ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse40832019.pdf
https://doi.org/10.30534/ijatcse/2019/40832019

Spectrograms and Scalograms with Correlation Filters for Anuran Vocalization Classification

Salina Abdul Samad, Aqilah Baseri Huddin
Centre for Integrated Systems and Advanced Technologies (INTEGRA)
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia
Email: salinasamad@ukm.edu.my, aqilah@ukm.edu.my

ABSTRACT

Correlation filters have been used with spectrograms to classify animals based on their vocalizations. The training images are a series of the animal vocalizations converted into spectrograms with each class having its own corresponding template. Cross-correlation is then performed with the test spectrogram images in order to classify them to the corresponding classes based on the correlation output parameters. In this paper, a similar approach is used using not only spectrograms, but also scalograms for comparisons in classifying anuran vocalization. The results show that although increasing the number of vocalization when constructing the templates increases the accuracy rate for both spectrograms and scalograms, spectrograms are much more suitable to be used for anuran vocalization classification due a better overall performance**.**

Key words: spectrogram; scalogram, correlation filter, anuran, classification.

1. INTRODUCTION

In the field of conservation of wetlands and floodplains, anuran population is regularly used as a bio-indicator for the health of habitats [1]. Anurans such as frogs and toads can be identified by their vocalization by human experts from field recordings. Another alternative is to use signal processing and intelligent algorithms to identify and classify anurans based on their vocalization. To this end, techniques usually employed for processing speech signals have been used which include signal segmentation [2], feature extraction [3] and classification using classifiers such as Support Vector Machines [2], Neural Networks [4] and many others [5]-[6].

Instead of one dimensional audio processing, the vocalization can be represented as spectrograms and methods popular with image processing can be used to process the signals. The short-time Fourier Transform (STFT) converts the audio signals into spectrograms and techniques have been used to extract features from them such as local peaks, ridges, and their derivatives, which are used as inputs to classifiers [7]-[8].

Researchers have also experimented with spectrograms using the whole spectrogram, instead of extracting features, as inputs to image based correlation filters (CFs) [9]. CFs have been extensively used in the field of biometrics and other object detection and recognition applications. Using training images, a template is designed for each class with a closed form solution for the filter. Cross-correlation with a query image is then performed in the frequency domain where the assigned class is determined based on the parameters of the output correlation plane.

Many types of CFs have been designed with good properties such as distortion tolerance, noise robustness, gradual degradation and shift-invariance. CFs evolved from matched filters used in detecting a known reference image in the presesnce of additive white Gaussian noise. However, matched filters exhibit low detection rates even with slight changes of the reference image in terms of scale, rotation and pose. With the intention to deal with these limitations, the Synthetic Discriminant Function (SDF) filter and the Equal Correlation Peak SDF (ECP SDF) filter are introduced. Several training images are utilized as a linear combination to represent a single correlation filter. In this case, a pre-specified value called peak constraint is obtained by ECP SDF filter that corresponds to the authentic class or imposter class when an image is tested.

However, the pre-specified peak values lead to misclassifications when the sidelobes are larger than the controlled values at the origin. To address this problem, advanced correlation filters are introduced including Minimum Average Correlation Energy (MACE), Unconstrained Minimum Average Correlation Energy (UMACE), Optimal Trade-off Synthetic Discriminant Function(OTSDF), Unconstrained Optimal Trade-off Synthetic Discriminant Function (UOTSDF), Average of Synthetic Exact Filter (ASEF), Minimum Output Sum of Squared Errors (MOSSE) and Maximum Margin Correlation Filters (MMCFs) [10]-[11].

In this paper the MMCFs are considered as they have been shown to perform better than other well-known types of CFs.

MMCFs provide not only good classification but also localization as demonstrated successfully in tasks such as vehicle recognition, eye localization and face classification [12]. Although spectrograms representation has been reported [9], it uses a different type of correlation filter which is the Unconstrained Minimum Average Correlation Energy (UMACE) filter. Scalograms used with correlation filters have not been shown. In this paper, conversions to spectrograms and scalograms of anuran vocalizations are performed to obtain the images in order to compare their performance with MMCFs in classifying two anuran species.

2. IMAGE CONSTRUCTION

2.1 Spectrograms

In order to compare the performance of spectrograms and scalograms, the same pre-processing techniques are applied to the anuran vocalization recordings prior to applying the relevant time-frequency transformation. This involved using the recordings of of anuran vocalizations sampled at 44.1 kHz and segmenting them into individual calls of 800 ms length with a sound editing tool. Each segment was then filtered with a high pass filter with a cut-off frequency of 250 Hz in order to eliminate the environmental noise. The resulting calls was centered with the peak amplitude at 400 ms for each segment.

Framing was performed on each segment with a frame length of 256 and a 75 percent overlap. These parameters were obtained after several trials with the objective of obtaining visually clear spectrograms. The Gaussian window was applied, selected due to its superior performance in eliminating energy leakage at the expense of more intensive computations compared to other windows. The Gaussian window can be written as

$$h(n) = e^{-\frac{1}{2}\left(\frac{2.5n}{N/2}\right)^2},\ 0 \le |n| \le \frac{N}{2} \quad (1)$$

where *N* is the window length.

Each windowed frame is then transformed from the time domain signal *f* into the frequency-domain signal by the STFT coefficients

$$F_{(a,b)} = \langle f, h_{(a,b)} \rangle = \sum_{n=0}^{N-1} f[n] h^*_{a,b}[n] \quad (2)$$

The spectrogram is constructed from the square magnitude of the coefficients for a particular frequency and time.

2.2 Scalograms

Unlike the Fourier Transform, which decomposes a signal into sinusoids, the continuous wavelet transform (CWT) is used to construct scalograms. It uses a combination of basis function that are located in both the real and the Fourier space. The same preprocessing technique as described for spectrogram was used. After centering the peak amplitude, the scalogram was constructed using the discretized CWT with

$$CWT_f^{\Psi}(\tau, s) = \Psi_f^{\psi}(\tau, s) \quad (3)$$

and

$$\Psi_f^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \sum_{t=1}^{N} f(t) \psi^* \left(\frac{t-\tau}{s}\right) \quad (4)$$

where τ is the translation, s is the scale and ψ is the mother wavelet. For this paper, the Morlet mother wavelet was selected as it has been reported to give acceptable results for sound recognition [13]. It is defined as

$$\psi(t) = e^{jat} e^{\frac{-t^2}{2s}} \quad (5)$$

where *a r*epresents the modulation parameter. The scalogram is obtained from the squared value of the coefficients plotted in a time-frequency representation.

3. CORRELATION FILTERING

A correlation filter acts as a template representing a class and is constructed with several training images. In this case, a filter template is generated from multiple spectrograms and scalograms for each class with each image of size 512x512. Figure 1 shows an example of the vocalization image.

Figure 1: Training Image Example.

The MMCF is synthesized in the Fourier domain using a closed form solution. The optimization of the MMCF template is [12]

$$\boldsymbol{H}_{MMCF} = \bar{\boldsymbol{T}}^{-1} \frac{1}{L} \left(\sum_{i=1}^{L} \bar{\boldsymbol{X}}_i \boldsymbol{g}_i\right) + \bar{\boldsymbol{T}}^{-1} \boldsymbol{A} \bar{\boldsymbol{V}} \boldsymbol{a} \quad (6)$$

where $\bar{\boldsymbol{T}}$ is the trade-off matrix, $\bar{\boldsymbol{X}}_i$is a *d x d* diagonal matrix form of the *i*th training image in the frequency domain with vector $\boldsymbol{x}_i$ along its diagonal, $\boldsymbol{g}_i$ is the 2D vector representation of expected correlation output for the *i*th training image, $\boldsymbol{A}$ is a *d x L* matrix whose columns are formed by *L* training image vectors $\boldsymbol{x}_i$, $\bar{\boldsymbol{V}}$ is a diagonal matrix with class label (1 for true class, 0 for false class) along its diagonal while the vector $\boldsymbol{a}$ is evaluated from the sequential minimum optimization technique.

Cross-correlation is performed with the filter template and a test set to determine if the query image belongs to the class that the template represents. The MMCF optimizes a criterion to produce a desired correlation output plane by a trade-off

matrix that minimizes the localization criterion expressed as the mean square error while maximizes the margin criterion similar to Support Vector Machines (SVM).

For the input test image *S(x,y)* and a correlation filter template *H(u,v)*, the correlation process is given by

$$c(x,y) = IFFT\{FFT(S(x,y)).H^*(u,v)\} \quad (7)$$

where the test image is first transformed to frequency domain and reshaped to be in the form of a vector. It is then convolved with the conjugate of the MMCF filter which is equivalent to cross correlating it with the MMCF filter. The output is transformed again to the spatial domain obtaining the correlation plane. This process is shown in Figure 2.

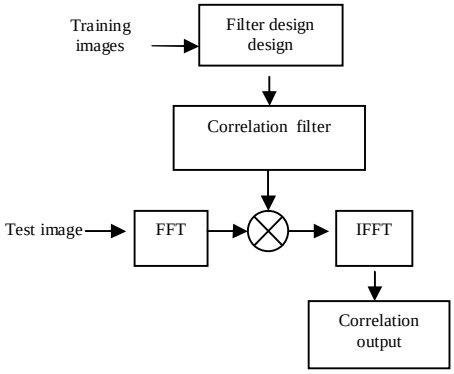

Figure 2: The Correlation Process

The resulting correlation plane produces a sharp peak at the origin while the values everywhere else are close to zero if the test image belongs to the same class as the template filter. The Peak-to-Sidelobe ratio (PSR) is used to measure the sharpness of the peak where

$$PSR = \frac{p-m}{\sigma} \quad (8)$$

where p is the largest value of the correlation output, σ is the standard deviation and m is the mean calculated from a sidelobe region excluding a central mask. The threshold value for each class is determined from the PSR values obtained with the cross-correlation process using the dataset. If a query image resulted in a PSR value that is greater than the threshold for the template class, it is classified as its true class, otherwise it is assigned as class false. The correlation plane output is shown in Figure 3.

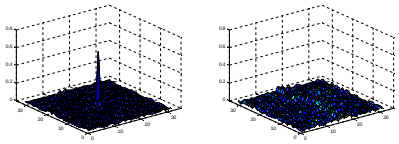

Figure 3: Examples of the Correlation Planes for Test Image from the True Class (Left) and False Class (Right).

4. EXPRIMENTAL RESULTS

Two species of anurans commonly found in Malaysia were considered for classification. Recordings from captured frogs were obtained for common grass frogs (*F. limnocharis Boie*) and mangrove frogs (*F. cancrivora Gravenhorst*) and the calls were subsequently processed as described in Section 2 to construct the spectrograms and scalograms. For each species, 40 images were obtained and the templates for each class were constructed as described in Section 3 using 10, 15 and 20 images from the training set. From the cross-correlation process of the test set and the templates, the accuracy rate is calculated defined as the ratio of correct classification to total number of test inputs. The results are tabulated in Table 1 and Table 2 for different number of training images per template for spectrograms and scalograms, repectively.

Table 1: Accuracy Rate for Spectrograms

Species	Number of training images		
	10	15	20
F. limnocharis Boie	45.9	67.5	70.6
F. cancrivora Gravenhorst	60.5	73.1	74.3
Average	53.2	70.3	72.5

Table 2: Accuracy Rate for Scalograms

Species	Number of training images		
	10	15	20
F. limnocharis Boie	20.7	43.7	47.5
F. cancrivora Gravenhorst	23.2	44.1	51.3
Average	22.0	43.9	49.4

The results show that spectrograms are better suited for anuran vocalization classification with MMCFs compared to scalograms. By using just 10 images, an accuracy rate of more than 50 percent is obtained for the spectrograms while scalograms cannot achieve this rate even with 20 images. In both cases however, the accuracy rate increases as more images are used to designed the template. The highest average accuracy of rate of 72.5 percent is obtained using spectrogram images. This is higher by more than 10 percent compared to results obtained by using UMACE filters with spectrograms as reported in [9].

Due to the limited dataset, it is not possible to test with a higher number of training images. Further experimentation with a longer dataset may increase the accuracy rate of both spectrograms and scalograms. Another possibility is to use a different mother wavelet for the scalograms in order to increase accuracy.

5. CONCLUSION

This paper has shown a comparison study between the use of spectrograms and scalograms with MMCFs to classify two species of anurans based on their vocalizations. In both cases, increasing the number of images to construct the template filter increases the accuracy rate. Despite using the Morlet mother wavelet to construct the scalograms, which have been found elsewhere to be suitable for sound classification, spectrograms with the Gaussian window are found more suitable for anuran vocalization classification. The accuracy rate is much higher in all cases considered using the parameters discussed for spectrograms compared to scalograms.

ACKNOWLEDGEMENT

The authors thank the Ministry of Education Malaysia for grant FRGS/1/2016/TK04/UKM/01/1.

REFERENCES

1. C.J. DeGarady and R.S Halbrook. **Using anurans as bioindicators of PCB contaminated streams**, *Journal of Herpetology*, vol 40, no.1, pp.127-130, 2006.
   https://doi.org/10.1670/30-05N.1
2. C.J. Huang, Y.J. Yang, D.X. Yang, and Y.J. Chen. **Frog classification using machine learning techniques**, *Expert Systems with Applications*, vol.36, no. 2, pp.3737-3743, 2009
   https://doi.org/10.1016/j.eswa.2008.02.059
3. B. Gingras and W.T. Fitch. **A three-parameter model for classifying anurans into four genera based on advertisement calls**, *The Journal of the Acoustical Society of America*, vol.133, no.1, pp.547-559, 2013.
   https://doi.org/10.1121/1.4768878
4. C.J. Huang, Y.J. Chen, H.M. Chen, J.J. Jian, S.C. Tseng, Y.J. Yang and P.A. Hsu. **Intelligent feature extraction and classification of anuran vocalizations**, *Applied Soft Computing*, vol.19, pp.1-7, 2014.
   https://doi.org/10.1016/j.asoc.2014.01.030
5. B. Gingras and W.T. Fitch. **A three-parameter model for classifying anurans into four genera based on advertisement calls**, *Journal of the Acoustical Society of America*, vol. 133, no.1, pp.547-559, 2013.
   https://doi.org/10.1121/1.4768878
6. W.P. Chen, S.S. Chen, C.C. Lin, Y.Z. Chen and W.C. Lin. **Automatic recognition of frog calls using a multi-stage average spectrum**, *Computers & Mathematics with Applications*, vol. 64, no.5, pp.1270-1281, 2012.
   https://doi.org/10.1016/j.camwa.2012.03.071
7. G. Grigg, A. Taylor, H. Mc Callum and G. Watson. **Monitoring frog communities: an application of machine learning**, in *Proc. Conf. Innovative Applications of Artificial Intelligence*, Portland Oregon, 1996, pp. 1564-1569,
8. J. Xie, M. Towsey, J. Zhang, X. Dong and P. Roe. **Application of image processing techniques for frog call classification**, in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4190-4194.
   https://doi.org/10.1109/ICIP.2015.7351595
9. S. Abdul Samad and A. Baseri Huddin. **Classifying anuran call spectrograms with correlation filters**, *International Journal of Engineering and Technology*, vol. 7, no. 4.16, pp.174-176, 2018.
10. Q. Wang, A. Alfalou and C. Brosseau. **New perspectives in face correlation research: A tutorial**, *Advances in Optics and Photonics*, vol. 9, no. 1, pp.1-78, 2017.
    https://doi.org/10.1364/AOP.9.000001
11. R.A. Kerekes and B.V. Kumar. **Selecting a composite correlation filter design: A survey and comparative study**, *Optical Engineering*, vol.47, no.6, p.067202, 2008.
    https://doi.org/10.1117/1.2943217
12. A. Rodriguez, V.N. Boddeti, B.V. Kumar, and A. Mahalanobis. **Maximum margin correlation filter: A new approach for localization and classification**, *IEEE Transactions on Image Processing*, vol. 22, no.2, pp.631-643, 2013.
    https://doi.org/10.1109/TIP.2012.2220151
13. M. Cowling and R. Sitte. **Comparisons of techniques for environmental sound recognition**, *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895-2907, 2003.
    https://doi.org/10.1016/S0167-8655(03)00147-8