%) compared to [9]. However this does come at the cost of a doubling in the feature dimension.

In Figure 4 we show the accuracy of the feature fusion of multiresolution of LBP with HOG compared to HOG and LBP only for the individual scenes. From Figure 4 the individual features had difficulties in discriminating scenes such as café, metro-Paris and pedestrian street. It was reported in [9] that the HOG feature was not able to totally capture the fine-grained discriminative features between some scenes which are difficult to distinguish such as metro-Paris, metro-Rouen, quiet street and pedestrian street. However, with the fusion of LBP with HOG, the accuracy for each scene class has been significantly improved. It is well known that HOG is excellent at capturing edges and corners in images. On the other hand, LBP is better at capturing the local patterns. As HOG and LBP emphasize different capabilities in image analysis, critical for ASC, this explains the successful fusion of these features in providing complementary information able to boost performance by up to 3% over LBP and HOG alone.

Table 2: Classification result of linear zoning

Features	Linear Zoning			
	# zone	Dimension	Accuracy (%)	Precision (%)
$LBP_{8,1}$	15	880	91.5	91.5
$LBP_{12,2}$	10	1350	90.5	90.4
HOG		1536	91.9	91.8

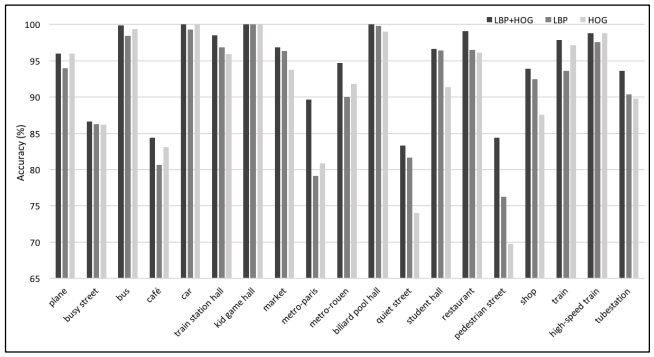

Figure 4: Accuracy result for each scene classes

The comparison of results between the application of feature level fusion and score level fusion is shown in Table 3. Simple concatenation of feature level fusion provides better results compared to the sum rule for score level fusion. The best accuracy and precision obtained is 95.2% and 95.1% respectively. These results are superior to the results reported by using HOG features alone [9] and comparable with the work in [8] using kernel non-negative matrix factorization over a sequence of CQT time slices.

Table 3: Result for features fusion

Features	Feature Level Fusion		Score Level Fusion	
	Accuracy (%)	Precision (%)	Accuracy (%)	Precision (%)
$LBP_{8,1}$+ $LBP_{12,2}$	92.9	92.7	92.4	92.4
$LBP_{8,1}$ + $LBP_{12,2}$ + HOG	**95.2**	**95.1**	94.9	94.9

4. CONCLUSION

This work has demonstrated the capability of LBP for time-frequency analysis for acoustic scene classification and the state of the art performance using time-frequency representations. To increase the classification accuracy, linear zoning and multiresolution LBP was introduced for more localized features. Furthermore, fusion with HOG features has provided the comparable reported result for ASC using time-frequency representations. In future we will consider other time-frequency representations and image-based visualization features.

5. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1713–1746, 2015.

[2] N. Sawhney and P. Maes, "Situational Awareness from Environmental Sounds," *Proj. Rep. Pattie Maes*, pp. 1–7, 1997.

[3] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification," *Signal Process. Mag. IEEE*, vol. 32, no. 3, pp. 16–34, 2015.

[4] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental Sound Recognition With Time- Frequency Audio Features," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, 2009.

[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 Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, 2006.

[6] B. Clarkson, N. Sawhney, and A. Pentland, "Auditory Context Awareness via Wearable Computing," in *In Proceedings of The 1998 Workshop On Perceptual User Interfaces*, 1998, pp. 1–6.

[7] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio Context Recognition Using Audio Event Histograms," in *European Signal Processing Conference*, 2010, pp. 1272–1276.

[8] V. Bisot, R. Serizel, S. Essid, and Gael Richard, "Acoustic Scene Classification With Matrix Factorization for Unsupervised Feature Learning," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6445–6449, 2016.

[9] A. Rakotomamonjy and G. Gasso, "Histogram of Gradients of Time–Frequency Representations for Audio Scene Classification," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 142–153, 2015.

[10] Topi Maenpaa and M. Pietikainen, "Texture Analysis With Local Binary Patterns," in *Handbook of Pattern Recognition and Computer Vision*, 2004, pp. 1–20.

[11] C. Schörkhuber and A. Klapuri, "Constant-Q Transform Toolbox for Music Processing," in *7th Sound and Music Computing Conference*, 2010, pp. 3–64.

[12] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[13] Y. M. G. Costa, L. S. Oliveira, a. L. Koerich, F. Gouyon, and J. G. Martins, "Music Genre Classification Using LBP Textural Features," *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, 2012.

[14] J. Kittler, M. Hater, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.