# ROBUST AUDIO SURVEILLANCE USING SPECTROGRAM IMAGE TEXTURE FEATURE

*Roneel V Sharan and Tom J Moir*

School of Engineering, Auckland University of Technology,
Private Bag 92006, Auckland 1142, New Zealand
roneel.sharan@aut.ac.nz, tom.moir@aut.ac.nz

## ABSTRACT

A sound signal produces a unique texture which can be visualized using a spectrogram image and analyzed for automatic sound recognition. In this paper, we explore the use of a well-known image texture analysis technique called the gray-level co-occurrence matrix (GLCM) for sound recognition in an audio surveillance application. The GLCM captures the distribution of co-occurring values at a given offset. Unlike most other similar research which derive features from the GLCM, we use the matrix values itself to form the feature vector with analysis carried out in sub-bands. When compared to a baseline feature from related work, the proposed spectrogram image texture feature (SITF) gives marginally lower results under clean and high signal-to-noise ratio (SNR) conditions but significantly better results are achieved at low SNR, where the baseline feature was seen to be less effective.

***Index Terms***—Audio surveillance, gray-level co-occurrence matrix, spectrogram image texture feature, sound recognition, support vector machine

## 1. INTRODUCTION

In recent years, there has been a growing interest in using features derived from the spectrogram image of a sound signal in sound recognition applications. In a sound event recognition application in [1], central moments are extracted as features from the short-time Fourier transform (STFT) spectrogram image of sound signals which produced significantly better results under noisy conditions when compared to mel-frequency cepstral coefficients (MFCCs), a common feature in speech and sound recognition applications. In an audio surveillance application in [2], we took a similar approach to this spectrogram image feature (SIF) method but with reduced feature dimensions, referred as reduced SIF (RSIF). Time-frequency images of a sound signal were also used for feature extraction in a hearing aid application [3]. While more than thirty features were extracted, eleven features were chosen through correlation analysis for classifying four classes.

In music genre recognition in [4], a slightly different approach is taken where the spectrogram representation is viewed as a texture image. Texture analysis is carried out using gray-level co-occurrence matrix (GLCM), also known as gray-tone spatial dependence matrix [5], which gives the spatial relationship of pixels in an image. From the fourteen textural descriptors proposed in [5], the following seven are extracted as features from the GLCM: entropy, correlation, homogeneity, third order momentum, maximum likelihood, contrast, and energy. The GLCM texture analysis technique using the fourteen textural descriptors of [5], a subset of these features, or with the addition of other textural descriptors has been successfully applied in other areas of research involving image classification such as insect recognition [6], evaluating fabric surface roughness [7], diagnosis of abdominal tumors using ultrasound images [8], and urban and agricultural land classification [9].

However, in a face recognition problem in [10], instead of extracting textural features from the GLCM, the matrix values itself form the feature vector. When tested with four different databases, under most experimental conditions, this approach was shown to give significantly better results than using the combined fourteen textural descriptors as features.

In this work, we apply the GLCM technique of texture analysis to sound signal spectrogram images for classification of sounds in an audio surveillance application. However, instead of extracting textural descriptors from the GLCM, we propose to concatenate the columns of the matrix to form the feature vector for a sound signal. We refer this as the spectrogram image texture feature (SITF). Unlike in [4, 10], we also evaluate the performance of the SITF under noisy conditions. In addition, we perform texture analysis in sub-bands. This essentially divides the spectrogram image into horizontal sections of different frequency bands. Analysis is performed independently in each frequency band and the final feature vector is a concatenation of the feature vectors from each sub-band. Due to the non-uniform nature of the sound signal spectrograms, this local feature extraction technique was shown to give higher results than global features in [4], which they referred as zoning.

The rest of this paper is organized as follows. Section 2 gives an overview of feature vector formation using GLCM method of spectrogram image texture analysis. Section 3 is on experiments, results, and discussions while conclusion and future recommendations are given in Section 4.

## 2. FEATURE EXTRACTION

The procedure for generating the time-frequency image and performing image texture analysis using GLCM are described below.

### 2.1. Grayscale Spectrogram

In this work, we consider the linear grayscale spectrogram only since it was found to be more noise robust than the log grayscale spectrogram in [2] and gave the best overall classification accuracy. Firstly, the discrete Fourier transform (DFT) is applied to the windowed signal as

$$X(k,t)=\sum_{n=0}^{N-1} x(n)\,w(n)\,e^{\frac{-2\pi ikn}{N}}, \qquad k=0,\ldots,N-1 \tag{1}$$

where $N$ is the window length, $x(n)$ is the time-domain signal, $X(k,t)$ is the $k^{th}$ harmonic corresponding to the frequency $f(k)=kF_s/N$ for the $t^{th}$ frame, $F_s$ is the sampling frequency, and $w(n)$ is the window function.

The linear values are then obtained as

$$S(k,t)=\left|X(k,t)\right|. \tag{2}$$

These values are then normalized in the range [0,1] which gives the grayscale image intensity values. The normalization is given as

$$I(k,t)=\frac{S(k,t)-\min(S)}{\max(S)-\min(S)}. \tag{3}$$

Illustration of spectrogram images, mapped to the HSV color space for better visualization, under clean conditions and with the addition of noise at 0dB signal-to-noise ratio (SNR) can be found in Fig. 1.

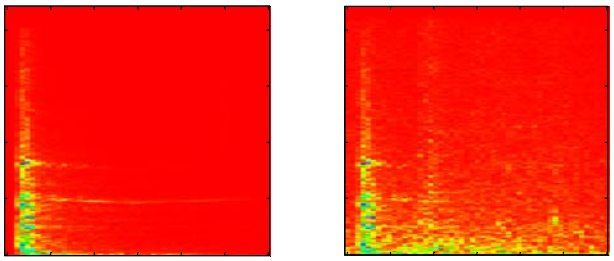
### 2.2. Feature Vector Formation Using GLCM

GLCM is a matrix of frequencies where each element $(i,j)$ is the number of times intensity value $j$ is located at a certain distance and angle, given by the displacement vector $[d_k\ d_t]$, where $d_k$ is the offset in the $y$ direction and $d_t$ is the offset in the $x$ direction, from intensity value $i$ in an $N_t \times N_k$ image $I$. Mathematically, this can be given as

$$P(i,j)=\sum_{k=1}^{N_k}\sum_{t=1}^{N_t}\begin{cases}1, & \text{if } I(k,t)=i \ \&\ I(k+d_k,t+d_t)=j\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

where the size of the output matrix is $N_g \times N_g$, $N_g$ is the number of quantized gray levels. The typical angles for computing the GLCM are $0°, 45°, 90°,$ and $135°$ corresponding to the displacement vector $[0\ d]$, $[-d\ d]$, $[-d\ 0]$, and $[-d\ -d]$, respectively.

The feature vector is then formed by concatenating the GLCM values into a column vector.



(a) (b)

**Fig. 1.** Spectrogram images for a sound signal from *construction* sound class. (a) Linear spectrogram image under clean conditions and (b) linear spectrogram image at 0dB SNR with factory noise.

## 3. EXPERIMENTAL EVALUATION

A description of the sound database used in this work is given first followed by an overview of the noise conditions and the experimental setup. We then present results using two baseline features: MFCCs and RSIF. Next, results using the proposed SITF are presented.

### 3.1. Sound Database

The sound database has a total of 1143 files belonging to 10 classes: *alarms*, *children voices*, *construction*, *dog barking*, *footsteps*, *glass breaking*, *gunshots*, *horn*, *machines*, and *phone rings*. The sound files are largely obtained from the Real World Computing Partnership (RWCP) Sound Scene database in Real Acoustic Environment [11] and the BBC Sound Effects library [12]. All signals in the database have 16-bit resolution and a sampling frequency of 44100 Hz. More details about the sound database and its comparison with that used in other similar work can be found in [2].

### 3.2. Noise Conditions

The performance of the different features are investigated under three different noise environments taken from the NOISEX-92 database [13]: *speech babble*, *factory floor 1*, and *destroyer control room*. The signals are resampled at 44100 Hz and the overall performance is measured in clean conditions and at 20dB, 10dB, 5dB, and 0dB SNR.

### 3.3. Experimental Setup

For all experiments, signal processing is carried out using a Hamming window of 512 points (11.61 ms) with 50% overlap. Support vector machine (SVM) is used for classification where the classification accuracy is given in percentage as *number of correctly classified test samples* divided by the *total number of test samples*. Being a binary classifier, we use the one-against-all (OAA) method [14, 15] for multiclass classification where the classifier that has the highest output function assigns the class. In a similar work [2], the OAA method was shown to give the best overall performance when compared to one-against-one (OAO) [16], decision directed acyclic graph (DDAG) [17], and adaptive directed acyclic graph (ADAG) [18] multiclass classification methods and against the K-nearest neighbor (KNN) classifier.

**Table I.** Classification accuracy using baseline features

| Feature | Clean | 20dB | 10dB | 5dB | 0dB | Average |
|---|---|---|---|---|---|---|
| Log-MFCC | 98.43 | 92.83 | 73.14 | 57.57 | 43.31 | 73.05 |
| Linear-MFCC | 99.21 | 93.53 | 86.09 | 70.87 | 47.16 | 79.37 |
| Linear-RSIF | 92.13 | 92.04 | 89.33 | 78.57 | 53.37 | 81.08 |

All results are reported using a nonlinear SVM with a Gaussian radial basis function kernel as it was found to give the best results. The classifier parameters, refer to [2], were tuned using cross validation where, instead of maximizing the classification accuracy under each noise condition, samples from all noise conditions were used at once to get the best overall classification accuracy. For all experimentations, the classifier is trained with two-third of the clean samples with the remaining one-third data used for testing under clean and noisy conditions.

### 3.4. Results and Discussions

*3.4.1. Baseline Features*

The first baseline method uses MFCCs as features. The feature vector for each frame is 39-dimensional: 13 MFCCs using a 20-filterbank system, plus deltas and accelerations. The overall size of the feature vector for a signal is $39 \times N_t$, where $N_t$ is the number of frames in the signal, which is different in each case depending on the length of the signal. After data normalization, the final feature vector is represented by concatenating the mean and standard deviation for each dimension. This results in a 78-dimensional final feature vector.

The second baseline method uses the SIF where the spectrogram image is divided into $9 \times 9$ blocks and second and third central moments are computed in each block. In [1], the features from each block were concatenated to form the final feature vector, resulting in a 162-dimensional feature vector. However, in [2], feature dimension reduction by concatenating the mean and standard deviation of the central moment values along the row and column of the blocks achieved comparable classification accuracy. This technique resulted in a 72-dimensional feature vector. Therefore, only results using the RSIF are presented.

The classification accuracy values using the baseline features are given in Table I. MFCCs give the highest classification accuracy under clean conditions and at 20dB SNR but the RSIF gives superior performance at 10dB, 5dB, and 0dB SNR and gives the best overall classification accuracy. A detailed analysis of the results can be found in [2].

*3.4.2. SITF*

For obtaining the SITF, we first apply the GLCM technique of texture analysis to the spectrogram images. The feature vector is then formed by concatenating the columns of the matrix. In preliminary experiments, we compared the classification accuracy with increasing values of $N_g$. The average classification accuracy decreased as $N_g$ increased, therefore, for all the experiments that follow, we use $N_g = 2$, which gave the highest average classification accuracy.

**Table II.** Classification accuracy using SITF – individual and combined feature vectors ($N_g = 2$ and $d = 1$)

| Angle | Clean | 20dB | 10dB | 5dB | 0dB | Average |
|---|---|---|---|---|---|---|
| 0° | 84.78 | 84.60 | 77.69 | 68.24 | 49.34 | 72.93 |
| 45° | 82.15 | 81.98 | 80.23 | 75.33 | 56.61 | 75.26 |
| 90° | 76.12 | 76.12 | 74.45 | 70.34 | 50.39 | 69.48 |
| 135° | 81.36 | 81.28 | 78.57 | 72.00 | 54.42 | 73.53 |
| All Angles | 86.09 | 85.74 | 81.45 | 74.89 | 55.03 | 76.64 |

We perform two sets of experiments using the SITF:

- compare the classification accuracy using feature vectors formed from application of GLCM analysis at angles of $0°, 45°, 90°,$ and $135°$ and then with combined feature vector,
- compare the classification accuracy with increasing number of frequency bands.

The results for the first set of experiments are given in Table II with $N_g = 2$ and $d = 1$. Comparing the average classification accuracy, for feature vectors using individual angles, the best average classification accuracy is achieved using an angle of $45°$ while the combined feature vector gives marginally better classification accuracy. In this experiment, the spectrogram image is not divided into sub-bands before feature extraction. Therefore, the feature vector dimension when analyzing at individual angles is $N_g^2 = 4$ and $4N_g^2 = 16$ when the feature vector from the four angles are combined. As such, while the feature vector dimension has quadrupled when combined, there isn't a considerable increase in the classification accuracy in comparison to the best performing individual feature vector. Also, while the best results at 0dB SNR using SITF are slightly higher, the overall performance using Linear–MFCCs and RSIF, as presented in Table I, are still better.

In the next set of experiments, we first look at the effect of performing GLCM analysis with increasing number of frequency bands on the classification accuracy. The spectrogram image is now divided into blocks of horizontal sections with equal number of frequency bins in each sub-band. The GLCM is computed in each sub-band which are then concatenated into one matrix. This matrix is then concatenated into a column vector which forms the final feature vector. The number of pixels in the spectrogram image along the vertical, or frequency axis, is $N/2 = 256$, therefore, various number of frequency bands, $N_b$, from $1$ to $256$ can be experimented with. We experimented with $N_b = 1, 2, 4, 8, 16, 32, 64,$ and $128$ at a time. The results presented in Table III use feature vector combined from all four angles.

While there isn't a significant change in the classification accuracy with increasing values of $N_b$ at lower values of $N_b$, there is notable increase in the classification

**Table III.** Classification accuracy using SITF (combined feature vector) – effect of increasing value of $N_b$ ($N_g = 2$ and $d = 1$)

| $N_b$ | Clean | 20dB | 10dB | 5dB | 0dB | Average |
|---|---|---|---|---|---|---|
| 1 | 86.09 | 85.74 | 81.45 | 74.89 | 55.03 | 76.64 |
| 2 | 85.04 | 84.51 | 81.98 | 75.42 | 57.83 | 76.96 |
| 4 | 85.04 | 84.34 | 82.59 | 77.17 | 59.58 | 77.74 |
| 8 | 83.99 | 83.99 | 82.33 | 77.34 | 59.49 | 77.43 |
| 16 | 86.09 | 86.09 | 84.16 | 82.15 | 62.38 | 80.17 |
| 32 | 87.93 | 87.66 | 86.79 | 86.00 | 69.03 | 83.48 |
| **64** | **90.29** | **89.68** | **89.59** | **87.75** | **73.40** | **86.14** |
| 128 | 88.71 | 88.63 | 88.10 | 85.83 | 72.27 | 84.71 |

accuracy from $N_b = 16$ onwards with the most improved results at 5dB and 0dB SNR. The highest classification accuracy under all noise conditions is at $N_b = 64$.

The GLCM method of texture analysis determines frequency of repeating patterns or intensity value combinations in the spectrogram image. The intensity value of pixels in a spectrogram image are dependent on the amplitude of the signal at any given time and frequency. The linear grayscale spectrogram only shows the dominant power frequencies and the more diffuse noise has only isolated effects on the time-frequency image, as shown in Fig. 1. As such, the inherent patterns largely remain intact with the addition of noise. In addition, using $N_g = 2$ essentially means that the grayscale image is transformed to a binary image for GLCM analysis. Therefore, small linear transformations caused by the noise may not change a pixel value as long as it does not cross the threshold for gray-level conversion, thereby maintaining the pattern. However, extracting global features from the spectrogram image means noise manipulated data is also captured in the only set of feature data. Extracting features from different frequency bands ensures feature data in sub-bands not affected by noise are mostly unchanged leading to a more robust performance.

When compared to results using the baseline features, the proposed SITF is not able to match the classification accuracy of MFCCs under clean conditions and at 20dB SNR but gives significantly better results at 10dB, 5dB, and 0dB SNR. When compared with the results using the RSIF, the SITF gives slightly lower results under clean conditions and at 20dB SNR, comparable at 10dB SNR, but, once again, significantly better classification accuracy is achieved at 5dB and 0dB SNR, increasing from 78.57% to 87.75% and 53.37% to 73.40%, an increase of 9.18% and 20.03%, respectively. Therefore, the key advantage of the SITF over the baseline features is its greater robustness at 5dB and 0dB SNR, or low SNR in general. To ensure that this improvement wasn't simply because of the different method of spectrogram image division before feature extraction, we applied the frequency sub-band analysis method to the SIF but there wasn't any significant change in the results.

However, the disadvantage of the proposed method is its high computational cost. The SITF dimension using sub-band analysis and with combined feature vector from all four angles can be given as $N_b \times 4N_g^2$. With $N_b = 64$, where the highest classification accuracy is achieved, the feature vector dimension is $1024$, which is about $6.32$ times more than the SIF [1], $13.13$ times more than MFCCs, and $14.22$ times more than the RSIF [2]. We also experimented with the sub-band analysis technique using each of the four angles. In general, it was observed that as $N_b$ increased in value, the difference in the classification accuracy with individual feature vectors and the combined feature vector got minimal. Table IV gives the classification accuracy using feature vectors from each of the four angles considered with $N_b = 64$. Best results were once again achieved with features extracted from angle of 45°. While the individual feature vectors give slightly lower classification accuracy than the combined features, these can be considered more effective since the feature vector dimension is much lower, reduced by 4 to 256. This could be an alternative if lower computational time is a priority.

**Table IV.** Classification accuracy using SITF with individual feature vectors ($N_b = 64$)

| Angle | Clean | 20dB | 10dB | 5dB | 0dB | Average |
|---|---|---|---|---|---|---|
| 0° | 88.45 | 88.45 | 87.84 | 84.95 | 69.29 | 83.80 |
| **45°** | **89.76** | **89.41** | **89.33** | **87.66** | **71.92** | **85.62** |
| 90° | 88.45 | 88.01 | 87.93 | 86.44 | 70.78 | 84.32 |
| 135° | 88.71 | 88.71 | 88.36 | 86.44 | 71.13 | 84.67 |

We also experimented with increasing values of $d$ and it was observed that while increasing $d$ from 1 to 2 increases the average classification accuracy for lower values of $N_b$, the difference between the two set of results got smaller as the value of $N_b$ increased. Eventually, the average classification accuracy with $d = 1$ surpassed those at $d = 2$, and, at $N_b = 64$, the highest classification accuracy was still achieved using $d = 1$.

## 4. CONCLUSION

The proposed SITF was found to be more noise robust than two baseline features: MFCCs and the RSIF. When compared to the RSIF, the best performing baseline feature, the proposed feature shows greater robustness at low SNR. Best results are achieved using sub-band analysis where the spectrogram image is divided into equal sized frequency bands and the feature vector from each frequency band is concatenated to form the final feature vector. The highest classification accuracy was achieved when the feature vector from GLCM analysis at angles of $0°, 45°, 90°,$ and $135°$ was combined with $N_g = 2, N_b = 64,$ and $d = 1$. However, the average classification accuracy with individual feature vectors was only slightly lower, particularly at 45°, and had the added advantage of a feature vector which was four times smaller in dimension. We chose increasing intervals in determining the optimal value for $N_b$ but smaller intervals could be experimented with for potentially better results. Feature vector reduction methods can also be considered for faster computational time.

## 5. REFERENCES


[1] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters,* vol. 18, no. 2, pp. 130-133, 2011.

[2] R. V. Sharan and T. J. Moir, "Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM," *Neurocomputing,* (In Press).

[3] K. Abe, H. Sakaue, T. Okuno, and K. Terada, "Sound classification for hearing aids using time-frequency images," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim)*, 2011, pp. 719-724.

[4] Y. M. G. Costa, L. S. Oliveira, A. L. Koericb, and F. Gouyon, "Music genre recognition using spectrograms," in *18th International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2011, pp. 1-4.

[5] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics,* vol. SMC-3, no. 6, pp. 610-621, 1973.

[6] L.-Q. Zhu and Z. Zhang, "Auto-classification of insect images based on color histogram and GLCM," in *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2010, pp. 2589-2593.

[7] X. Wang and N. D. Georganas, "GLCM texture based fractal method for evaluating fabric surface roughness," in *Canadian Conference on Electrical and Computer Engineering (CCECE '09)*, 2009, pp. 104-107.

[8] D. Mitrea, M. Socaciu, R. Badea, and A. Golea, "Texture based characterization and automatic diagnosis of the abdominal tumors from ultrasound images using third order GLCM features," in *4th International Congress on Image and Signal Processing (CISP)*, Shanghai, 2011, pp. 1558-1562.

[9] M. Umaselvi, S. S. Kumar, and M. Athithya, "Color based urban and agricultural land classification by GLCM texture features," in *IET Chennai 3rd International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2012)*, 2012, pp. 1-4.

[10] A. Eleyan and H. Demirel, "Co-occurrence matrix and its statistical features as a new approach for face recognition," *Turkish Journal of Electrical Engineering & Computer Sciences,* vol. 19, no. 1, pp. 97-107, 2011.

[11] 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 *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000, pp. 965–968.

[12] *BBC Sound Effects Library*. Available: http://www.leonardosoft.com

[13] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication,* vol. 12, no. 3, pp. 247-251, Jul. 1993.

[14] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[15] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, *et al.*, "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image Processing*, 1994, pp. 77-82.

[16] U. H. G. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 255-268.

[17] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems 12 (NIPS-99)*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge MA: MIT Press, 2000, pp. 547-553.

[18] B. Kijsirikul, N. Ussivakul, and S. Meknavin, "Adaptive directed acyclic graphs for multiclass classification," in *PRICAI 2002: Trends in Artificial Intelligence*. vol. 2417, M. Ishizuka and A. Sattar, Eds. Berlin Heidelberg: Springer, 2002, pp. 158-168.