Environmental Sounds Classification Based on Visual Features

Sameh Souli[1] and Zied Lachiri[1,2]

[1] Signal, Image and pattern recognition research unit
Dept. of Genie Electrique, ENIT
BP 37, 1002, Le Belvédère, Tunisia
soulisameh@yahoo.fr
[2] Dept. of Physique and Instrumentation, INSAT
BP 676, 1080, Centre Urbain, Tunisia
zied.lachiri@enit.rnu.tn

Abstract. This paper presents a method aimed at classification of the environmental sounds in the visual domain by using the scale and translation invariance. We present a new approach that extracts visual features from sound spectrograms. We suggest to apply support vector machines (SVM's) in order to address sound classification. Indeed, in the proposed method we explore sound spectrograms as texture images, and extracts the time-frequency structures by using a translation-invariant wavelet transform and a patch transform alternated with local maximum and global maximum to pursuit scale and translation invariance. We illustrate the performance of this method on an audio database, which composed of 10 sounds classes. The obtained recognition rate is of the order 91.82 % with the multiclass decomposition method: One-Against-One.

Keywords: Environmental sounds, Visual features, Translation-invariant wavelet transform, Spectrogram, SVM Multiclass.

1 Introduction

The environmental sound classification has for purpose the identification of some everyday life sound classes. It is about an elementary task participant in the conception of remote monitoring systems for the securing urban transport, the assistance to the old persons, etc.

For a long time, choosing suitable features for environmental sounds is a basic problem in audio signal processing. The environmental sound classification system can achieve important results for surveillance and security applications. Many previous works [9], [10] and [11] have concentrated on classification of environmental sound, which used in extraction phase an audio feature vector with a very limited components number like Line Spectral Frequencies (LSF's), spectral energy distribution, Linear-Frequencies Cepstral Coefficients (LFCCs). Many other studies [1] and [2], used a combination of audio features such as wavelet-based features, MFCCs, individual temporal and frequency features. The majority of these

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 459–466, 2011.

studies focus on the acoustic features derived from linear models of sound production. Indeed, this work presents a classification of the environmental sounds in the visual domain by processing the time-frequency representation sounds as texture images. In the time-frequency plan the descriptors extraction method is based on using the wavelet technique followed by a local maximum application of the obtained wavelet coefficients. A patch transform is then applied to group together the similar time-frequency geometries, followed by a research for a global maximum to select a representative time-frequency structure. The classification phase is realized by using SVM's with the multiclass One-Against-One and One-Against-Rest methods.

This paper is organized in four parts. Section 2 presents the advantage of using sound environmental spectrogram, describes the visual feature extraction method and depicts the classification algorithm. Classification results are given in Section 3. Finally conclusions are presented in Section 4.

2 Description of the Classification System

In this paper, first we apply a time-frequency transformation on the signal to obtain the spectrogram. Then, we pass into the phase of characteristics extraction from the resulting spectrogram. This extraction uses the scale and translation invariance [3]. Finally, we adopt the SVM's for the classification phase.

2.1 Visual Features Extraction

A spectrogram is an energy representation of signal, obtained by Short-time Fourier transform, it displays several distinctive characteristics [12]. Therefore, a spectrogram is compact and the most efficient representation to observe the complete spectrum of environmental sounds and to express sound by combining the merit of time and frequency domains [13]. Furthermore, we can easily identify the spectrograms of environmental sounds by their contrast, since they are considered as different textures [14]. These observations show that the spectrograms contain characteristics that can be used to differentiate between different classes of environmental sounds [15].

After the signal spectrogram calculation [14], we extracted visual features based on translation-invariant wavelet transform, followed by a particular patch transform and a global selection operation [16].

In this paper, the algorithm is based on the following steps:

Step 1 : Translation-invariant wavelet transform. Let $S[x, y]$ be a spectrogram of the size $N_1 \times N_2$ .We used the translation-invariant wavelet transform. The resulting wavelet coefficients will be defined by:

$$Wf(u, v, j, k) = \sum_{x=1}^{N_1} \sum_{y=1}^{N_2} S[x, y] \frac{1}{2^j} \psi^k \left(\frac{x-u, y-v}{2^j} \right). \quad (1)$$

Where $k = 1,2,3$ is the orientation (horizontal, vertical, diagonal), $\psi^k(x, y)$ is the wavelet function. Indeed, to build a translation- invariant wavelet representation, the scale is made discrete but not the translation parameter. The scale is sampled on a

dyadic analysis $\{2^j\}_{j\in Z}$. The use of the translation-invariant wavelet transform creates a redundancy of information that allows keeping the translation-invariance at all levels of factorization [7]. The scale invariance is carried out by normalization, using the following formula:

$$S_1(u,v,j,k) = \frac{|Wf(u,v,j,k)|}{\|S\|^2_{\sup p(\psi_j^k)}} . \quad (2)$$

Where $\|S\|^2_{\sup p(\psi_j^k)}$ is the energy of detail wavelet coefficients of a spectrogram.

Step 2 : Local Maximum. The continuation of translation invariance [3] and [16], is done by calculating the local maximum of S_1:

$$C_1(u,v,,j,k) = \max_{u'\in[2^j(u-1)+1,2^j u),v'\in[2^j(v-1)+1,2^j v)} S_1(u',v',j,k) . \quad (3)$$

The C_1 section is obtained by a subsampling of S_1 using a cell grid of the $2^j \times 2^j$ size that is then followed by the local maximum. Generally, the maximum being taken at each j scale and k direction of a spatial neighborhood of a size that is proportional to $2^j \times 2^j$. The resulting C_1 at the j scale and the k direction is therefore of the $N_1 / 2^j \times N_2 / 2^j$ size, where $j = 1,2,3$.

Step 3 : Patch Transform. The idea consists of selecting N prototypes P_i of C_1, then the scalar product is calculated between the prototypes P_i and the C_1 coefficients, then followed by a sum [11] . For every patch, we get only one scalar at the end.

$$S_2(u,v,j,i) = \sum_{u'=1}^{N_1/2^j} \sum_{v'=1}^{N_2/2^j} \sum_{k=1}^{3} C_1(u',v',j,k) P_i(u'-u,v'-v,k) . \quad (4)$$

Where P_i of size $M_i \times M_i \times 3$ are the patch functions that group 3 wavelet orientations. The patch functions are extracted by a simple sampling at a random scale and a random position of the C_1 coefficients of a spectrogram [3], for instance a P_0 patch of the $M_0 \times M_0$ size contains $M_0 \times M_0 \times 3$ elements, M_0 may take the following values $(M_0 = 4,8,12)$.

Step 4: Global Maximum. The C_2 coefficients are obtained by the application of the max function on S_2:

$$C_2(i) = \max_{u,v,j} S_2(u,v,j,i) \ . \tag{5}$$

In this work, the obtained result is a vector of NC_2 values, where N corresponds to the number of extracted patches. In this way, the C_2 obtained coefficients constitute the parameter vector for the classification.

2.2 SVM Classification

The classification is performed using a new technique of statistic learning: Support Vector Machines. The SVM's is a tool for creating practical algorithms for estimating multidimensional functions [4].

Let a set of data $(x_1, y_1),\ldots,(x_m, y_m) \in \Re^d \times \{\pm 1\}$ where $X = \{x_1,\ldots, x_m\}$ a dataset in $\Re^d$ where each x_i is the feature vector of a signal. In the nonlinear case, the idea is to use a kernel function $k(x_i, x_j)$, where $k(x_i, x_j)$ satisfies the Mercer conditions [5]. Here, we used a Gaussian RBF kernel whose formula is:

$$k(x, x') = \exp\left[-\left\| x - x' \right\|^2 / 2\gamma^2 \right] . \tag{6}$$

Where $\|.\|$ indicates the Euclidean norm in $\Re^d$.

Let Ω be a nonlinear function which transforms the space of entry $\Re^d$ to an intern space H called a feature space. Ω allows to perform a mapping to a large space in which the linear separation of data is possible [8].

$$\begin{aligned} &\Omega : \Re^d \rightarrow H \\ &(x_i, x_j) \mapsto \Omega(x_i)\Omega(x_j) = k(x_i, x_j) \end{aligned} \ . \tag{7}$$

The H space is a reproducing kernel Hilbert space (RKHS).

Thus, the dual problem is presented by a Lagrangian formulation as follows:

$$\max W(\alpha) = \sum_{i=0}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j k(x_i, x_j), i = 1,\ldots,m \ . \tag{8}$$

Under the following constraints:

$$\sum_{i=1}^{m} \alpha_i y_i = 0, \ 0 \le \alpha_i \le C \ . \tag{9}$$

They α_i are called Lagrange multipliers and C is a regularization parameter which is used to allow classification errors. The decision function will be formulated as follows:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b). \quad (10)$$

We hence adopted two approaches of multiclass classification: One-against-the-Rest and One-against-One. The first method one-against-the-rest builds K models of binary SVM. The SVM model assigns the label '1' for the class C_k and the supplementary label '-1' to all the remaining classes. The second method one-against-one, consists of creating a binary classification of each possible combination of classes, the result for K classes $K(K-1)/2$. The classification is then carried out in accordance with the majority voting scheme [6].

3 Experimental Results

Our corpus of sounds comes from commercial CDs [18]. Among the sounds of the corpus we find: explosions, broken glass, door slamming, gunshot, etc. We used 10 classes of environmental sounds as shown in Table 1.

All signals have a resolution of 16 bits and a sampling frequency of 44100 Hz that is characterized by a good temporal resolution and a wide frequency band. Most of the signals are impulsive, we took 2/3 for the training and 1/3 for the test. Each spectrogram is segmented into 8 non-overlapping segments. Each segment is composed of 64 samples.

For each signal, firstly we apply a time-frequency transformation, then the resultant spectrogram passes by the various stages of the proposed-visual characteristic extraction method. Finally, the obtained feature vector passed for the classification phase by using SVM's. Among the big problems met during the classification by the SVM's is the choice of the values of the kernel parameter γ and the constant of regularization C. To resolve this problem we used the cross-validation method. Indeed, according to [17], this method consists in setting up a grid-search for γ and C. For the implementation of this grid, it is necessary to proceed

Table 1. Classes of sounds and number of samples in the database used for performance evaluation

Classes	Train	Test	Total number
Door slams	208	104	312
Explosions	38	18	56
Class breaking	38	18	56
Dog barks	32	16	48
Phone rings	32	16	48
Children voices	54	26	80
Gunshots	150	74	224
Human screams	48	24	72
Machines	38	18	56
Cymbals	32	16	48
Total	670	330	1000

iteratively, by creating a couple of values γ and C. In this work, we use the following couples C, γ : C=[$2^{(0)}$, $2^{(1)}$, …, $2^{(16)}$] et γ=[$2^{(15)}$, $2^{(-14)}$, …, $2^{(2)}$].

In Table 2, we present the results obtained with various classes of sound and different C, γ settings of Gaussian RBF kernel. After learning phase, we test firstly the train data then the test data. We remark that the classification rate is different from one class to another. We were able to achieve an averaged accuracy rate of the order 91.82% in ten classes with one-against-one approach. There are seven classes that have a classification rate higher than 90%. But with one-against-all approach we obtained an averaged accuracy rate of the order 87.90%.

The obtained results by our classification system in nine classes of environmental sounds, with one-against-one approach, is satisfactory. Indeed, we reached a recognition

Table 2. Recognition rates for visual feature applied to one-vs-all and one-vs-one SVMs based classifiers

Classes	Kernel	Parameters (C, γ)	Multiclass Approach	Classification rate(%)		Execution Time(s)
				Train	Test	
Door slams	Gaussien	$(2^{(1)}, 2^{(-7)})$	One-vs-all	91.42	85.71	147.45
	RBF	$(2^{(1)}, 2^{(-7)})$	One-vs-One	94.28	90.47	7.91
Explosions	Gaussien	$(2^{(-5)}, 2^{(-15)})$	One-vs-all	91.28	90.47	113.19
	RBF	$(2^{(3)}, 2^{(-5)})$	One-vs-One	94.28	95.23	7.96
Class breaking	Gaussien	$(2^{(1)}, 2^{(-9)})$	One-vs-all	98,46	97,43	147.94
	RBF	$(2^{(1)}, 2^{(-7)})$	One-vs-One	97.94	97.43	7.96
Dog barks	Gaussien	$(2^{(0)}, 2^{(-8)})$	One-vs-all	90.00	83.33	113.02
	RBF	$(2^{(1)}, 2^{(-7)})$	One-vs-One	93.33	88.88	7.91
Phone rings	Gaussien	$(2^{(0)}, 2^{(-6)})$	One-vs-all	90.00	77.77	111.01
	RBF	$(2^{(2)}, 2^{(-6)})$	One-vs-One	93.33	83.33	8.01
Children voices	Gaussien	$(2^{(3)}, 2^{(-7)})$	One-vs-all	94.00	90.00	112.57
	RBF	$(2^{(1)}, 2^{(-7)})$	One-vs-One	96.00	93.33	7.88
Gunshots	Gaussien	$(2^{(5)}, 2^{(-5)})$	One-vs-all	97,85	96,42	112.77
	RBF	$(2^{(0)}, 2^{(-15)})$	One-vs-One	98.57	97.61	7.97
Human screams	Gaussien	$(2^{(0)}, 2^{(-4)})$	One-vs-all	93.33	88.88	141.32
	RBF	$(2^{(1)}, 2^{(-7)})$	One-vs-One	95.55	92.59	7.69
Machines	Gaussien	$(2^{(12)}, 2^{(2)})$	One-vs-all	91.42	85.71	112.35
	RBF	$(2^{(5)}, 2^{(-3)})$	One-vs-One	94.28	90.47	7.80
Cymbals	Gaussien	$(2^{(8)}, 2^{(-2)})$	One-vs-all	90.00	83.33	122.20
	RBF	$(2^{(2)}, 2^{(-6)})$	One-vs-One	93.33	88.88	8.56

rate of the order 92.14% with visual descriptors, but in [2] the obtained classification rate is of 90.23%, whose method is to extract from the signal the following descriptors: MFCCs, Energy and Log energy. Furthermore, by combining wavelet-based features, MFCCs, individual temporal and frequency features, "Rabaoui, et al " [2] have attained a recognition rate of the order 93.22% with the same classes and the same classification approach that we have adopted. It proves that our result is efficient compared to the number of the used characteristics parameters.

The adjustment of the extraction method of visual features, used in image processing, to the special characteristics of the environmental sounds has given satisfactory and improved classification results. Furthermore, the used feature vector represent all relevant time-frequency information in the signals to recognize.

4 Conclusion

This paper presents a new approach for environmental sound classification in the visual domain by processing spectrogram as texture images. Indeed this approach is based on the use of wavelet technique followed by a local maximum then a patch transform, and finally by a global maximum. The obtained results are very satisfactory (91.82 % with the method one-against-one and 87.90 % with the method one-against-all). The proposed approach can be improved while digging deeply into the visual domain.

Acknowledgments. We are grateful to G. Yu for many discussions by mail.

References

1. Chu, S., Narayanan, S., Kuo, C.C.J.: Environmental Sound Recognition with Time-Frequency Audio Features. IEEE Trans. on Speech, Audio, and Language Processing 17, 1142–1158 (2009)
2. Rabaoui, A., Davy, M., Rossignol, S., Ellouze, N.: Using One-Class SVMs and Wavelets for Audio Surveillance. IEEE Transactions on Information Forensics and Security 3, 763–775 (2008)
3. Schulz-Mir, H., Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust Object Recognition with Cortex-Like Mechanisms. IEEE Transactions Pattern Analysis and Machine Intelligence 29, 411–426 (2007)
4. Vladimir, V., Vapnik, N.: An Overview of Statistical Learning Theory. IEEE Transactions on Neural Networks 10, 988–999 (1999)
5. Vapnik, V., Chapelle, O.: Bounds on error expectation for support vector machines. Neural Computation 12 (2000)
6. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 13, 415–425 (2002)
7. Mallat, S.: A Wavelet Tour of Signal Processing, 2nd edn. Academic Press (1999)
8. Scholkopf, B., Smola, A.: Learning with Kernels. MIT Press (2001)
9. El-Maleh, K., Samouelian, A., Kabal, P.: Frame-Level Noise Classification in Mobile Environments. In: Proc. ICASSP, Phoenix, AZ, pp. 237–240 (1999)

10. Dufaux, A., Besacier, L., Ansorge, M., Pellandini, F.: Automatic Sound Detection and Recognition For Noisy Environment. In: Proceedings of European Signal Processing Conference (EUSIPCO), Tampere, FI, pp. 1033–1036 (2000)
11. Fleury, A., Noury, N., Vacher, M., Glasson, H., Serigna, J.-F.: Sound and Speech Detection and classification in a Health Smart Home. In: 30th Annual Int. Conf. IEEE, Engineering in Medicine and Biology Society (EMBS), Canada, pp. 4644–4647 (2008)
12. He, L., Lech, M., Maddage, N.: Stress and Emotion Recognition Using Log-Gabor Filter Analysis of Speech Spectrograms. In: 3rd Int. Conf. Affective Computing and Intelligent Interaction and Workshops, ACII, Amsterdam, pp. 1–6 (2009)
13. Xinyi, Z., Jianxiao, Y., Qiang, H.: Research of STRAIGHT Spectrogram and Difference Subspace Algorithm for Speech Recognition. In: Int. Congress on Image and Signal Processing (CISP 2009), IEEE DOI Link 0910, pp. 1–4 (2009)
14. Yu, G., Slotine, J.J.: Audio Classification from Time-Frequency Texture. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Taipei, pp. 1677–1680 (2009)
15. He, L., Lech, M., Maddage, N.C., Allen, N.: Stress Detection Using Speech Spectrograms and Sigma-pi Neuron Units. In: Fifth Int. Conf. on Natural Computation, pp. 260–264 (2009)
16. Yu, G., Sloine, J.J.: Fast Wavelet-based Visual Classification. In: Proc. IEEE ICPR, Tampa (2008)
17. Hsu, C.-W., Chang, C-C., Lin, C-J.: A practical Guide to Support Vector Classification. Department of Computer Science and Information Engineering National, Taipei, Taiwan (2009)
18. Leonardo Software, Santa Monica, CA 90401, `http://www.leonardosoft.com`