ACOUSTIC FEATURE EXTRACTION BY STATISTICS BASED LOCAL BINARY PATTERN FOR ENVIRONMENTAL SOUND CLASSIFICATION

Takumi Kobayashi and Jiaxing Ye

National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Japan

ABSTRACT

Classification of environmental sounds is a fundamental procedure for a wide range of real-world applications. In this paper, we propose a novel acoustic feature extraction method for classifying the environmental sounds. The proposed method is motivated from the image processing technique, local binary pattern (LBP), and works on a spectrogram which forms two-dimensional (time-frequency) data like an *image*. Since the spectrogram contains noisy pixel values, for improving classification performance, it is crucial to extract the features which are robust to the fluctuations in pixel values. We effectively incorporate the local statistics, mean and standard deviation on local pixels, to establish robust LBP. In addition, we provide the technique of L_2-Hellinger normalization which is efficiently applied to the proposed features so as to further enhance the discriminative power while increasing the robustness. In the experiments on environmental sound classification using RWCP dataset that contains 105 sound categories, the proposed method produces the superior performance (98.62%) compared to the other methods, exhibiting significant improvements over the standard LBP method as well as robustness to noise and low computation time.

Index Terms— environmental sound, classification, spectrogram, local binary pattern

1. INTRODUCTION

In daily life, there are a variety of sounds generated by not only human but also the environmental objects. Those environmental sounds contain rich information for understanding the surrounding situations; for example, door knocking sounds imply that someone comes to the house. Thus, the recognition of the environmental sounds endows a wide range of real-world applications, such as acoustic surveillance.

While the methods for classifying speech and music have been intensively developed for decades, those for the environmental sounds are studied with keen attention in recent years [1, 2, 3, 4]. The environmental sounds are different from the speech and music in that the acoustic signals are not stationary nor well-structured; characteristics in these types of sounds are discussed in [5]. Therefore, the conventional speech recognition methods might not be suitable for the environmental sounds [1, 2, 6], and much research effort has been made especially on the feature extraction that effectively characterize them.

Cowling et al. [1] extensively investigated various types of acoustic features as well as classification methods, concluding that the continuous wavelet transform with dynamic time warping produces the promising performance. Wavelet packet analysis is applied to extract acoustic feature in generalized sound classification [7]. Recently, the method of matching pursuit that sparsely decomposes the signal by using over-complete dictionaries has been successfully applied to classify the environmental sounds [2, 5]. As in the matching pursuit, the non-negative matrix factorization (NMF) also works on sparse factorization of signals with learning the dictionary; Cotton and Ellis [8] employ the NMF to construct acoustic event-based patch features from a spectrogram. Ye et al. [3] utilize the acoustic subspace extracted from sound clips in the kernel-based framework.

On the other hand, once an acoustic signal is transformed into a spectrogram which forms two-dimensional (time-frequency) data like an *image*, the visual recognition methods are applicable to extract the acoustic features. Guo et al. [6] extract the time-frequency intersection patterns from the spectrogram similarly to image projection profile. Dennis et al. [4] establish the moment-based spectrogram features.

In this study, we propose a novel method to extract acoustic features from a spectrogram. The proposed method is based on the image recognition technique, local binary pattern (LBP) [9], to capture (local) characteristics on the spectrogram. The LBP method exhibits high discriminative power for classifications, but it is highly affected by the fluctuations in pixel values, though Costa et al. [10] employ the LBP for music genre classification. Thus, we effectively incorporate the local statistics, mean and standard deviation on local pixels, into the LBP so as to improve the robustness to those fluctuations frequently observed in the spectrogram *images*. The proposed statistics-based LBP efficiently exploits the local geometric characteristics in the spectrogram for classifying the environmental sounds with high robustness. In addition, the technique of L_2-Hellinger normalization is applied to the proposed features so as to further enhance the discrim-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

inative power as well as robustness. The proposed method is tested on RWC dataset that contains 9,772 samples in 105 environmental sound categories, exhibiting favorable performance compared to the other methods.

2. ACOUSTIC FEATURE EXTRACTION

The LBP [9] has produced promising performance especially in image classifications. We develop it by incorporating the local statistics to establish the acoustic feature extraction method that is robust to fluctuations on an audio spectrogram. In the followings, let the spectrogram be represented by the image $I(\boldsymbol{r}), \boldsymbol{r} = [t, f]$ on the time-frequency (discrete) domain; the pixel value $I(\boldsymbol{r})$ indicates the power of the frequency f at the time t.

2.1. Proposed method

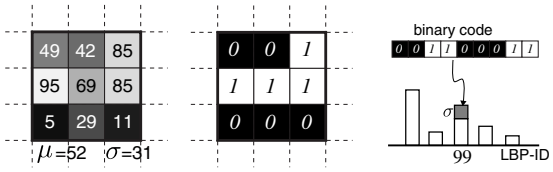
By incorporating statistics of local pixel values into LBP, the local patterns are effectively characterized with robustness to fluctuations in pixels. The LBP [9] always categorizes the local patterns into one of the (binary) codes, no matter how the patterns are less significant. That is, the coding procedure in the LBP takes into account only the magnitude relation between the pixel values in disregard of the difference (margin) between them. However, the fluctuations on the pixels whose values are close to each other easily change the binary code, breaking up the magnitude relation. Thus, the binary codes on those pixels of small margins are vulnerable to pixel fluctuations, resulting in unstable LBP features.

In this study, we exploit the local statistics of pixel values for extracting the stable features with high robustness to the fluctuations, as shown in Figure 1. As in LBP, the proposed method operates on a local patch; let $\boldsymbol{c} \in \mathbb{R}^2$ denote the center position in the patch and $\mathcal{L}$ be the set of pixels in the local patch centered at $\boldsymbol{c}$. First, the mean pixel value in the local patch, $\mu_{\boldsymbol{c}} = E_{\boldsymbol{r}\in\mathcal{L}}\{I(\boldsymbol{r})\}$, is employed for partitioning the pixel values into binary codes.

$$\forall \boldsymbol{r} \in \mathcal{L},\ \ \texttt{code}(\boldsymbol{r}; \mu_{\boldsymbol{c}}) = \begin{cases} 1 & \text{if } I(\boldsymbol{r}) > \mu_{\boldsymbol{c}} \\ 0 & \text{if } I(\boldsymbol{r}) \le \mu_{\boldsymbol{c}} \end{cases}. \quad (1)$$

Thresholding by the mean $\mu_{\boldsymbol{c}}$ is robust to local variations in contrast to the standard LBP that directly employs the center pixel value $I(\boldsymbol{c})$ as the threshold. Note that in (1) the center pixel $\boldsymbol{c}$ is also encoded and the feature dimensionality results in $2^{|\mathcal{L}|}$; there are binary states in each of $|\mathcal{L}|$ pixels.

Next, we measure the significance of the local pattern based on the standard deviation in the patch, $\sigma_{\boldsymbol{c}} = \sqrt{E_{\boldsymbol{r}\in\mathcal{L}}\{(I(\boldsymbol{r}) - \mu_{\boldsymbol{c}})^2\}}$. The standard deviation is regarded as the averaged margin between the pixel values and the mean $\mu_{\boldsymbol{c}}$. The local pattern of large $\sigma_{\boldsymbol{c}}$ is stable since the magnitude relation in (1) is rarely broken up by the fluctuations (Figure 2). Therefore, the standard deviations reflect the significances of the local patterns and we employ them for



(a) local patch (b) binary code (c) weighted voting

Fig. 1. Proposed method. Each local patch (a) is encoded into binary code (b) by comparing to the mean ($\mu = 52$) and votes the weight of $\sigma = 31$ to the pattern code histogram (c).

(a) low significance (b) high significance

Fig. 2. Local statistics. The black dot indicates the pixel value in the local patch. The pattern of the large standard deviation σ is stable since the magnitude relation to the mean μ is rarely broken up by fluctuations in pixel values.

weighting the binary codes. In the proposed method, the local patch is scanned over the region of interest $\mathbb{D}$ to accumulate the weight $\sigma_{\boldsymbol{c}}$ of the binary patterns into the histogram which is the final feature vector describing the image region $\mathbb{D}$;

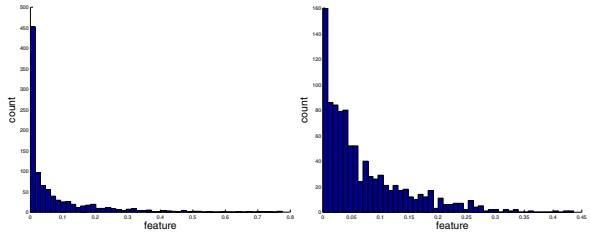
$$x_i = \sum_{\boldsymbol{c}\in\mathbb{D}} [\![\texttt{code}(\mathcal{L}; \mu_{\boldsymbol{c}}) = i-1]\!]\sigma_{\boldsymbol{c}}, \ \ i \in \{1, \cdots, 2^{|\mathcal{L}|}\}. \quad (2)$$

where x_i is the i-th proposed feature, code is the coding operator (1) to produce $|\mathcal{N}_{\boldsymbol{c}}|$ bit code on the basis of $\mu_{\boldsymbol{c}}$ (see Figure 1), and $[\![\cdot]\!]$ indicates the identity function that equals to 1 if the equation in the brackets holds and to 0 otherwise. Through weighting by the standard deviation, the significant patterns are favorably counted, while the less-significant ones hardly contribute to the feature.

2.2. Properties of the proposed method

The proposed feature is robust to (constant) additive variations as in LBP since the local magnitude relations between pixel values and the mean are also invariant to such variations. On the other hand, the multiplicative variation slightly affects it via the weighting by the standard deviation. Those variations, however, are suppressed by applying the normalization described in the next section.

The proposed method effectively extracts the geometrical characteristics on the spectrogram. From geometrical viewpoint, the two-dimensional spectrogram is composed of gradients (lines) and curvatures (corners). Those fundamental characteristics are represented by the local binary patterns describing how the pixel values are distributed in the local patch. The patterns are weighted by the standard deviation in a manner similar to interest point detectors [11] that pick up the geometrically distinct points. Thus, the geometrical characteristics, not only lines but also corners, are efficiently exploited



(a) L_2 (b) L_2-Hellinger

Fig. 3. Distribution on the proposed features normalized by (a) L_2 normalization and by (b) L_2-Hellinger normalization.

by the proposed method, which enables us to discriminatively distinguish the two-dimensional spectrogram.

2.3. L_2-Hellinger normalization

The proposed method produces the histogram feature which is regarded as a discrete probability distribution over the binary patterns. The Hellinger (Bhattacharya) kernel is effective for measuring the similarity between probability distributions [12]. Such kernel can be embedded in a (linear) dot product of the feature vectors normalized by the following L_2-Hellinger normalization [13]; $\hat{\boldsymbol{x}} = \sqrt{\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}}$. Note that the proposed features are non-negative and the normalized feature vector $\hat{\boldsymbol{x}}$ has a unit L_2 norm ($\|\hat{\boldsymbol{x}}\|_2 = 1$); linear SVM effectively works on the features that have a unit L_2 norm [14].

The L_2-Helinger normalization enhances the discriminative power of the features. In the proposed method, only a small amount of binary patterns are activated with high significance (σ), which results in peaky feature distribution around zero; Figure 3a shows the distribution of a certain component of the features simply normalized in a unit L_2 norm. By applying the L_2-Hellinger normalization, the feature distribution becomes favorably dispersed as shown in Figure 3b, and thereby the features can be classified more discriminatively.

3. SOUND CLASSIFICATION SCHEME

In this section, we describe the procedure to classify the environmental sounds by using the proposed features; the overall flow is shown in Figure 4.

An input acoustic signal is first processed by applying short term Fourier transform (STFT) with time window of the length τ to produce the spectrogram. The time-frequency spectrogram is viewed as the 2-D image on which we extract the proposed features (Section 2.1) at each frequency bin via summation along the time axis, *i.e.*, by setting $\mathbb{D}_f = \left\{\boldsymbol{r} = (t', f') \middle| \forall t', f' = f\right\}$ to produce frequency-wise feature vector $\boldsymbol{x}(f)$ in (2). The temporal information is marginalized out in order to make the feature invariant to temporal shift, while the local patterns characterize the local temporal dynamics which are effective clues to classify the non-stationary sound.

To reduce the frequency dimensionality, we subsequently apply the simple filter banks that are equally spaced along the frequency axis. Suppose we have K filter banks denoted by $\omega_k(f), k = 1, .., K$, and the acoustic features are obtained as

$$\bar{\boldsymbol{x}}_k = \sum_f \omega_k(f)\boldsymbol{x}(f), \quad \bar{\boldsymbol{x}} = [\bar{\boldsymbol{x}}_1^\top, \cdots, \bar{\boldsymbol{x}}_K^\top]^\top. \tag{3}$$

Then, L_2-Hellinger normalization (Section 2.3) is applied to $\bar{\boldsymbol{x}}$, followed by the linear SVM classification [15]. To cope with the multi-class problems, the one-versus-rest classification approach is employed in this study.

4. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed method on environmental sound classification by using RWCP dataset [16].

The RWCP dataset [16] contains 9,772 sound clips in 105 categories, such as *hand clapping* and *bell ringing*; there are some similar sound categories, *e.g.*, three types of *phone bell*, which makes the classification quite challenging. The acoustic signals are recorded by 48kHz with 16bit resolution; the averaged recording length is about 1 sec. We set the analysis window length in STFT by $\tau = 512$ frames (about 10 msec) with half overlapping, and we use $K = 50$ filter banks. The performance is measured by classification accuracy averaged over 10-fold cross validations.

4.1. Performance analysis

We first investigate the effect of the settings in the proposed method on the classification performance.

Local patch $\mathcal{L}$. The size of local patch $\mathcal{L}$ defines the local patterns that we extract and consequently the dimensionality of the features. The performance results on various patch sizes are shown in Table 1. We can see that the performance increases along the frequency axis more than along the time axis. This result shows that the local relationship among the frequencies is more important for classifying the environmental sounds. On the other hand, too large local patch deteriorates the performance, since the feature of such patch captures too detailed variations of spectrogram patterns, degrading the generalization performance. The dimensionality of the feature is exponential with respect to the area size of the patch, and the patches whose area sizes are around 9 pixels exhibit better performance. The best performance is obtained on 2×4 (time $\times$ freq.); in the following experiments, we use the local patch of 2×4.

Normalization. We then compared the L_2-Hellinger normalization to the other types of normalizations; L_1 and L_2 normalizations. The performance comparison is shown in Table 2a. The L_2-Hellinger significantly improves the performance by enhancing the discriminative power of the feature as described in Section 2.3. The commonly used L_2 normalization outperforms L_1, but is inferior to L_2-Hellinger.

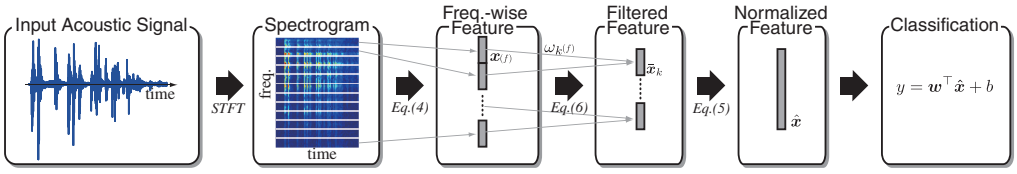
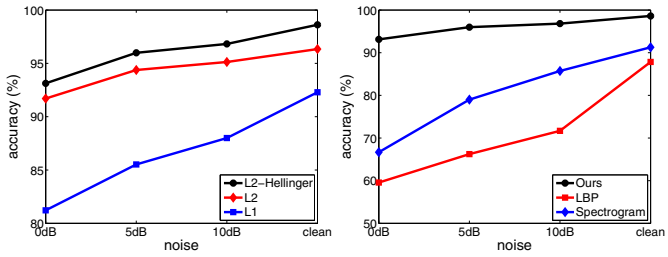


Fig. 4. Overview of the sound classification scheme. In classification, $\boldsymbol{\omega}$ and b are learned by linear SVM.

Table 1. Performances on various local patch sizes. Horizontal and vertical axes mean the time and frequency, respectively The best performance is is 98.62% at 2×4 (time×freq.).

frequency \ time				
N/A	88.68	95.24	96.62	97.15
96.30	97.73	98.27	98.26	98.07
97.20	98.29	98.30	98.23	97.88
97.44	**98.62**	98.22	98.00	97.31
97.94	98.47	98.22	97.20	96.42



(a) normalization methods (b) comparison to other methods

Fig. 5. Robustness to noise in sound signals.

Table 2. Performance results on RWC dataset.

(a) Normalization

Normalization	L_1	L_2	L_2-Hellinger
Acc. (%)	92.29	96.34	**98.62**

(b) Comparison

Method	Ours	LBP	Spectrogram	Ye et al. [3]
Acc. (%)	**98.62**	87.83	91.28	94.41

(c) Averaged computation time per sound clip which is 853 msec on an average

Procedure	Spectrogram (STFT)	Feature	Classification
time (msec)	4.11	**9.55**	0.35

4.2. Comparison to the other methods

Next, the proposed method is compared to the other methods; the standard LBP features, spectrogram features averaged over the time axis, and the method proposed by Ye et al. [3] who reported the performance on the whole set of 105 classes. The performance results are shown in Table 2b. Some other works have also reported the classification performance on the RWCP dataset, but they used the subset of the dataset; *e.g.*, the works of [5, 6, 17, 18] reported around 90% only on $10 \sim 20$ sound categories, and Dennis et al. [4] exhibited the performance of 98.1% which is close to our result though their method was evaluated only on 60 categories, half subset of ours. Therefore, we can say that the proposed method achieves the state-of-the-art performance on the whole RWCP dataset. Table 2b also shows the proposed method significantly outperforms the method of LBP, demonstrating that the local statistics incorporated into the proposed method are quite effective for classification.

4.3. Robustness to noise

Finally, we show the robustness of the methods to noises. The RWCP dataset is recorded in clean setting suppressing noise. We added the noise signals of "factory (floor1 and floor2)" in NOISEX'92 dataset to the original sound signals in RWCP dataset with various noise intensities (denoted by dB). The performance results are shown in Figure 5. As the signals are more noisy, the performances are accordingly degraded. In the proposed method, however, the performances are stably high with the low performance decay, while the LBP is significantly affected by the noise. These results show that the proposed method is robust to noise in signals while producing high performances.

The computation time is shown in Table 2c on 3.33GHz Xeon PC using MATLAB. The computation time required in the feature extraction is quite low, which contributes to speed up the whole procedure for classifying the sounds.

5. CONCLUSION

In this paper, we have proposed the novel method to extract acoustic features for classifying the environmental sounds. The proposed method characterizes a spectrogram, which forms two-dimensional data like an image, by means of LBP. We effectively incorporate the local statistics in the spectrogram, mean and standard deviation on local pixels, to improve the robustness of LBP-based features to fluctuations in pixel values. We also provided the effective normalization technique, L_2-Hellinger, for enhancing discriminative power and robustness of the proposed features. In the experiments on environmental sound classification using RWCP sound dataset, the proposed method exhibited the state-of-the-art performance compared to the other methods, demonstrating the robustness to noise and low computation time.

6. REFERENCES

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

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

[3] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Kernel discriminant analysis for environmental sound recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2013 (accepted).

[4] Jonathan Dennis, Huy Dat Tran, and Haizhou Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Transactions on Signal Processing Letters*, vol. 18, no. 2, pp. 130–133, 2011.

[5] Nobuhide Yamakawa, Tetsuro Kitahara, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Effects of modelling within- and between-frame temporal variations in power spectra on non-verbal sound recognition," in *INTERSPEECH*, 2010, pp. 2342–2345.

[6] Xuan Guo, Yoshiyuki Toyoda, Huankang Li, Jie Huang, Shuxue Ding, and Yong Liu, "Environmental sound recognition using time-frequency intersection patterns," *Applied Computational Intelligence and Soft Computing*, 2012.

[7] Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "Exploiting temporal feature integration for generalized sound recognition," *EURASIP Journal on Advances in Signal Processing*, 2009.

[8] Courtenay V. Cotton and Daniel P.W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 69–72.

[9] Matti Pietikäinen, Guoying Zhao, Abdenour Hadid, and Timo Ahonen, *Computer Vision Using Local Binary Pattern*, Springer, 2011.

[10] 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.

[11] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[13] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010, pp. 143–156.

[14] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[15] V.N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[16] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *International Conference on Language Resources and Evaluation*, 2000, pp. 965–968.

[17] Héctor Lozano, Inmaculada Hernáez, Artzai Picón, Javier Camarena, and Eva Navas, "Audio classification techniques in home environments for elderly/dependant people," in *International Conference on Computers Helping People with special needs*, 2010, pp. 320–323.

[18] Yang Zhang, Shun Nishide, Toru Takahashi, Hiroshi G. Okuno, and Tetsuya Ogata, "Cluster self-organization of known and unknown environmental sounds using recurrent neural network," in *International Conference on Artificial Neural Networks*, 2011, pp. 167–175.