Bird Species Classification Using Visual and Acoustic Features Extracted from Audio Signal

Diego Rafael Lucio
Programa de Pós Graduação em Ciência da Computação
Universidade Estadual de Maringá
Avenida Colombo, 5790 - Jardim Universitário,
Maringá - Paraná - Brasil
Email: diegorafaellucio@gmail.com

Yandre Maldonado e Gomes da Costa
Programa de Pós Graduação em Ciência da Computação
Universidade Estadual de Maringá
Avenida Colombo, 5790 - Jardim Universitário,
Maringá - Paraná - Brasil
Email: yandre@din.uem.br

***Abstract*—This work aims to present a system for automatic bird species classification based on acoustic and visual features extracted from the birdsong. The visual features are extracted from spectrogram images generated from the birdsong audio, while the acoustic features are taken directly from the audio. Texture descriptors were used to describe the spectrogram content, as this is the main visual content found in this kind of image. The texture operators used are Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Robust Local Binary Pattern (RLBP), Gray-Scale Level Co-occurrence Matrix (GLCM) and Gabor filters. The acoustic features are, in turn, described using Rhythm Histogram (RH), Rhythm Patterns (RP) and Statistical Spectrum Descriptor (SSD.) Aiming to perform more fare comparisons, the experiments performed were made on a similar database already used in other works. In the classification step, SVM classifier was used and the final results were taken by using 10-fold cross validation. And over all performed tests the combination between acoustic and visual features produce the best rate of this work 91.08%.**

***Index Terms*—Bird species classification, Spectrograms, Pattern recognition, Signal processing, Machine learning, Information retrieval, Information extraction.**

I. Introdução

Devido ao fato de haver a necessidade da conservação da fauna e da flora, alguns estudos foram realizados tendo por finalidade estudar a biodiversidade das espécies de pássaros. Fato este se deve pelos importantes papéis desempenhados pelos mesmos, visto que estes são responsáveis por tarefas como: controle de insetos [1][2], dispersão de sementes[3][4] e polinização [5][6][7].

O conhecimento da biodiversidade supracitada é alcançado quando se identifica as espécies de pássaros presentes em um dado bioma. Sendo assim, o uso de técnicas baseadas em bioacústica passou a ser utilizado na identificação dos pássaros a partir dos registros de áudio capturados na natureza, o que é caracterizado pela eficácia apresentada pelas mesmas [8] [9].

Neste contexto, é oportuno o desenvolvimento de pesquisas relacionadas ao reconhecimento automático de espécies de pássaros, baseadas em registro de áudio, tendo como base estes registros para a criação de uma base de dados para o desenvolvimento de um sistema de classificação. Sendo assim, o presente trabalho é voltado para a tarefa de classificação automática de espécies de pássaros fazendo o uso da combinação de características acústicas e visuais, obtidas a partir dos cantos dos pássaros.

978-1-5090-3339-3/16/$31.00 ©2016 IEEE

As características visuais são obtidas a partir de um espectrograma, que é uma representação visual do espectro das frequências do som. No seu formato mais comum, é representado por um gráfico em que o eixo horizontal representa o tempo e o eixo vertical representa a frequência. A amplitude é representada em uma terceira dimensão, descrita pela intensidade da cor de cada ponto da imagem.

Já as características acústicas são obtidas diretamente a partir do sinal de áudio tendo por finalidade analisar a essência de uma amostra, para assim descrever características como: o ritmo, o timbre e o tom da mesma.

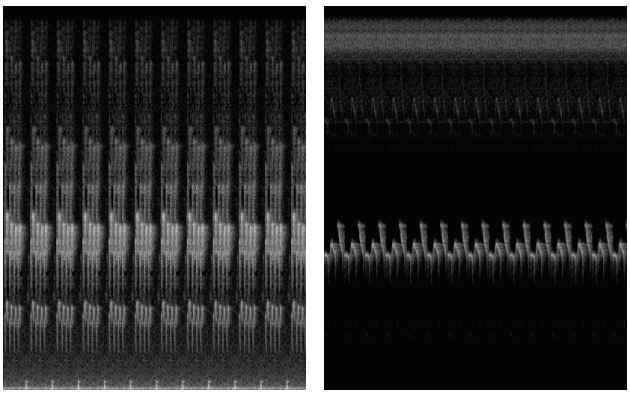
A Figura 1 ilustra dois exemplos de espectrogramas obtidos a partir do sinal de áudio das espécies Automolus Leucophthalmus e Sittasomus Griseicapillus, respectivamente. Como se pode observar as imagens geradas para cada uma das espécies refletem características diferentes, relacionadas inclusive a diferenças entre conteúdos harmônicos.

De acordo com a problemática supra citada foi definida que a hipótese deste trabalho é a de que é possível representar o canto de um pássaro através de características acústicas e visuais, com o propósito de criar um sistema de reconhecimento automático de espécies. Investigou-se também a hipótese de que, neste cenário a natureza do sinal é integralmente abstraída, sendo possível alcançar resultados similares ou superiores aos alcançados por métodos utilizados em trabalhos já publicados tendo por objetivo apresentar a complementariedade encontrada quando se combinou classificadores obtidos por meio do uso de características acústicas com classificadores obtidos com a utilização de características visuais.

A utilização de características acústicas e visuais é justificada pelo fato de que cada tipo de característica poder capturar informações diferentes decorrente do fato de trabalharem em domínios diferentes, o que leva a crer que pode haver certa complementaridade entre as técnicas, o que pode acarretar a criação de uma metodologia de reconhecimento mais eficaz.

A principal contribuição do presente trabalho foi identificar uma nova abordagem para a classificação de espécies

de pássaros. Por meio da realização das seguintes etapas: obtenção de imagens de espectrogramas a partir de um sinal de áudio, extração das características de textura dos espectrogramas, extração das características acústicas a partir do sinal de áudio e a utilização de classificadores sobre as características extraídas.



(a) Espectrograma criado a partir de uma amostra de áudio da espécie de pássaro Automolus Leucophthalmus

(b) Espectrograma criado a partir de uma amostra de áudio da espécie de pássaro Sittasomus Griseicapillus

Figura 1: Exemplos de espectrogramas gerados a partir dos cantos dos pássaros

Este trabalho encontra-se organizado da seguinte forma: a seção II, apresenta a revisão da literatura realizada buscando encontrar o estado da arte acerca do tema de reconhecimento automático de espécies de pássaros, por sua vez a seção III apresenta os fundamentos teóricos utilizados para desenvolver o método de classificação proposto. A seção IV apresenta uma breve descrição sobre a base de dados utilizada, e a seção V apresenta o método de classificação elaborado. Enquanto a seção VI apresenta os resultados obtidos e as discussões acerca dos mesmos, e por fim a seção VII apresenta as conclusões obtidas com a execução das atividades aqui apresentadas.

II. Conceitos e Trabalhos Relacionados

O interesse no reconhecimento de espécies de pássaros baseado na sua vocalização tem aumentado e muitos estudos recentes têm sido publicados. O reconhecimento de espécies de pássaros é um típico problema de reconhecimento de padrões e a maioria dos estudos incluem processamento de sinais, extração de características e elaboração de um sistemas de classificação.

Os trabalhos apresentados por Anderson et al. [12] e Kogan and Margoliash [14] estão entre as primeiras tentativas para o reconhecimento automático de espécies de pássaros por meio de sons emitidos pelos mesmos. O primeiro trabalha com *Dynamic Time Warping* (DTW) que é um algoritmo utilizado para comparar sequências que variam com o tempo, enquanto o segundo faz o uso da técnica citada, assim como também utiliza *Hidden Markov Models* (HMM) para o reconhecimento automático de duas espécies de pássaros. Em ambos os estudos são criados *templates* das amostras de áudio com base na vocalização das espécies, que é segmentada em sílabas, parágrafos e frases. Após a criação dos templates os mesmos são utilizados como entrada em um sistema de *template matching*, que faz o uso de DTW e HMM para a classificação. Os resultados apresentados pelos trabalhos foram respectivamente 97,00% em uma base de dados composta por 2 espécies e 82,00% em uma base de dados composta por 6 espécies.

McIlraith and Card [13], Selouani et al. [15], Cai et al. [18], e Chou and Liu [22] fazem o uso de redes neurais para realizar a classificação das espécies de pássaros. [13] realizaram o reconhecimento de sons de seis espécies de pássaros. Nesse trabalho foram utilizados como características parâmetros temporais e espectrais obtidos por meio da utilização da *Fast Fourier Transform* (FFT), sendo que sua melhor taxa de acerto foi de 82,00% em uma base de dados composta por 6 espécies.

Selouani et al. [15] utilizaram uma abordagem de rede neural chamada *Time Delay Neural Network* (TDNN) com *Autoregressive Backpropagation* para realizar a classificação das espécies de pássaros. Como entrada do sistema foram utilizados *templates* extraídos dos registros de áudio das 16 espécies de pássaros presentes na base de dados utilizada, sendo que a melhor taxa de acerto foi de 83,00%.

O trabalho apresentado por Cai et al. [18] utilizou *Mel-frequency Cepstral Coefficients* (MFCC) como características para o sistema de classificação para as duas bases de dados utilizadas: uma composta por 4 e a outra por 14 espécies de pássaros, as taxas de acerto para cada uma das bases foram respectivamente, 98,70% e 86,80%.

Briggs et al. [21] também utilizaram o MFCC como características das amostras de áudio. Além do MFCC os autores também fizeram o uso do do Mean Frequency Bandwidth (MFB) e da densidade do espectro como características. A etapa de classificação fez uso de K-NN, SVM e distâncias de KULLBAK-LEIBLER e HELLINGER, sendo que a melhor taxa de reconhecimento obtida para a classificação das 6 espécies presentes na base de dados foi de 91,10%.

No trabalho apresentado por Chou and Liu [22] também fez-se o uso do MFCC como característica para as amostras de áudio das 420 espécies presentes na base de dados utilizada obtendo-se uma taxa de acerto de 18,72%.

Kwan et al. [16] utilizaram *Guassian Mixture Model* (GMM) para a etapa de classificação, sobre uma base de dados composta por 11 espécies, que tiveram seus cantos representados por MFCC. Neste trabalho também foi apresentado um sistema para monitoramento automático de pássaros na natureza, com a taxa de acerto de 90%.

Chou et al. [19] também fizeram o uso de HMM assim como os trabalhos citados acima, no entanto, este é utilizado como conjunto de características do sistema de classificação, e não como classificador. Os autores utilizaram uma base de dados composta por 420 espécies, sendo que a taxa de acerto alcançada foi de 78,20%.

Um outro exemplo da utilização do GMM na classificação foi apresentado por Lee et al. [20], no qual foram utilizados como características o *Two-dimensional Mel-frequency*

Tabela I: Relação de trabalhos e seus respectivos conjuntos de características e classificadores

Trabalho	Características	Classificador	Qtde de espécies	Melhor Acerto
Anderson et al.	Templates criados a partir das amostras de áudio	HMM e DTW	2 espécies	97,00%**
McIlraith and Card	Características extraídas com FFT	Redes Neurais	6 espécies	82,00%**
Kogan and Margoliash	Templates criados a partir das amostras de áudio	HMM e DTW	2 espécies	98,70%**
Selouani et al.	Templates criados a partir das amostras de áudio	TDNN com Autoregressive Backpropagation	16 espécies	83,00%**
Kwan et al.	MFCC	GMM	11 espécies	90,00%**
Vilches et al.	Características acústicas	VQ, ID3, J4.8 e Naive Bayes	3 espécies	98,39%**
Fagerlund	MFCC e parâmetros descritivos de sílabas	SVM	6 espécies 8 espécies	93,00%** 97,00%**
Cai et al.	MFCC	Redes Neurais	4 espécies 14 espécies	98,70%** 86,80%**
Chou et al.	HMM	Algoritmo de Viterbi	420 espécies	78,20%**
Lee et al.	TDMFCC e DTDMFCC	GMM e VQ	28 espécies	84,06%**
Briggs et al.	MFCC, MFB e Spectrum Density	K-NN, SVM, Distância de KULLBAK-LEIBLER e Distância de HELLINGER	6 espécies	91,10%**
Chou and Liu	MFCC	Redes neurais	420 espécies	18,72%**
Lopes et al.	Características acústicas	Naive Bayes, KNN (k=3), J4.8, MLP, SMO(Polynomial) e SMO(Pearson)	Xeno-Canto 3 espécies 5 espécies 8 espécies 12 espécies 20 espécies	 95,10%* 89,30%* 89,30%* 82,90%* 78,20%*
Lucio and Costa	Características visuais: LBP, LPQ, e Filtros de Gabor	SVM	Xeno-Canto 46 espécies	 77,65%*

* Resultados apresentados com o uso de F-measure
** Resultado apresentado com o uso de acurácia

Cepstral Coefficients (TDMFCC), *Dynamic Two-dimensional Mel-frequency Cepstral Coefficients* (DTDMFCC) dos cantos das especies. A base utilizada no trabalho é composta de 28 espécies e a melhor taxa de acerto foi de 84,06%.

No trabalho apresentado por Tyagi et al. [24] foi apresentada uma nova forma de representar as sílabas dos cantos dos pássaros que tem como base a média do espectro do som sobre o tempo e a classificação é baseada na combinação de padrões. A taxa de acerto foi de 100% sobre uma base de dados constituída por 15 espécies.

Vilches et al. [17] apresentaram uma abordagem de classificação de espécies de pássaros baseadas em características acústicas, tais como: harmônia, timbre e ritmo. A base de dados utilizadas neste trabalho é composta por três espécies de pássaros e a melhor taxa de acerto obtida foi de 98,39%.

Fagerlund [10] também utilizou o MFCC das amostra de áudio como características para o sistema de classificação, no entanto assim como Briggs et al. [21] fez o uso de SVM para a classificação das amostras de áudio das duas bases de dados utilizados no projeto uma contendo 6 espécies e a outra 8 espécies, sendo que os melhores resultados alcançados foram de 93,00% para a base composta por 6 espécies e de 97,00% para a base de dados composta por 8 espécies.

Lopes et al. [23], fez o uso de características acústicas das amostras de áudios das espécies de pássaros, as bases de dados utilizadas pelos autores, são compostas por um subconjunto das amostras de áudio disponibilizadas pelo site Xeno-Canto. O mesmo trabalhou 5 bases de dados compostas por 3, 5, 8, 12 e 20 espécies. Sendo que os resultados apresentados foram, 95,10%, 89,30%, 89,30%, 82,90% e 78,20% , para as respectivas bases de dados.

Lucio and Costa [11] apresentaram a classificação de espécies de pássaros utilizando espectrogramas gerados a partir dos sons disponibilizados pelo Xeno-Canto. Os descritores de textura LBP, LPQ e Filtros de Gabor foram utilizados para extrair características a partir de espectrogramas gerados a partir de uma base composta por 46 espécies divididas em 10 *folds*. A melhor acurácia encontrada foi de 77,65% usando o SVM para realizar a classificação. Porém, todas os sinais de áudio utilizados foram segmentados manualmente a fim de encontrar as regiões de interesse com cantos de pássaros e descartar ruídos externos.

III. Fundamentação Teórica

O problema de classificação, pode ser descrito como o processo pelo qual padrões ou sinais recebidos são distribuídos por um número prescrito de classes com o uso de alguma técnica de aprendizagem. A classificação representa um amplo conjunto de problemas de grande significado prático[25]. Em sua forma mais comum, um sistema de classificação é dividida em etapas bem definidas, sendo as principais: pré-processamento, extração de características e classificação [25, 26, 27], conforme ilustrado na Figura 2.

A etapa de pré-processamento, na tarefa de classificação, compreende a aplicação de várias técnicas para captação, organização, tratamento e a preparação dos dados. É uma etapa

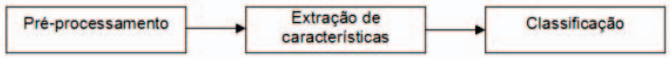


Figura 2: Principais etapas para o desenvolvimento de um sistema de reconhecimento de padrões

que possui fundamental relevância no processo. Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados. A etapa de extração de características depende fundamentalmente do tipo de sinal que está sendo processado, onde técnicas específicas para cada tipo de sinal são aplicadas sobre a base de dados. Na última etapa, algoritmos de classificação bastante conhecidos são utilizados sobre os descritores extraídos a fim de se atribuir uma classe para cada padrão submetido ao sistema[25].

As próximas seções apresentam uma breve descrição dos conceitos de extração de características, zoneamento da imagem, classificação, fusão de classificadores utilizados no desenvolvimento deste trabalho.

A. Extração de Características

A extração de características é uma etapa de grande importância para o desenvolvimento de um sistema de reconhecimento de padrões. E decorrente do objetivo deste trabalho estar relacionado a extração de características a partir do sinal do áudio dos sons de pássaros sejam estas acústicas ou visuais, as seções seguintes apresentam algumas das abordagens apresentadas na literatura para a extração de características que podem ser utilizadas para a criação de um sistema de classificação.

1) Características Acústicas: Esta seção tem por finalidade apresentar uma descrição sobre os conjuntos de características acústicas utilizados neste trabalho.

Rhythm Pattern (RP): de acordo com Rauber and Frühwirth [28] e Rauber et al. [29] o *Rhythm Patterns* (RP) descreve a amplitude da modulação para um intervalo de frequências de modulação presentes nas zonas críticas do sistema de audição humano. Os parágrafos seguintes descrevem o sistema de extração de características adotado pelo RP.

Em um primeiro momento, espectrogramas de segmentos de áudio de 44 kHz com duração de aproximadamente 6 segundos é processado utilizando a *Short-Time Fourier Transform* com uma janela de Hanning de 1024 amostras e um *overlap* de 50%.

Em seguida a escala de Bark, uma escala contínua com grupos de frequência para zonas críticas para a audição humana, é aplicado ao espectrograma, agregando a este 24 zonas de frequência[30].

Posteriormente o espectrograma gerado pela escala de Bark é transformado na escala de Decibel, para posteriormente se aplicar as transformações psicoacústicas: É realizado então o cálculo da escala de Phon para incorporar curvas de ruído que são utilizadas para calcular diferentes percepções de ruído de diferentes frequências, e transformações na escala de Sone para calcular ruídos. O sonograma resultante da escala de Bark especifica a sensação de ruído de um seguimento de áudio para a audição humana [30].

Em um segundo momento, a variação da energia nas zonas críticas do espectrograma na escala de Bark é considerada como uma modulação da amplitude do sinal sobre o tempo também conhecida como *cepstrum* é obtida através do uso da transformada de Fourier. O resultado é um sinal da invariante no tempo que contém a magnitude de modulação para frequência nas zonas críticas. Esta matriz representa um RP, indicando a ocorrência de ritmo como barras verticais, mas também descrevendo pequenas flutuações em todas as zonas de frequência do sistema de audição humano.

Subsequentemente, modulações de amplitude são avaliadas de acordo com a função da sensação humana através da modulação da frequência acentuando valores em torno de 4 Hz, e eliminando frequência maiores que 10 Hz. A aplicação de um filtro gradiente e da suavização Gaussiana melhora a similaridade entre RPs. A matriz final de características com dimenões de 24×60 é computada pela mediana dos RP segmentados.

Rhythm Histogram (RH): agrega valores de modulação de amplitude de 24 zonas críticas individuais computadas por um *rhythm pattern*, apresentando a magnetude da modulação para 60 frequências de modulação entre $0,17$ e 10Hz [31].

O que acaba por caracterizar um descritor geral para características rítmicas em uma amostra de áudio. Um RH é computado para cada segmento de 6 segundos em uma amostra de áudio e o vetor de características é então calculado pela mediana dos é valores cálculados para cara um dos segmentos.

Statistical Spectrum Descriptor (SSD): o *Statistical Spectrum Descriptor* (SSD) é um descritor de características acústicas que tem por finalidade computar específicas sensações de ruído nas 24 zonas da escala de Bark, análogamente ao *Rhythm Patterns*. Subsequentemente, medidas estatísticas são calculadas para cada uma das zonas críticas, descrevendo assim variações em cada uma das zonas estatísticamente. O SSD assim descreve flutuações nas zonas críticas e captura informação de timbre adicional que não foi coberto por outros conjuntos de características, tal como *Rhythm Pattern*, desta forma capturando e descrevendo muito bem conteúdo acústico [31].

2) Características Visuais: Esta seção tem por finalidade apresentar uma descrição sobre os conjuntos de características visuais utilizados neste trabalho.

Local Binary Pattern (LBP): o método define que a textura de uma imagem é descrita levando-se em consideração cada um pixel central C, com seus P vizinhos equidistantes considerando uma distância R, com pode ser visto na Figura 3. O histograma h de padrões LBP é sintetizado utilizando-se as diferenças de intensidade entre cada pixel central C e seus P vizinhos. De acordo com [32], boa parte da informação sobre características de textura é preservada na distribuição T descrita na Equação 1.

$$T \approx (g_0 - g_c, ..., g_{P-1} - g_c) \tag{1}$$

onde g_c é a intensidade de nível de cinza do pixel central C e g_0 a g_{p-1} correspondem as intensidades de nível de cinza dos vizinhos. Quando um vizinho não corresponde exatamente à posição de um pixel, seu valor é obtido por interpolação.

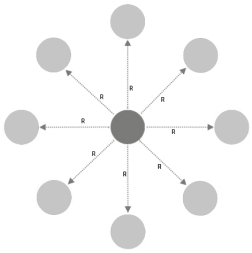
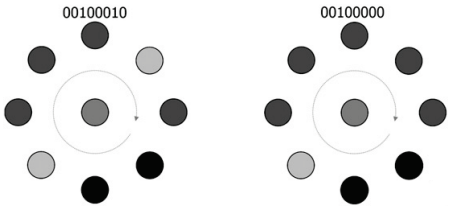


Figura 3: O operador LBP. O *pixel C*, escuro no centro, e os *pixels* claros são os *P* vizinhos



(a) Representação de um padrão não uniforme

(b) Representação de um padrão uniforme

Figura 4: Uniformidade do padrão LBP

Considerando o sinal resultante da diferença entre o pixel central C e cada um dos seus P vizinhos, como descrito na Equação 2, é definido que: se o sinal é positivo, o resultado é igual a um; caso contrário, o resultado é igual a zero, como descrito na Equação 3.

$$T \approx (s(g_0 - g_c), ..., s(g_{P-1} - g_c) \tag{2}$$

na qual

$$s(g_i - g_c) = \begin{cases} 1 & \text{se } g_i - g_c \geq 0 \\ 0 & \text{se } g_i - g_c \leq 0 \end{cases} \tag{3}$$

na qual $i = [0, P]$ é o índice dos vizinhos de C.
Com isto, o valor do padrão LBP referente ao pixel C pode ser obtido através da multiplicação dos elementos binários por um coeficiente binomial. Associando-se um peso binomial 2^P a cada $s(g_p - g_c)$, as diferenças presentes na vizinhança são transformadas em um único código LBP, um valor $0 \leq C \leq 2^P$. A Equação 4 descreve como este código é obtido.

$$LBP_{P,R}(X_C, Y_C) = \sum_{P=0}^{P-1} s(g_P - g_C)2^P \tag{4}$$

assumindo que $X_C \in \{0, ..., No - 1\}$

O conceito de uniformidade da sequência obtida no padrão LBP, é baseado no número de transições entre zeros e uns presente na sequência associada ao padrão[32]. Um código LBP binário e considerado uniforme se o número de transições é menor ou igual a dois, considerando inclusive que o código é tratado como uma lista circular. Assim, o código representado pela sequência 00100100 não é considerado uniforme, já que contém quatro transições. Por outro lado, o código 00100000 é considerado uniforme, já que apresenta apenas duas transições, como pode ser visto na Figura 4.

Assim ao invés de utilizar integralmente o histograma de padrões LBP, cujo tamanho é 2^P, é possível utilizar apenas os valores associados a padrões uniformes, constituindo um vetor com menor dimensionalidade, com apenas 59 características. De acordo com [32], além das 58 combinações uniformes, todos os padrões não uniformes encontrados devem participar de uma coluna adicional no histograma gerado. Devido a este fato, o vetor de características LBP para na configuração de 8 vizinhos com distância 2 possui 59 características em sua constituição.

Local Phase Quantization (LPQ): é uma técnica para descrição de textura apresentada por Ojansivu, Ville and Heikkilä [33] originalmente utilizadas em imagens que apresentam borramento, todavia, é interessante observar que, embora o método tenha sido criado com este propósito, ele também produz resultados muito bons para imagens que não apresentam borramento.

O descritor, denominado *Local Phase Quantization* (LPQ) é baseado na propriedade de invariância ao borramento do espectro de fase de Fourier. Ele utiliza a informação de fase local extraída utilizando a 2D DFT computada sobre uma vizinhança retangular, chamada janela local, para cada pixel da imagem. A informação da fase local de uma imagem de tamanho $N \times N$ é dada pela *Short-time Fourier Transform* (STFT) descrita na equação 5.

$$\hat{f}_{u_i}(x) = (f \times \phi_{u_i})x \tag{5}$$

sendo o filtro ϕ_{u_i} dado pela equação

$$\phi_{u_i} = e^{-j2\pi u_i^T} | y \in \mathbb{Z}^2 || y || \infty \leq r \tag{6}$$

na qual $r = (m - 1)/2$ é do tamanho da janela local e u_i é um vetor de frequências 2D.

No LPQ são considerados apenas quatro coeficientes complexos que correspondem às frequências 2D: $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, $u_1 = [a, -a]^T$, em que $a = 1/m$. Por conveniência, a STFT(equação)5 é expressa através do vetor de notação conforme a equação 7

$$\hat{f}_{u_i}(x) = w_{u_i}^T f(x) \tag{7}$$

sendo $F = [f(x_1), f(x_2), \cdots, f(x_{x^2})]$ denotado como uma matriz $m^2 \times N^2$ que compreende a vizinhança de todos os pixels da imagem e $w = [w_R, w_I]$, em que $w_R = Re[W_u1, W_u2, W_u3, W_u4]$ e $w_I = Im[W_u1, W_u2, W_u3, W_u4]$. O $Re[]$ e $Im[]$ representam, respectivamente, as partes reais e imaginárias de um número complexo e a matriz de transformação $(8 \times N^2)$ é dada por $\hat{F} = wF$.

[33] assumem que a função $f(x)$ de uma imagem é resultado de um processo de primeira ordem de Markov, em que o coeficiente de correlação entre dois pixels x_i e x_j é relacionada exponencialmente com a sua distância L^2. Para o vetor f é definida uma matriz de covariância C de tamanho $m^2 \times m^2$, dada pela equação 8. A matriz de covariância dos coeficientes de Fourier pode ser obtida por $DwCw^T$. Considerando que D não é uma matriz diagonal, os coeficientes são correlatos e podem deixar de ser através de $E = C^T\hat{F}$, sendo V uma matriz ortogonal derivada do valor de decomposição singular (SVD - *Singular Value Decomposition*) da matriz D, com $D' = V^T DV$.

$$C_{i,j} = \sigma^{||x_i - x_j||} \tag{8}$$

Os coeficientes são quantizados usando-se a equação 9, em que e_{ij} são componentes de E Estes elementos são transformados de binário para decimal através da equação 10 e caracterizam valores inteiros compreendidos entre 0 e 255. Então, através de todas as posições da Imagem, é composto

o vetor de 256 posições que correspondem ao histograma do LPQ.

$$q_{ij} = \begin{cases} 1 & \text{se } e_{ij} \geq 0 \\ 0 & \text{caso contrário} \end{cases} \tag{9}$$

$$b_j = \sum_{i=0}^{7} q_{ij} 2^i \tag{10}$$

Filtros de Gabor: durante muito tempo um sinal podia ser representado em função do tempo ou, alternativamente, em função da frequência através da transformada de Fourier. Entretanto esta abordagem possuía a limitação de permitir a extração de informações apenas no domínio da frequência e não em função do tempo. Em 1946, Denis Gabor apresentou os filtros de Gabor, que permitem extrair informações no domínio da frequência e tempo. Em seu trabalho original Gabor buscava a síntese do sinal e preocupou-se em como um sinal poderia ser construído através da combinação linear de funções lineares [34]. Os filtros de Gabor correspondem à um conjunto de funções senoidais complexas, bidimensionais, moduladas por uma função Gausiana também bidimensional com propriedades muito úteis para a finalidade de classificação de imagens. Na análise de sinais em processamento de imagens, a extração de características exerce um papel importante no qual o principal objetivo é saber "o que está aonde" . com os princípios de Gabor, informações relacionadas a frequência pode informar "o que" , enquanto as ligadas ao tempo podem informar "aonde".

A segmentação da textura é uma tarefa difícil e muito importante em muitas aplicações de análise de imagens ou visão computacional e filtros de Gabor têm sido utilizados com êxito para estes propósitos. Existem muitas formas de se implementar filtros de Gabor apresentadas na literatura. Uma possível forma para filtros de Gabor bidimensionais no domínio espacial, portanto apropriados para imagens digitais, é dada pelas equações 11 e 11.

$$\Psi(x,y) = exp\left(-\left(\frac{x^2+Y^2}{2\sigma^2}\right)\right) exp\left(\frac{j2\pi x}{\lambda}\right) \tag{11}$$

na qual j é a unidade imaginária, , σ é o desvio padrão da função Gausiana e λ é o comprimento de onda.

Para uma imagem I de tamanho $M \times N$, e considerando $\Psi(x,y)$ conforme descrito na equação 11, a saída do filtro de Gabor é obtida pela convolução da imagem de entrada com o filtro de Gabor apresentado na equação 12.

$$\sum_x \sum_y I(m-x, n-y)\Psi(x,y) \tag{12}$$

Filtros de Gabor podem ser utilizados para detectar linhas. Uma vez que a imagem pode conter linhas com diferentes espessuras, é necessário construir filtros de Gabor com diferentes fatores de escala, variando λ. Adicionalmente, o filtro de Gabor pode detectar somente linhas verticais, o que não é suficiente em muitos casos, já que é comum a ocorrência de linhas com diferentes orientações nas imagens. Assim, pode-se rotacionar $\Psi(x,y)$ com um ângulo θ para construir $\Psi(x',y')$ para a detecção de linhas com diferentes orientações. Neste caso, x' e y' podem ser encontrados pelas equações 13 e 14 respectivamente.

$$x' = x\cos\theta + y\sin\theta \tag{13}$$

$$y' = x\sin\theta + y\cos\theta \tag{14}$$

Robust Local Binary Pattern: idealizado com a finalidade de suprir a deficiência apresentada pelo LBP quando de trata de extrair características de imagens que apresentam alta taxa de ruídos [35].

Assim como o LBP, consiste em analisar um determinado número de pixels vizinhos P, baseado em um pixel central C levando-se em consideração uma distância R.

O que difere uma abordagem da outra é a forma como a uniformidade da sequência do padrão é obtida. Como foi citado anteriormente um código LBP binário é considerado válido se e somente se o número de transicões é menor ou igual a dois, tomemos então como exemplo a sequência 00100100, ao analisarmos o número de transições entre os valores de 0 e 1 podemos ver claramente que este valor é igual a 4 sendo assim é plausível dizer que temos então uma sequência não uniforme.

O RBLP tem por finalidade analisar a representação binária de um pixel central para com seus vizinhos, buscando identificar *substrings* que seguem os padrões 101 e 010 e assim substitui-ls por 111 e 000 respectivamente, e assim porteriormente realizar a análise de uniformidade.

Para enteder melhor a idéia da substituição de *substrings* proposta tomemos novamente como exemplo a representação binária 00100100, sendo que após aplicar a ídeia apresentada obtem-se a seguinte representação 00000000 e ao se realizar a análise de uniformidade contasta-se que se trata de um padrão uniforme.

Gray-Level Co-Occurence Matrix: é uma abordagem apresentada no trabalho [36], como o próprio nome sugere a mesma faz o uso de matrizes de co-ocorrência para a caracterização da textura de uma imagem através de medidas estatísticas obtidas a partir da contagem de ocorrências dos níveis de cinza presentes nos pixels da imagem ou a obtidas através da forma com pixels de diferentes níveis de cinza se relacionam no espaço bidimensional de uma imagem.

A metodologia da extração das características consiste em se construir as matrizes de co-ocorrência para sem seguida realizar a extração das medidas estatísticas inerentes as mesmas. As matrizes construídas sobre a imagem são da ordem $N \times N$, onde N corresponde ao número de tons de cinza utilizados na representação da imagem. Sendo que para cada posição da matriz é armazenada a probabilidade de que dois valores de intensidades de cinza estejam envolvidos por uma determinada relação espacial.

Em cada posição da matriz é armazenada a probabilidade de que dois valores de intensidades de cinza estejam envolvidos por uma determinada relação espacial. Parâmetros como a distância d entre os pixels e o ângulo θ que caracteriza a orientação de uma reta que passa pelos mesmos definem uma relação espacial. As possíveis orientações do ângulo θ são 0 °, 45 °, 90 °, 135 °, como pode ser observado na Figura 5.

A Figura 6 apresenta a matriz de pixels de uma imagem com itensidades de cinza variando entre 0 e 3, por meio da mesma iremos apresentar um exemplo da geração da matriz de co-ocorrência.

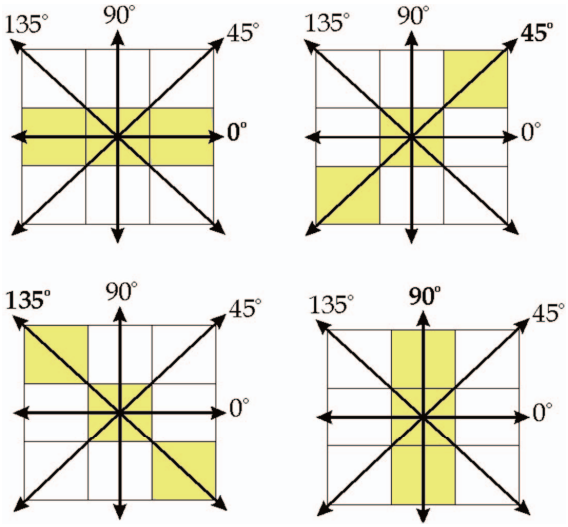


Figura 5: Orientações utilizadas para criação da matriz de co-ocorrência

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Figura 6: Exemplo de matriz de pixels de uma imagem

A partir da matriz de pixels apresentada na Figura 6, foi considerada a orientação $\theta = 0°$ e distância $d = 1$, para criar a matriz de co-ocorrência. Sendo que de acordo com o método proposto por no trabalho [36], a matriz de co-ocorrência registra nas posições (i, j) o número de ocorrências de relação espacial entre um pixel de intensidade i e um pixel com intensidade j considerando a distândia d e a orientação θ independente do sentido da relação. Assim, a presença de um pixel de intensidade j imediatamente a direita de um pixel de intensidade i seria contabilizada na matriz com $d = 1$ e $\theta = 0°$ da mesma forma como a ocorrência da intensidade j imediatamente à esquerda de i seria contabilizada. com isso, a matriz de co-ocorrência que se forma é simétrica. Depois de contadas as quantidades das relações espaciais, elas são transformadas em probabilidades para a relação dos processos de extração de características subsequentes, conforme mostra a Figura 7.

	0	1	3	3
0	4	2	1	0
1	2	4	0	0
2	1	0	6	1
3	0	0	1	2

→

	0	1	2	3
0	0,25	0,12	0,04	0
1	0,12	0,25	0	0
2	0,06	0	0,37	0,06
3	0	0	0,06	0,12

Figura 7: Matrix de co-ocorrência de distância um e ângulo zero

No trabalho [36] foram propostas originalmente 14 medidas de características de texturas possíveis de se extrair das matrizes de co-ocorrência. Essas características são calculadas a partir de algumas equações que utilizam as probabilidades associadas as posições da matriz de co-ocorrência.

Das 14 características originalmente propostas, sete se consolidaram como características relevantes em processos de descrição de textura. Essas características são: contraste, energia (ou uniformidade), entropia, homogeneidade, momento de terceira ordem, probabilidade máxima e correlação. Sendo G o número de intensidade de cinza utilizado na representação da imagem e $p(i, j)$ a probabilidade de relacionamento entre as intensidades i e j, as equações de 15 a 21 representam as características citadas anteriormente.

$$\text{Contraste} = \sum_{i=1}^{G}\sum_{j=1}^{G}(i-j)^(2)(p(i,j)) \tag{15}$$

$$\text{Energia} = \sum_{i=1}^{G}\sum_{j=1}^{G}((p(i,j))^(2) \tag{16}$$

$$\text{Entropia} = \sum_{i=1}^{G}\sum_{j=1}^{G}p(i,j)\ \log\ p(i,j) \tag{17}$$

$$\text{Homogeneidade} = \sum_{i=1}^{G}\sum_{j=1}^{G}\left(\frac{p(i,j)}{1+(i-j)^(2)}\right) \tag{18}$$

$$\text{Momento de terceira ordem} = \sum_{i=1}^{G}\sum_{j=1}^{G}p(i,j)(i-j)^(3) \tag{19}$$

$$\text{Probabilidade máxima} = \sum_{i=1}^{G}\sum_{j=1}^{G}\ \max(p(i,j)) \tag{20}$$

$$\text{Correlação} = \left(\frac{p(i,j)-\mu_x\mu_y}{\sigma_x^2\sigma_y^2}\right) \tag{21}$$

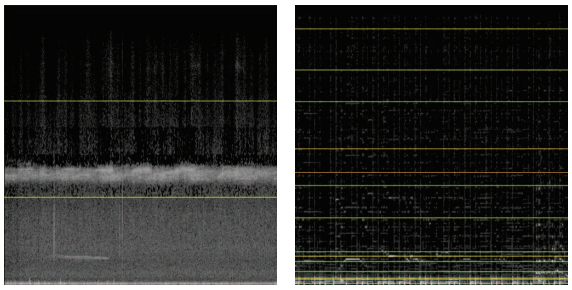
na qual $\mu_x = \sum_{i=1}^{G} i \times p_x(i)$, $p_x(i) = \sum_{j=1}^{G} p(i,j)$, $\sigma_x^2 = \sum_{i=1}^{G}(i-\mu_x)^2 p_x(i)$, $\mu_y = \sum_{j=1}^{G} j \times p_y(j)$, $p_y(j) = \sum_{i=1}^{G} p(i,j)$ e $\sigma_y^2 = \sum_{j=1}^{G}(j-\mu_y)^2 p_y(j)$.

B. Divisão da Imagem em Zonas

O zoneamento da imagem tem como objetivo preservar as informações locais presentes em regiões específicas da imagem [37]. Além da preservação de informações locais, a estratégia de divisão das zonas é bastante oportuna por permitir naturalmente a criação de um *pool* de classificadores, pois um classificador é criado para cada uma das zonas. De acordo com a literatura existem alguns sistemas de zoneamento que diferem entre si, podendo estes serem lineares ou não [37]. As próximas subseções, descrevem algumas metodologias de zoneamento existentes.

Divisão em Zonas Lineares: com a divisão linear, são estabelecidas na imagem do espectrograma zonas de igual tamanho que correspondem a bandas de frequência. O limites de cada banda criada dependem da quantidade de zonas definidas e do limite de frequência até o qual o sinal das músicas utilizadas apresenta informação relevante. A Figura 8a apresenta um espectrograma dividido linearmente em 3 zonas

Divisão pela escala de Mel: é uma escala psicoacústica apresentada por S. S. Stevens [38], esta diretamente relacionada as frequências percebidas pelos humanos, assim como a escala Bark, no entanto são estabelecidas 15 bandas de frequência, cujos limites em Hz são: 0, 40, 161, 200, 404,



(a) Exemplo de zoneamento utilizando escala linear (b) Exemplo de zoneamento utilizando escala de Mel

Figura 8: Exemplos de zoneamento do espectrograma

693, 867, 1000, 2022, 3393, 4109, 5526, 6500, 7743 e 12000. Nas aplicações que envolvem a classificação a partir de imagens de espectrograma, o número de zonas criadas, e consequentemente o número de classificadores, depende do limite de frequência até o qual a imagem do espectrograma apresenta informações relevantes. A Figura 8b apresenta um espectrograma dividido entre os limites de frequência supracitados, sendo que a primeira linha na parte inferior da imagem representa o limite de frequência mais baixo e a última linha na parte superior da imagem representa o maior limite de frequência.

C. Combinação de Classificadores

Muitos esquemas de classificação são criados tendo por base a utilização de um único classificador para resolver um determinado problema. No entanto, a qualidade das hipóteses induzidas por estes depende da quantidade dos exemplos dos conjuntos de treinamento. Por outro lado, muitos dos sistema de aprendizado de máquina conhecidos não estão preparados para trabalhar com uma grande quantidade de exemplos. Uma maneira para resolver este problema consiste em realizar a combinação de classificadores, podendo proporcionar um melhor resultado se comparado aos resultados individuais de cada classificador [39].

O que fundamenta o ganho de desempenho ao se realizar a combinação de classificadores é que, os conjuntos de padrões classificados incorretamente por diferentes classificadores não necessariamente se sobrepõem. Isto sugere que diferentes projetos de classificadores podem oferecer informação complementar sobre os padrões a serem classificados, fato este, que pode acarretar na melhora do desempenho do sistema de classificação [40].

De acordo com Kittler et al. [40], o método mais comumente utilizado para a combinação de classificadores é dada por meio das seguintes funções matemáticas: máximo, mínimo, produto, soma e média. As equações 22 a 26 descrevem as fórmulas utilizadas das funções supra citadas.

$$Regra\,do\,Maximo(v) = \arg\max_{k=1}^{c} \max_{i=1}^{n} P(\omega_k|l_i(v)) \tag{22}$$

$$Regra\,do\,Minimo(v) = \arg\max_{k=1}^{c} \min_{i=1}^{n} P(\omega_k|l_i(v)) \tag{23}$$

$$Regra\,do\,produto(v) = \arg\max_{k=1}^{c} \prod_{i=1}^{n} P(\omega_k|l_i(v)) \tag{24}$$

$$Regra\,da\,Soma(v) = \arg\max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k|l_i(v)) \tag{25}$$

$$Regra\,da\,Média(v) = \frac{1}{n} \arg\max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k|l_i(v)) \tag{26}$$

na qual, v representa o padrão que será classificado, n é o número de classificadores, l_i representa a saída do i-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_1, ..., \omega_c$ e $P(\omega_k|l_i(v)$ é a probabilidade de que a amostra v pertença a classe ω_k encontrada pelo i-ésimo classificador.

D. Medidas de Avaliação

Esta seção apresenta os critérios comumente utilizados para avaliar a eficiência de sistemas de classificação, sendo estes: *precison*, *recall*, *F-measure* e *Macro-F*. As subseções seguintes apresentam os critérios de avaliação citados [41] .

1) Precision: É o total de exemplos corretamente classificados como uma classe C sobre o total de exemplos classificados como a classe C, a desvantagem desta métrica é que ela não leva em consideração os exemplos de deveriam ter sido reprovados, mas foram aprovados. Sua fórmula é expressa pela equação 27.

$$Precision(C_i) = \frac{M(C_i, C_i)}{M(*, C_i)} \tag{27}$$

2) Recall: É o total de exemplos corretamente classificados como uma classe C sobre o total de exemplos pertencentes a classe C presentes co conjunto de dados, a desvantagem desta métrica é que ela não leva em consideração todas as medidas. Sua fórmula é expressa pela equação 28.

$$Recall(C_i) = \frac{M(C_i, C_i)}{M(C_i, *)} \tag{28}$$

3) F-measure: É a média harmonica das medidas de Precison e Recall, sendo uma forma de expressar as duas medidas com um único valor, sua fórmula é expressa pela equação 29.

$$F - measure(C) = \frac{2 \times recall(C) \times precision(C)}{recall(C) + precision(C)} \tag{29}$$

em que C é a classe sobre a qual o valor está sendo calculado.

4) Macro-F: É a média aritmética das F-measures de todas as classes presentes no conjunto de dados, sua fórmula é expressa pela equação 30.

$$Macro - F(h) = \frac{1}{k} \sum_{i=1}^{k} F - measure(C_i) \tag{30}$$

em que k é o total de classes presentes no conjunto de amostras.

IV. Base de Dados dos Cantos

A base de dados empregada neste trabalho foi a mesma utilizada no trabalho apresentada por Lucio and Costa [11], sendo que a mesma é composta por 2814 amostras de áudio divididas entre 46 espécies.

V. Método Proposto

Esta seção apresenta o método de classificação utilizado no desenvolvimento do trabalho até o presente momento. No que diz respeito especificamente ao método proposto neste trabalho, foram realizadas as seguintes etapas para realizar a tarefa de classificação: geração dos espectrogramas, extração das características, treinamento utilizando SVM para a criação de modelos de classificação com otimização de parâmetros do classificador. Cada uma das etapas apresentadas anteriormente podem ser vistas nas seções seguintes.

A. Geração do Espectrograma

Para a geração dos espectrogramas foi utilizado o sofware Sox 14.4.1 (*Sound eXchange*), um utilitário disponível em http://sox.sourceforge.net que permite a realização de conversões entre vários formatos diferentes de representação de áudio. Este permite a utilização de alguns parâmetros que irão impactar na aparência do espectrograma gerado, por meio deste pode-se delimitar a altura e a largura, e a amplitude do sinal de áudio a ser considerada. Alguns parâmetros deste software foram empiricamente ajustados a fim de produzir espectrogramas com conteúdo de textura destacado.

B. Divisão da Imagem em Zonas

Para o zoneamento da imagem foi adotada a estratégia apresentada na seção III-B, na qual foram realizados testes utilizando zoneamento linear onde foi realizada a variação do número de zonas com o propósito de se verificar que quantidade de zonas apresenta uma taxa de acerto mais elevada e e não linear fazendo-se o uso da escala de Mel. Os testes realizados com a escala de Mel foram realizados removendo as zonas inferiores da mesma decorrente do fato de não haver uma quantidade de pixels suficiente para se extrair as características utilizadas no sistema de classificação sendo assim das 15 bandas de frequência presentes na escala Mel apenas 13 foram empregadas.

C. Extração de Características

1) Características Acústicas: Esta seção apresenta as configurações utilizadas pelos descritores de características acústicas utilizados neste trabalho.

Esta seção tem por finalidade apresentar uma descrição sobre os conjuntos de características acústicas utilizadas neste trabalho. A mesma encontra-se dividida em subseções nas quais são apresentadas as configurações utilizadas sobre os descritores para gerar os vetores de características.

SSD: os vetores de características gerados a partir do SSD possuiam 161 elementos, estes foram obtidos através da biblioteca *Rhythm and Timbre Feature Extraction from Music* [1].

RH: os vetores de características gerados a partir do RH possuiam 60 elementos, e assim como o SSD estes foram obtidos através da biblioteca *Rhythm and Timbre Feature Extraction from Music*.

[1]http://ifs.tuwien.ac.at/mir/musicbricks/index.html, acessado em 05/05/2016

RP: os vetores de características gerados a partir do RP possuiam 1380 elementos, e assim como os outros dois descritores citados anteriormente também foram gerados através da biblioteca *Rhythm and Timbre Feature Extraction from Music* .

2) Características Visuais: Esta seção apresenta as configurações utilizadas pelos descritores de características visuais utilizados neste trabalho.

LBP: a geração dos vetores de características extraídos através do uso do LBP levou em consideração o $LBP_{8,2}$, ou seja este leva em consideração 8 pixels com uma distância 2 pixels do pixel central, fato este que proporcionou a criação de vetores de características compostos por 59 elementos.

LPQ: para a criação dos vetorores de características a partir do LPQ se fez o uso de uma janela de dimensão 5×5 e ao se percorrer todos os pixels da imagem com está foi obtido um vetor composto por 256 elementos.

RLBP: a criação dos vetores de características com o RLBP assim como o LBP levou em consideração o $RLBP_{8,2}$, ou seja este levou em consideração 8 pixels com uma distância 2 pixels do pixel central, fato este que proporcionou a criação de vetores de características compostos por 59 elementos.

Filtro de Gabor: os vetores de características obtidos pelo filtro de gabor tem sua dimensão dependente do número de fatores de escala, da quantidade de rotações utilizadas e do número de medidas estatísticas escolhidas, sendo que o total de elementos presentes em um vetor é dado pela sequinte relação *Número de rotações* $\times$ *Número de fatores de escala* $\times$ *Número de rotações*.

GLCM: os vetores de características utilizado com uso do GLCM eram compostos por 28 elementos, calculados a partir das 4 orientações possíveis sendo estas $0°$, $45°$, $90°$ e $135°$.

D. Sistema de Classificação

Para realizar as tarefas de classificação será utilizado o SVM, um modelo de algoritmos de aprendizagem apresentado por Vapnik [46], seu uso é muito difundido em toda comunidade científica em trabalhos que envolvem a análise de dados e reconhecimento de padrões, sendo que este tem apresentado bom desempenho em vários trabalhos publicados recentemente. Para a construção dos modelos de classificação SVM, foi utilizado o *kernel Radial Basis Function* (RBF) e os parâmetros C e γ foram otimizados utilizando um procedimento *grid-search*.

O esquema de classificação proposto consiste na seguinte sequência de passos: divisão da base de dados em folds, geração dos espectrogramas, extração das características. Após extraídas as características foi realizada a classificação por meio do SVM através da biblioteca LIBSVM [47]. A técnica consiste na utilização de dois conjuntos de dados, sendo um para treino e outro para teste. Com o objetivo de obter um resultado mais consistente foi utilizada a técnica de validação cruzada, na qual um dos folds criados é utilizado como conjunto de teste e os demais para treinamento, sendo que o processo é repetido até que todos os folds criados tenham sido utilizados como conjunto de teste [40].

Ao final, toma-se como medida de desempenho a taxa de acerto média obtida entre todas as situações testadas.

VI. Resultados e Discussão

A Tabela II apresenta a síntese dos melhores resultados encontrados em todos os testes realizados para a escrita deste trabalho.

Analisando a Tabela II é possível observar, alguns aspectos envolvendo cada uma das etapas de testes realizadas. Na primeira etapa quando o espectrograma foi analisado em sua completude o melhor resultado foi dado através do uso do Filtro de Gabor como descritor de textura, as medidas de avaliação alcançadas com a utilização do mesmo foram: *Recall* 76,44%, *Precision* 80,80% e *F-Measure* 79,09%.

Ao utilizarmos o zoneamento linear dos espectrogramas obtemos resultados melhores dos que os que foram encontrados sem a utilização do zoneamento da imagem para 4 descritores de textura, sendo estes, LBP, LPQ, RLBP e Filtro de Gabor. Todavia quando se fez o uso do zoneamento do espectrograma da imagem pela escala Mel todos os resultados foram inferiores ao teste em que não foi utilizado nenhuma técnica de zoamento. O melhor resultado encontrado ao utilizar o zoneamento da imagem foi dado pelo descritor RLBP dividido em 4 zonas lineares, obtendo as seguintes medidas de avaliação: *Recall* 79,01%, *Precision* 87,42% e *F-Measure* 81,87%.

Quanto aos testes em que foram utilizados descritores de características acústicas, podemos constatar que os melhores resultados foram encontrados ao se utilizar o SSD, que trabalha diretamente sobre o timbre das amostras de áudio, fato este que nos leva a crer que os descritores que trabalham com timbre de amostras de áudio apresentam grande relevância para tarefas que envolvem a classificação automática de espécies de pássaros. O resultados apresentado pelo SSD apresentou taxas de acerto superiores as encontradas nos descritores de textura, sendo estas: *Recall* 79,68%, *Precision* 83,87% e *F-Measure* 81,83%.

No momento em que se combinou os resultados dos melhores classificadores baseados em características acústicas com características visuais, foi possível observar que para certas combinações de características há complementariedade entre os resultados proporcionando alcançar valores medidas de avaliação superiores a 90% nos melhores casos. Nestes experimentos, o que apresentou as medidas de avaliação dos resultados mais elevadas foi aquele em que se combinou o classificador obtido com o RLBP sem zoneamento com o classificador obtido para o SSD, sendo que os valores alcançados foram: *Recall* 91,08%, *Precision* 94,02 e *F-Measure* 92,34%.

O resultados encontrados pela combinação do RLBP com o SSD não somente apresentou o melhor resultado dos testes envolvendo zoneamento como também apresentou o melhor resultado geral do trabalho, o que caracteriza que o RLBP combinado com o SSD é uma boa metodologia para a classificação automática de espécies de pássaros.

Muito embora, os resultados aqui empresentados apresentem taxas de acerto semelhantes as encontradas durante a revisão da literatura, não foi possível realizar uma comparação justa com a maior parte dos trabalhos, decorrente do fato de as bases de dados empregadas serem diferentes, visto que não se obteve acesso as bases utilizadas por outros autores. Todavia podemos realizar uma comparação direta com o trabalho apresentado por Lucio and Costa [11], visto que a mesma base de dados é empregada em ambos os trabalhos.

No trabalho citado anteriormente a melhor taxa de acurácia encontrada foi $77,65\%$ como pode ser visto na tabela I, já no presente trabalho a melhor taxa de acurácia é de $94,02\%$, fato esse que demonstra a evolução alcançada quando passou a se utilizar outros tipos de descritores de características visuais aliados a descritores acústicos e também ao zoneamento dos espectrogramas.

VII. Considerações Finais e Trabalhos Futuros

O presente trabalho foi desenvolvida sobre a premissa da utilização de descritores de características acústicas e visuais, para a criação de um sistema de classificação automático de espécies de pássaros, tendo por finalidade explorar também a existência da complementariedade entre classificadores baseados nesses diferentes tipos de características.

Nos experimentos iniciais foi constatado que o uso do Filtro de gabor como descritor de caractrísticas visuais apresentou os melhores resultados nos testes em que nenhum tipo de zoneamento foi aplicado sobre os espectrogramas, quando comparado a alguns dos principais descritores de textura, de diferentes abordagens, descritos na literatura. Nos testes seguintes onde foram aplicadas técnicas de zoneamento sobre os espectrogramas e dentre todos os descritores de textura utilizados os melhores resultados foram alcançados com o uso do RLBP.

Quanto ao uso dos descritores de características acústicos, estes foram escolhidos por apresentarem bons resultados em trabalhos préviamente descritos na literatura, tanto relacionados à classificação de espécie de pássaros, quanto a outros domínios de aplicação. Com o uso destes foi constatado que o uso do SSD apresentou as melhores taxas de acerto. E devido a este fato o SSD foi empregado na combinação com características visuais, com a execução destes foi verificado que havia complementariedade entre os dois conjuntos de características distintos, sendo que o melhor resultado encontrado foi dado pela combinação do SSD com o RLBP que apresentou as melhores taxas de acerto do trabalho com as medidas de avaliação *Recall*: 91,08%, *Precision*: 94,02% e *F-Measure*: 92,34%.

Tendo por base as explanações apresentadas acima podemos verificar a eficácia do método de classificação de espécies de pássaros aqui proposto assim como também a complementariedade entre os descritores de características acústicas e visuais como pode ser visto na seção VI. O que caracteriza a validade da hipótese apresentada por este trabalho.

VIII. Trabalhos futuros

Para trabalhos futuros, tem-se como objetivo sintetizar novas bases de dados com um número maior de espécies, tendo

Tabela II: Resultados obtidos com o uso do LBP

Descritor	Zonas	Regra de Fusão	Recall	Precison	F-Measure	Descritor	Zonas	Regra de Fusão	Recall	Precison	F-Measure
LBP	Nenhum	-	74,70%	79,01 %	76,39%	RH	Nenhum	-	28,36%	33,47 %	28,13%
LPQ	Nenhum	-	66,65%	73,06 %	68,52%	LBP e SSD	Nenhum	-	89,02%	93,09 %	90,68%
RLBP	Nenhum	-	74,94%	80,26 %	76,80%	LBP e SSD	Linear	Soma	89,42%	88,25 %	90,81%
GABOR	Nenhum	-	76,44%	80,80 %	79,09%	LBP e SSD	Escala Mel	Soma	83,79%	88,25 %	85,59%
GLCM	Nenhum	-	45,59%	56,34 %	44,37%	LPQ e SSD	Nenhum	-	52,12%	60,87 %	53,06%
LBP	Linear	Produto	78,29%	85,51 %	80,74%	LPQ e SSD	Linear	Soma	89,28%	92,55 %	90,60%
LPQ	Linear	Soma	72,10%	81,09 %	77,95%	LPQ e SSD	Escala Mel	Máximo	80,96%	85,61 %	82,80%
RLBP	Linear	Produto	79,01%	87,42 %	81,87%	RLBP e SSD	Nenhum	-	**91,08%**	**94,02** %	**92,34%**
GABOR	Linear	Produto	76,97%	86,25 %	79,56%	RLBP e SSD	Linear	Soma	90,57%	93,62 %	90,57%
GLCM	Linear	Produto	19,13%	54,98 %	21,90%	RLBP e SSD	Escala Mel	Soma	83,64%	88,32 %	85,47%
LBP	Escala Mel	Máximo	73,09%	78,92 %	75,03%	GABOR e SSD	Nenhum	-	90,67%	93,78 %	91,96%
LPQ	Escala Mel	Mediana	2,17%	30,68 %	3,92%	GABOR e SSD	Linear	Soma	88,93%	92,51 %	90,35%
RLBP	Escala Mel	Máximo	72,32%	79,83 %	74,66%	GABOR e SSD	Escala Mel	Soma	83,26%	88,23 %	85,20%
GABOR	Escala Mel	Máximo	69,62%	78,02 %	73,18%	GLCM e SSD	Nenhum	-	80,47%	90,68 %	86,92%
GLCM	Escala Mel	Mediana	4,54%	30,68 %	3,92%	GLCM e SSD	Linear	Soma	78,99%	88,57 %	82,43%
SSD	Nenhum	-	79,68%	83,87 %	81,83%	GLCM e SSD	Escala Mel	Máximo	80,96%	85,61 %	82,80%
RP	Nenhum	-	62,29%	67,76 %	64,00%	-	-	-	-	-	-

por finalidade verificar a eficácia do método de classificação aqui proposto, sobre um conjunto amostral maior. Também se estuda a possibilidade de aplicar outras técnicas de extração de características, aliadas a mecânismos de classificação diferentes do SVM, visando realizar uma comparação direta com os resultados apresentados neste trabalho, contribuindo o densevolvimento contínuo do estado da arte.

ACKNOWLEDGMENT

Os autores agradecem ao Departamento de Informática da UEM, pela infraestrutura disponibilizada, a CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e ao Centro de Tecnologia da Universidade Estadual de Maringá por fornecer o apoio financeiro necessário para o desenvolvimento do presente trabalho.

REFERÊNCIAS

[1] R. T. Holmes, *Ecological and evolutionary impact of bird predation on forest insects: an overview*. Studies in Avian Biology, 1990, pp. 6–13.

[2] R. T. Holmes, J. C. Schultz, and P. J. Nothnagle, "Bird predation on forest insects: an exclosure experiment," vol. 206, pp. 462–463, 1979.

[3] D. W. Snow, "Evolutionary aspects of fruit-eating by birds," *Ibis*, vol. 113, no. 2, pp. 194–202, Apr. 1971.

[4] ——, "Tropical frugivorous birds and their food plants: a world survey," *Biotropica*, pp. 1–14, 1981.

[5] F. L. Carpenter, "A spectrum of nectar-eater communities," *American Zoologist*, vol. 18, no. 4, pp. 809–819, 1978.

[6] P. Feinsinger and R. K. Colwell, "Community organization among neotropical nectar-feeding birds," *American Zoologist*, vol. 18, no. 4, pp. 779–795, 1978.

[7] M. Proctor, P. Yeo, A. Lack *et al.*, *The natural history of pollination.* HarperCollins Publishers, 1996.

[8] F. Straube, "Newsletter bocev (bird observers club of the european valley," *Oct*, 2005.

[9] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524 – 1534, 2010, pattern Recognition of Non-Speech Audio.

[10] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[11] D. R. Lucio and Y. M. G. Costa, "Bird species classification using spectrograms," in *Computing Conference (CLEI), 2015 Latin American*. IEEE, 2015, pp. 1–11.

[12] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *The Journal of the Acoustical Society of America*, vol. 100, no. 2 Pt 1, pp. 1209–1219, Aug. 1996.

[13] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2740–2748, 1997.

[14] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study," *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2185–2196, 1998.

[15] S. Selouani, M. Kardouchi, E. Hervet, and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *2005 ICSC Congress on Computational Intelligence Methods and Applications*, 2005, pp. 6 pp.–.

[16] C. Kwan, K. C. Ho, G. Mei, Y. Li, Z. Ren, R. Xu, Y. Zhang, D. Lao, M. Stevenson, V. Stanford, and C. Rochet, "An automated acoustic systemto monitor and classify birds," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 52–52, Jan. 2006.

[17] E. Vilches, I. Escobar, E. Vallejo, and C. Taylor, "Data mining applied to acoustic bird species recognition," in *18th International Conference on Pattern Recognition, 2006. ICPR 2006*, vol. 3, 2006, pp. 400–403.

[18] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *3rd International Conference on Intelli-*

gent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007, Dec. 2007, pp. 293–298.

[19] C.-H. Chou, C.-H. Lee, and H.-W. Ni, "Bird species recognition by comparing the HMMs of the syllables," in *Second International Conference on Innovative Computing, Information and Control, 2007. ICICIC '07*, Sep. 2007, pp. 143–143.

[20] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1541–1550, Nov. 2008.

[21] F. Briggs, R. Raich, and X. Z. Fern, "Audio classification of bird species: A statistical manifold approach." in *ICDM*, W. W. 0010, H. Kargupta, S. Ranka, P. S. Yu, and X. Wu, Eds. IEEE Computer Society, 2009, pp. 51–60.

[22] C.-H. Chou and P.-H. Liu, "Bird species recognition by wavelet transformation of a section of birdsong," in *Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC'09. Symposia and Workshops on*. IEEE, 2009, pp. 189–193.

[23] M. Lopes, L. Gioppo, T. Higushi, C. Kaestner, C. Silla, and A. Koerich, "Automatic bird species identification for large number of species," in *2011 IEEE International Symposium on Multimedia (ISM)*, Dec. 2011, pp. 117–122.

[24] H. Tyagi, R. M. Hegde, H. A. Murthy, and A. Prabhakar, *Automatic Identification of Bird Calls Using Spectral Ensemble Average Voice Prints*, 2006.

[25] R. Semolini, "Support vector machines, inferência transdutiva e o problema de classificação," Ph.D. dissertation, Universidade Estadual de Campinas, 2002.

[26] R. O. Duda, P. E. Hart *et al.*, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[28] A. Rauber and M. Frühwirth, *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001 Darmstadt, Germany, September 4-9, 2001 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, ch. Automatically Analyzing and Organizing Music Archives, pp. 402–414.

[29] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity," in *Proc. ISMIR*, 2002, pp. 71–80.

[30] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[31] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *ISMIR*, 2005, pp. 34–41.

[32] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[33] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and signal processing*. Springer, 2008, pp. 236–243.

[34] W. Li, K. Mao, H. Zhang, and T. Chai, "Selection of gabor filters for improved texture feature extraction," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Sept 2010, pp. 361–364.

[35] J. Chen, V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Rlbp: Robust local binary pattern," in *BMVC*, 2013.

[36] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.

[37] Y. M. Costa, "Reconhecimento de gêneros musicais utilizando espectrogramas com combinação de classificadores," Ph.D. dissertation, Universidade Federal do Paraná, 2013.

[38] J. V. S. S. Stevens, "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[39] A. Pacheco and C. Pacheco, "Combinação de classificadores," 2013.

[40] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, Mar 1998.

[41] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[42] J. G. Martins, Y. Costa, D. Bertolini, and L. Oliveira, "Uso de descritores de textura extraıdos de glcm para o reconhecimento de padroes em diferentes domınios de aplicaçao," *XXXVII Conferencia Latinoamericana de Informática*, pp. 637–652, 2011.

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

[44] Y. Costa, L. Oliveira, A. Koerich, and F. Gouyon, "Music genre recognition using gabor filters and lpq texture descriptors," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, ser. Lecture Notes in Computer Science, J. Ruiz-Shulcloper and G. Sanniti di Baja, Eds. Springer Berlin Heidelberg, 2013, vol. 8259, pp. 67–74.

[45] Y. Costa, L. Oliveira, A. Koerich, F. Gouyon, and J. Martins, "Music genre classification using {LBP} textural features," *Signal Processing*, vol. 92, no. 11, pp. 2723 – 2737, 2012.

[46] V. N. Vapnik, *The nature of statistical learning theory*, Second Edition ed., ser. Statistics for engineering and information science. New York: Springer, 2000.

[47] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.