ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

IJIRSET

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering (ICIET'14)
On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Automatic Identification of Bird Species from the Recorded Bird Song Using ART Approach

Deepika M, Nagalinga Rajan A

Infant Jesus College of Engineering, Keela Vallanadu., India
Infant Jesus College of Engineering, Keela Vallanadu.,India

ABSTRACT— our goal is to automatically identify which species of bird is present in an audio recording using spectrogram features. Birds song recognition approach is used to provide automatic investigation and remote monitoring of bird species population, which can provide the relevant agencies with sound information to habitat conservation as well as rare\endangered species survival plans and actions. The set of acoustic features developed for bird's song recognition was generally inspired by feature representations used in speech/speaker recognition or audio/music classification fields. In general these acoustic features are based on the acoustic model of speech production or the perceptual model of the human auditory system. Each spectrogram can be viewed as an image. Each spectrogram can be viewed as grey-level images. The new MPEG-7 Angular Radial Transform (ART) descriptor can be efficiently describing the grey-level variations within an image region in both angular and radial directions, to extract the shape features from the spectrogram image. Bird's song having distinct frequency and temporal variations will exhibit different shapes in their spectrogram. It extracts the shape features from the sound spectrograms of fixed duration bird's song segments. A sector expansion algorithm is proposed to transform its spectrogram image into sector image. It will align with the radial and angular directions of the ART basis function. A classification algorithm then employed to exactly identify the bird species based on the extracted features. GMM (Gaussian mixture models) then used for the classification to determine the bird species associated with the input birds song segment. For the classification of bird species using ART descriptor, is better than the traditional descriptor such as LPCC, MFCC and TDMFCC.

KEYWORDS—Spectrogram, Birds Song Recognition, MPEG-7, Angular and Radial Transform (ART), Sector Expansion Algorithm, Grey-Level Image, Gaussian Mixture Models (GMM).

I. INTRODUCTION

In daily life we can hear a variety of creatures including human speech, dog barks, birdsongs, frog calls, etc. Many animals generate sounds either for communication or as a by product of their living activities such as eating, moving, flying, mating etc. Bird species identification is a well-known problem to ornithologists, and it is considered as a scientific task since antiquity. Technology for Birds and their sounds are in many ways important for our culture. They can be heard even in big cities and most people can recognize at least a few most common species by their sounds. Biologists tried to investigate species richness, presence or absence of indicator species, and the population sizes of rare/ endangered species in a site. Collecting data from wild animals in outdoor environment and analysing them has been a tedious task that some field workers have little enthusiasm for , considering it difficult , dull and old-fashioned work. However modern technology has enabled scientists to accomplish this task in a way that was not possible some years ago. Many sounds archives consisting of a huge number of animal sound recordings, recorded in the fields by a lots of skilled and experienced investigators or biologists have been created for acoustic analysis. Birds are numerous and sensitive to environmental changes; also and are easier to monitor than other species. Birds are a good indicator for assessing habitat changes because they are distributed over a wide range of areas, are easy to detect in comparison with other animals and a significant amount of knowledge on bird diversity and behaviour was discovered through field observations bird diversity and behaviour was discovered through field observations by experienced birdwatchers and expert ornithologists. So that the spectrogram was a turning point in bird song research and it made it possible to analyse, measure, classify and recognize the different sound a bird makes. The set of acoustic features developed for birds song recognition was generally inspired by feature representations used in speech/speaker recognition or audio/music classification fields. Technology for sound-based identification of birds would be a significant addition to the research methodology on ornithology, and biology in general. There is also significant commercial potential for such systems because bird watching is a popular hobby in many countries. Bird vocalizations have been evolved to be species and individual recognition based on acoustic features. These acoustic features are based on the acoustic model of speech production or the perceptual model of the human auditory system. Some researchers have attempted to describe and spectral features of the acoustic signal. In general bird song not only corresponds to short-time timbrel characteristics but also to temporal structure of the sound signal i.e., the time evolution of bird sounds will provide some discriminating information for bird's song recognition. Therefore it would be valuable to apply image analysis methods to the identification of bird species based on their spectrogram images. Then MPEG-7 Angular and Radial Transform (ART) to extract the shape feature from the sound spectrograms of fixed duration birds song segments. It will extract discriminating features for bird song recognition.

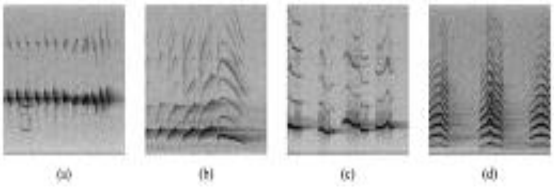

Fig .1. Different shapes of the spectrogram of different bird species. (a) Taiwan Firecrest. (b) Taiwan Sibia. (c) Vivid Niltava. (d) Crested Goshawk.

II. RELATED WORK

Traditionally visual inspection of the sound spectrogram or sonograms was one of the primary means for analysing bird song typically relies on the subjective judgements of experts. This process is extremely laborious, time consuming, and not entirely objective. Therefore it is impractical for large scale and long-term study of the population trends of different bird species. Traditional system uses different descriptor such as LPCC, MFCC, and TDMFCC[9]. Multi-Instance Multi-Label (MIML)[1] have been proposed for detecting the set of bird species present in an audio recording using the MIML framework, and propose a method to transform an audio recording into a representation suitable by using MIML algorithms. In this paper necessary to transform the data from its original representation into bag-of-instances representation and the proposed representation uses 2D time frequency segmentation of the audio signal, which can separate bird sounds that overlap in time. The main disadvantage of this system is MIML classifier can only predict labels that appear in their training data, and can't detect when something does not belong one of the training classes. Hence it is not clear how to handle unexpected sounds. Due to the high-noise environment, and birds vocalizing far from the microphone, it is often difficult for a human labeller to determine all of the species present in a recording. Consequently, some of the segments or instances may come from species that are not present in the training label set. The Markova chain frame independent syllable (MCFIS)[2] model is introduce the temporal structure within the syllable provides the significant amount of discriminative information and a new probabilistic models for identifying bird species from audio recordings, the independent syllable model and consider two ways of aggregating frame level features within a syllable. Each syllable as a probability distribution of its frame level features. The independent frame independent syllable (IFIS)[2] model allows us to distinguish syllable whose feature distributions are different from one another. Derive the bayas risk minimizing classifier for each model and show that it can be approximated as a nearest neighbour classifier. The main disadvantage is the independent syllable model does not capture temporal structure across syllables. This can be incorporated by assuming a Markova model for the distribution of syllable within an interval. The spectral peak track method appears to work as designed for isolated bird syllable and to conventional Dynamic Time Wrapping (DTW) [10] and Hidden Markova Model (HMM)[6] and Short- Time Fourier Transform (STFT)[5][7] methods used to classify the same database [3]. Manually extract one syllable from the recorded bird sounds and save the data in a separate sound file. The spectral analysis data demonstrate the usefulness of the sum-of-sinusoids model for rapid automatic recognition of isolated bird syllables. Set of spectral features by time variant analysis of the recorded bird vocalizations, then perform a calculation of the degree to which the derived parameters match a set of stored templates that were determined from a set of reference bird vocalizations. The main disadvantage is conventional methods based on a linear prediction model have difficulty with the sparse spectrum of the bird syllables in the test database are insufficient information is present in each syllable recording to create a unique and easily distinguishable LPC[6] model. The proposed method is inappropriate for use with bird vocalizations containing a periodic noise like components because the assumption of connected peak track is violated in these cases. The conventional model parameters are quite sensitive to existing back-ground noise, reverberation and competing sounds in the recordings. The segmentation of the songs was performed by analysing the autocorrelation and the roll-off of the songs. The bioacoustics methods developed for encoding birdsongs as sequences of discrete syllables are presented and syllables are extracted from segmented song recordings and their spectrograms are encoded. They are then classified using a new kind of artificial neural network which utilizes dynamic time wrapping for the learning stage. Perform sequence comparison is to use alignment algorithm that minimize a distance function between a pair of sequences because it allow us to perform unsupervised classification. Pair wise syllable distance measure, calculated on the basis of Mel-Cepstrum Co-efficient dynamic time wrapping[4] will be introduced where this distance is used for classifying syllable into different clusters. Retrieving cluster centres' from syllable data sets can be achieved using evolving neural networks and a distance measure based on dynamic time wrapping. Self-organizing maps are a kind of neural network designed for data representation and classification. The main disadvantage is during the learning stage, the network is trained with the data samples. However this process can be very time consuming and the size of the map has to be arbitrarily chosen. Bird song syllable classification is a difficult task and the development of automatic methods acknowledges that the true classification is unknown. Some other technique such as Multi Layer Perception (MLP)[8] , Sequential Minimization Algorithm (SMO), TDMFCC[9] can be used for bird species classification.

III. PROPOSED BIRD SONG RECOGNITION SYSTEM

Our goal is to develop algorithms that can predict which species of bird is present in an audio recording. In this proposed system it mainly focuses on automatic recognition of fixed duration of bird song segment. This segment may be 3 or 5 seconds. Each bird song segment is an elementary unit for the identification of bird species. Each elementary unit can be divided into several overlapping texture window. To reduce the noise or background sound, each texture window is first computed.

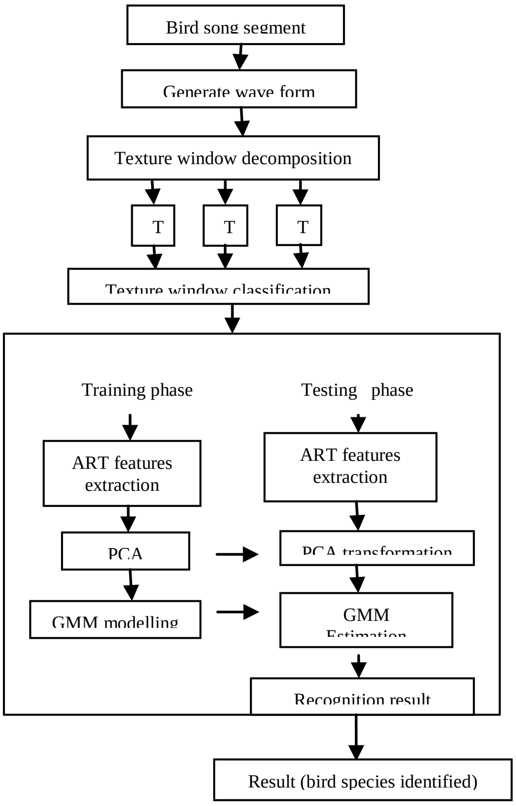

Fig .2. The block diagram of the proposed bird song recognition system.

The texture window duration is 2-second and the offset between two neighbouring texture windows is 0.25 second. Therefore each 3 or 5 –second bird song segment consists of 5 or 13 texture windows. The texture window and its power less than the largest one among all training texture window by 40 dB is called as silent texture window. Otherwise it is called as an active texture window. The classification of each texture window consists of two phases: the training phase and the recognition phase or testing phase. The training phase consists of 3 modules that are: ART feature extraction, principal component analysis (PCA), and GMM modelling. The testing phase consists of 3 modules that are ART feature extraction, PCA transformation, GMM likelihood estimation.

IV. ART FEATURE EXTRACTION

The MPEG-7 standard comity proposes a new region based shape descriptor that is Angular Radial Transform (ART). ART is a moment-based image description method adopted in MPEG-7 as a region-based shape descriptor. It gives a compact and efficient way to express pixel distribution within a 3D object region. The ART is a complex orthogonal unitary transform defined on a unit disk that consists of the complete orthogonal sinusoidal basis functions in polar coordinates. The ART coefficients, F_{nm} of order n and m are defined by

$$F(n,m) = (V_{n,m}(\rho,\theta), I_S(\rho,\theta))$$

$$=\int_0^{2\pi}\int_0^1 V_{n,m}(\rho,\theta)\, I_S(\rho,\theta)\, \rho d\rho d\theta$$

To capture both spectral and temporal structure present in spectrogram of each bird song and it will exploit the MPEG-7 ART descriptor

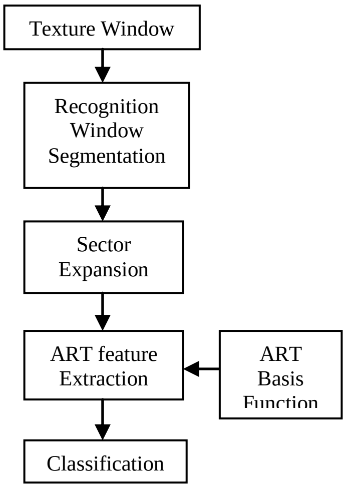

Fig.3. The block diagram for ART feature Extraction.

to extract the discriminating features for bird song recognition. First given an active texture window, it is first divided into a sequence of overlapping frames of length 512 samples (i.e., 11.6 ms for sampling frequency 44,100Hz) and offset by 256 samples between two neighbouring frames. From the frame with the maximum power and its preceding 127 frames and succeeding 128 frames are segmented to form a fixed length recognition window (fig). If the number of preceding frames is less than 128, zero frames are appended to obtain a fixed length recognition window of 256 frames.

The DFT magnitude of each frame consisting of 256 bins for 256 frame recognition window is computed first. Collection of DFT magnitudes of all frames within the recognition window will form the spectrogram. The spectrogram can be viewed as a 256×256 image is called the recognition image, which consists of acoustic intensity within the texture window. The ART descriptor will be extracted by convolving a set of complex ART basis functions with the transformed sector image. It has real and imaginary parts of set of 12×12 ART basis function Fig.4 From this basis function, the ART descriptor captures the shape variations within an image region in both angular (m) and radial direction (n). To capture the shape variations using ART is called sector expansion algorithm, to map the recognition image into another transformed image called sector image, such that the frequency axis and temporal axis of the recognition image will respectively align with the radial variable *n*, and angular variable *m* of the ART basis function Fig. 4. For example the *r*th row of the recognition image, consisting of the *r*th DFT frequency bins of 256 consecutive frames, will be mapped to the circle of radius (256-*r*) in the sector image; each radial line in the sector image corresponds to the magnitude spectrum of a specific frame i.e., vertical column in the recognition image. That is the magnitude spectrum of the first frame becomes the radial line with angle 0, the magnitude spectrum of the second frame is the radial line with angle (2π/256), the magnitude spectrum of the third frame is the radial line with angle 2× (2π/256) and so on...

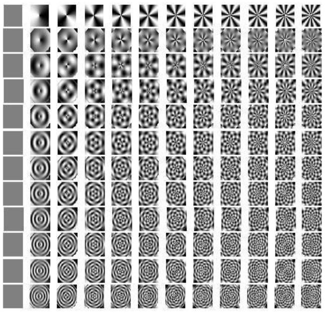

Fig. 4. The 12×12 (N = 12 and M= 12) imaginary part of complex ART basis function.

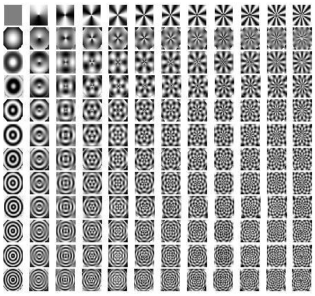

Fig .5. The 12×12 (N = 12 and M = 12) real part of complex ART function.

As a result, each vertical stripe in the recognition image will be mapped to a sector in the sector image, and each horizontal stripe will be mapped to a ring and the whole recognition image will be mapped to the disk of radius 256 Fig.3. To effectively exploit the merits of the ART basis functions with the transformed sector image instead of the recognition image.

A. Sector Image Generation

Let I(f,t) denotes the recognition image segmented from the active texture window , where f(0 ≤ f ≤ 255) and t(0 ≤ t ≤ 255) denote the frequency variable and temporal variable/frame respectively. In this paper, the proposed sector expansion algorithm to amp the recognition image I(f,t) into the sector image $I_s(u,v)$, for 0≤u, v≤511, such that the frequency variable (f) and temporal variable (t) will be respectively associated with the radial variable

($\rho, 0 \le \rho \le 255$) and the angular variable ($\theta, 0 \le \theta \le 2\pi$) of the ART basis functios. The sector image is of size 512×512, which is larger than that the recognition image (256×256) and based on this mapping operation, the variations in the frequency direction as well as the temporal direction in the recognition image can be effectively captured by convolving the ART basis functions with the transformed sector image. The transformation of the recognition image I(f,t) to the sector image $I_s(u,v)$ is defined as follow Fig.6.

$$u = 256 - \Delta u$$
$$v = 256 + \Delta v$$

where

$$\Delta u = \rho \sin\theta$$
$$\Delta v = \rho \cos\theta$$

and

$$\rho = 256 - f$$
$$\theta = 2\pi \times (t/256)$$

The implementation of the mapping operation is realized by using the following inverse mapping function where the pixel value locating at (u,v) in the sector image is determined by finding its corresponding pixel value locating at (f,t) in the recognition image:

$$f = 256 - \rho = 256 - \sqrt{(\Delta u)^2 + (\Delta v)^2}$$

$$= 256 - \sqrt{(256-u)^2 + (256-v)^2}$$

$$t = \frac{256}{2\pi} \times \theta = \frac{256}{2\pi} \tan^{-1}\frac{\Delta u}{\Delta v}$$

$$= \frac{256}{2\pi} \tan^{-1}\left(\frac{256-u}{v-256}\right)$$

According to this inverse mapping function, the grey value of each pixel in the sector image can be determined by finding the corresponding pixel value in the recognition image.

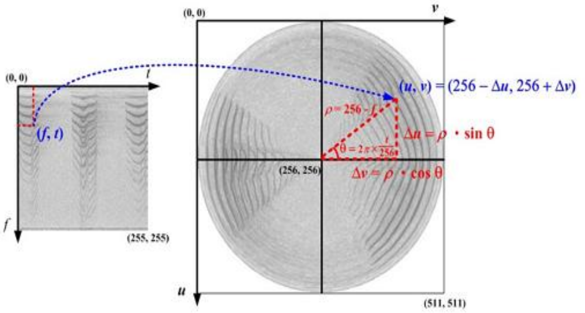

Fig .6.The transformation of the recognition image to the sector image.

B. ART Descriptor Extraction

The ART descriptor is defined as a set of normalized magnitudes of the set of ART coefficients. Feature extraction is a special form of dimensionality reduction. When an input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this ART descriptor instead of the full size input. In the ART descriptor to achieve rotational invariance, an exponential function is used for the angular basis function and the radial basis function is defined by a cosine function.

$$A_m(\theta) = (1/2\pi)e^{in\theta}$$

$$Rn(\rho) = \begin{cases} 1, & n = 0 \\ 2\cos(2\pi n\rho), & n \neq 0 \end{cases}$$

Assuming that the sector image is represented in polar form, denoted by $I_s(\rho,\theta)$ and the radius variable ρ is normalized in the range of [0,1]. To derive the ART descriptor, it can convolve a set of ART basis function $V_{n,m}(\rho,\theta)$, $0 \le n \le N$ and $0 \le m \le M$,with the sector image:

$$F(n,m) = (V_{n,m}(\rho,\theta), I_s(\rho,\theta))$$

$$= \int_0^{2\Pi} \int_0^1 V_{n,m}(\rho,\theta)\, I_S(\rho,\theta)\rho d\rho d\theta$$

where $V_{n,m}(\rho,\theta)$ is the ART basis function of order n and m, which is separable along the angular and radial directions:

$$V_{n,m}(\rho,\theta) = A_m(\theta)R_n(\rho)$$

As a result, N×M ART coefficients, denoted by F(n,m) for $0 \le n < N$ and $0 \le m < M$, can be obtained. The magnitude of each ART coefficient, |F(n,m)|, is then divided by |F(0,0)|:

$$\frac{|F(n,m)|}{|F(0,0)|},\ 0 \le n < N, 0 \le m < M$$

Finally, these scaled ART coefficients, except f(0,0), will form the ART descriptor:

$$f_{ART} = [f_{ART}(1), f_{ART}(2), \ldots\ldots\ldots f_{ART}(N \times M-1)]^T$$

$$= [f(0,1), \ldots\ldots\ldots\ldots. f(0,M-1),\ f(1,0), \ldots.\ f(1,M-1),\ \ldots\ldots$$

$$f(N-1,0),\dots\dots\dots f(N-1,M-1)]^T$$

Fig.6 shows the recognition image, sector images and the 3D plots of the magnitudes of the ART coefficients . The horizontal stripes in the recognition or two rings in the sector image, which represents the temporal rhythmic information of the birdsong. Thus, the3D plot of the ART coefficients exhibits some peak values for angular variable *m*=16. Therefore the 3D plot of the ART coefficients exhibits some peak values when the angular variable *m* is a multiple of 4 and the radial variable *n* is a multiple of 5. As the result of the ART descriptor can capture both radial and angular variations within the sector image correspondingly the spectral harmonic and temporal rhythmic information in the input birdsong.

V. **PRINCIPAL COMPONENT ANALYSIS (PCA)**

PCA has been widely used for dimensionality reduction. Feature extraction is also special form of dimensionality reduction. It involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfit the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Best results are achieved when an expert constructs a set of application–dependent features. PCA is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components should be lesser than or equal to the number of original variable. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate is called the first principal component, the second greatest variance on the second coordinate and so on... PCA transforms the data from a higher dimensional vector space into a lower dimensional space such that the variance of the projected data is maximized. The D×D covariance matrix Σ is computed for the set of D-dimensional training feature vectors. The eigenvectors and corresponding eigen values of the covariance matrix Σ are computed and sorted in a decreasing order of the eigen values.

Let eigen vector V_i be associated with eigen value ʎ_i, $1 \leq i \leq D$. The first d eign vectors having the largest eign values will form the d columns of the D×d transformation matrix A_{PCA};

$$A_{PCA}=[V1,V2........Vd]$$

The number of selected eign vectors d can be determined by finding the minimum integer that satisfies the following criterion:

$$\sum_{j=1}^{d} \text{ʎj} \geq \alpha \sum_{j=1}^{D} \text{ʎj}$$

where α is a parameter that determines how many percentage of information need to be preserved. The projected vector can be obtained by using the transformation matrix A_{PCA}:

$$X_{PCA}=A^{T}{}_{PCA}{}^{X}$$

VI. **GAUSSIAN MIXTURE MODELS (GMM)**

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm. In general, the texture windows segmented from the same bird species may exhibit distinct spectral and/or temporal characteristics. That is, two texture windows segmented from identical bird species might differ significantly. Consequently, the feature vectors extracted from the texture windows of identical bird species will reveal many isolated manifolds in the feature space. Gaussian Mixture Model (GMM) has been extensively used as the speaker model in speaker recognition systems in which the feature vectors of each speaker is represented by a number of Gaussian components. GMM will be used to model the bird song of each bird species as a weighted combination of Gaussian Component probability density functions.

In GMM, the mixture density of a feature vector $X \, \mathcal{E} \, R^d$ can be represented as a mixture of G unimodal Gaussian densities:

$$P(X|\Theta) = \sum_{g=1}^{G} \pi_g \, \rho(X|\theta_g)$$

where $\Theta=\{ \pi_g,\theta_g \; g=1, \dots\dots, G\}$ is the set of parameters of the GMM model, π_g, the mixture weight associated with the g^th Gaussian density $p(X \; \theta_g)$, is subject to the following constraints:

$$\pi_g \geq 0 \text{ and } \sum_{g=1}^{G} \mathit{\Pi}_g=1$$

the g^th Gaussian component with parameters $\theta_g=\{\mu_g, \Sigma_g\}$ is modelled by a d-variate Gausssian distribution:

$$p(x|\theta_g) = (1/((2\pi)^{d/2}|\Sigma_g|^{1/2})\times\exp(-1/2(x-\mu_g)^T(\Sigma_g)^{-1}(x-\mu_g)$$

where μ_g and Σ_g are respectively the mean vector and covariance matrix of the g_{th} Gaussian component. To reduce the computational complexity, use diagonal covariance matrix for each Gaussian component. Each GMM is specified by the set of parameters $\Theta = \{\pi_g, \mu_g, \Sigma_g, \mid g = 1.....G\}$. each bird species has a unique GMM, representing the particular distributions of its feature vectors.

Give an initial set of parameters Θ^0, the Em algorithm iteratively re-estimates the parameters to obtain Θ_{ML}. At each iteration, the EM algorithm consists of two steps: the expectation step (E-step) and the maximization step (M-step), which are alternatively implemented until the log-likelihood of the training set X modelled by the GMM with parameter set Θ converges to a local minimum. The performance of the EM algorithm depends on the choice of the initial set of parameters Θ^0. The *k* - means clustering algorithm is used to find the initial parameters. The EM algorithm for learning GMM parameter set is followed:

Step 1: **Initialization.** Divide the set of training vector

Step 2: **Expectation - Step.** For each vector , X_i^s, computing the posterior probability associated with each Gaussian component.

Step 3: **Maximization – Step.** Update the GMM parameters using the posterior probability.

Step 4: If the log-likelihood function log $p(X^s|\Theta^i)$ converges, exit the EM algorithm. Otherwise, set i = i+1 and go to step 2.

VII. EXPERIMENTAL RESULT

A new descriptor is proposed to identify the bird species in the recorded bird song. The birdsong database, which contains the bird sounds of bird species, collected from commercially available on compact disks and the internet. Most of the sounds are field recordings with additional background sounds or noises. Some recordings contain bird song vocalized by multiple individuals. The proposed ART descriptor is individually independent for bird species identification, the testing and training data are selected from different recordings in this experiment. The sampling frequency is 44100Hz with each sample digitized in 16-bit accuracy. In this paper experiments on audio segments with different lengths, including 3-second and 5-seconds audio segments are tested to measure the performance. The performance is measured in terms of the classification accuracy (CA) defined as follow:

$$CA = (N_{CA}/N_{TW}) \times 100$$

Where N_{CA} is the number of texture windows which were recognized correctly and N_{TW} is the total number of test texture windows. Fig.7 shows the screen shot for texture window and sector image and Fig.4 and 5 shows the screen shot for ART imaginary and real part basis function. The best classification accuracy is 98.6% using newly proposed ART descriptor.

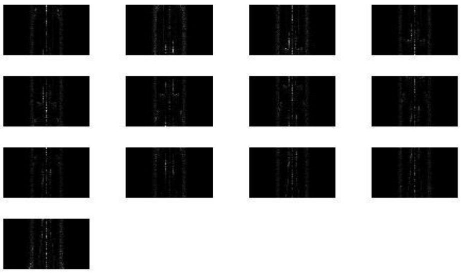

Fig. 7. Texture Window for bird song.

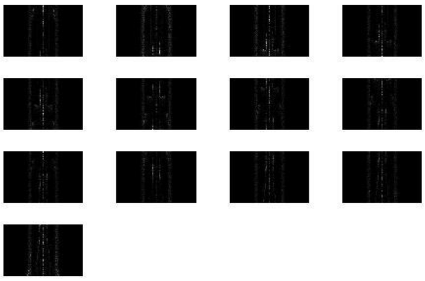

Fig. 8. Sector Image for bird song.

VIII. CONCLUSIONS

A new feature descriptor is proposed for the identification of bird species from the birdsong they vocalize. A fixed song duration of bird song segment is an elementary unit for identification of bird species. The input of bird song segment is first divided into a several overlapping texture windows. Each spectrogram image can be viewed as an image. The ART image descriptor is proposed to extract the discriminating features, which can be describe both spectral and temporal structure of each texture window, from the transformed spectrogram image and that can be converted into sector image using sector expansion algorithm such that the frequency and temporal axes of the spectrogram image will align with the radial and angular directions of the ART basis function respectively. Sector expansion algorithm is proposed to map the recognition image into the sector image such that the frequency variable and the temporal variable will be respectively associated with the radial and angular variable of the ART basis functions and that mapping operation extract the variations in the frequency direction as well as the temporal direction in the recognition image can be effectively captured by convolving the ART basis functions with the transformed sector image. Then PCA has been widely used for dimensionality reduction and defined as the orthogonal projection that transforms the data from a higher dimensional vector space onto a lower dimensional space. After reduce the dimensionality from the resultant data, GMM is then employed to classify the set of ART descriptors extracted from the input bird song segment in a minimal likelihood estimation sense. Based on the results of this study, it seems that most bird species are easily to be recognized using the proposed ART descriptor, where as some species result in higher classification errors. Due to the limited sets of training and testing data, the experimental results cannot be considered representative. To make the results more convincing, it is necessary to collect a larger set of training and testing recordings in the near future. In general, most bird sounds are recorded in noise-intense and adverse environments. However, from the experimental results, found that the proposed ART descriptor is not robust regarding background sounds/noises due to the fact that it is obtained by convolving the ART basis functions with the noisy spectrogram image. Thus the future research direction will focus on reducing the impact of noise, by using acoustical model compensation approaches to reduce the noise present in each birdsong, or by designing a noise robust ART descriptor. In addition, no temporal information among texture windows is captured in the proposed system.

REFERENCES

[1] F.Briggs, R.Raich, and X.Z.Fern, "Audio classification of bird species: A statistical manifold approach," inProc. 9th IEEE Int.Conf.Data Mining, 2009, pp.51-60.

[2] B. Lakshminarayanan, R. Raich, and X. Fern, "A syllable-level probabilistic framework for bird species identification," inProc. IEEE Int. Conf. Mach. Learn. Appl., 2009, pp. 53–59.

[3] Zhixin Chen and Robert C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," J. Acoust. Soc. Amer.,vol.120,no. 5, pp. 2974–2984, Nov. 2006.

[4] E.E.Vallejo, M.L.Cody, and C.E.Taylor, "Unsupervised acoustic classification of bird species using hierarchical self-organizing maps," in Proc. 3rd Australian Conf. Progress Artificial Life, 2007, pp.212–221.

[5] Aki Härmä, "Automatic Identification Of Bird Species Based On Sinusoidal Modeling Of Syllables", IEEE Int. Conf. Acoust., Speech,Signal Process, 2009. paper 11.3.4, p. 109.

[6] Wei Chu and Daniel T. Blumstein, "Noise robust bird song detection using syllable pattern-based hidden markov models",. in Proc. of International Conference Acoustic Speech Signal Process, 2011.

[7] Edward F. Connor, Shidong Li and Steven Li ," Automating identification of avian vocalizations using time–frequency information extracted from the Gabor transform", J. Acoust. Soc. Am.132(1), July 2012, pp. 507–517.

[8] Marcelo T. Lopes, Lucas L. Gioppo, Thiago T. Higushi, Celso A. A. Kaestner , Carlos N. Silla Jr and Alessandro L. Koerich, "Automatic Bird Species Identification for Large Number of Species ", 2011 IEEE International Symposium on Multimedia.

[9] Chang-Hsing Lee, Jau-Ling Shin , Sheng-Bin Shiu, " Automatic Recognition Of Bird Species From Continuous Birdsong Recordings ", inProc. inProc. Int. MultiConference Eng. Comput. Scientists,Hong Kong, 2010.

[10] Mosamee Gund, Aarti Bang, "Classification of Bird Species", International Journal of Electronics, Communication & Soft Computing Science and Engineering ISSN: 2277-9477, Volume 2, Issue 4, 2010.

[11] Yushan National Park, CD Sound of the Mountain V: The songs of Wild Birds, Taiwan, 1996.

[12] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models,"IEEE Trans.Speech Audio Process., vol. 3, no. 1, pp. 72–83, Jan. 1995.

[13] R.Duda,P.Hart,andD.Stork, Pattern Classification.NewYork:Wiley, 2000.

[14] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification,"IEEE Trans. Signal Process., vol. 52, no. 10, pp. 3023–3035, Oct. 2004.

[15] B. S. Manjunath, P. Salembier, and T. Sikora,IntroductiontoMPEG-7: Multimedia Content Description Interface. New York: Wiley, 2002.

[16] C. H. Lee, J. L. Shih, K. M. Yu, and J. M. Su, "Automatic music genre classification using modulation spectral contrast feature," in Proc.IEEE Int. Conf. Multimedia and Expo, 2007, pp. 204–207.

[17] C.H.Lee,J. L. Shih,K.M.Yu,andH.S.Lin,"Automatic music genre classification based on modulation spectral analysis of spectral and cep-stral features,"IEEE Trans. Multimedia, vol. 11, no. 4, pp. 670–682, Jun. 2009.

[18] R. Bardeli, "Similarity search in animal sound database,"IEEE Trans. Multimedia, vol. 11, no. 1, pp. 68–76, Jan. 2009.

[19] T. A. Parker, III, "On the use of tape recorders in avifaunal surveys," Auk, vol. 108, pp. 443–444, 1991.

[20] H.Tyagi,R.M.Hegde,H.A.Murthy,andA.Prabhakar,"Automatic identification of bird calls using spectral ensemble average voice prints," inProc. 13th European Signal Process. Conf. (EUSIPCO'06), Florence, Italy, Sep. 2006.

[21] S.E.Anderson,A.S. Dave,and D. Margoliash,"Template-based auto-matic recognition of birdsong syllables from continuous recordings," J. Acoust. Soc. Amer., vol. 100, no. 2, pp. 1209–1219, Aug. 1996.

[22] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," inProc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2004, vol. 5, pp. 825–828.

[23] A. L. McIlraith and H. C. Card, "Birdsong recognition using backprop-agation and multivariate statistics,"IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2740–2748, Nov. 1997.

[24] A.L.McIlraithandH.C.Card,"Birdsongidentification using artifi-cial neural networks and statistical analysis," inProc. Canadian Conf. Elect. Comput. Eng., 1997, vol. 1, pp. 63–66.

[25] J. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study,"J. Acoust. Soc. Amer., vol. 103, no. 4, pp. 2187–2196, Apr. 1998.

[26] C. H. Lee, Y. K. Lee, and R. Z. Huang, "Automatic recognition of bird songs using cepstral coefficients,"J. Inf. Technol. Appl., vol. 1, no. 1, pp. 17–23, May 2006.

[27] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition,"IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 6, pp. 2252–2263, Nov. 2006.

[28] A. Fagerlund, "Bird species recognition using support vector ma-chines,"EURASIP J. Adv. Signal Process., vol. 2007, no. Article ID 38637, 8 pp.