ANDRÉIA MARINI

CLASSIFICAÇÃO AUTOMÁTICA DE ESPÉCIES DE PÁSSAROS USANDO ESTRATÉGIAS SUPERFICIAIS E PROFUNDAS

Tese apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Doutor em Informática.

Curitiba
2014

ANDRÉIA MARINI

CLASSIFICAÇÃO AUTOMÁTICA DE ESPÉCIES DE PÁSSAROS USANDO ESTRATÉGIAS SUPERFICIAIS E PROFUNDAS

Tese apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Doutor em Informática.

Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. Alessandro L. Koerich

Curitiba
2014

Marini, Andréia
M339c Classificação automática de espécies de pássaros usando estratégias
2014 superficiais e profundas / Andréia Marini ; orientador, Alessandro L. Koerich.
– 2014.
147 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná,
Curitiba, 2014
Bibliografia: f. 129-135

1. Processamento de imagens. 2. Pássaro - Identificação. 3. Recuperação de imagem. 4. Informática. I. Koerich, Alessandro Lameiras. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em informática. III. Título.

CDD 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

PUCPR
GRUPO MARISTA

ATA DE DEFESA DE TESE DE DOUTORADO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ÁREA DE CONCENTRAÇÃO: CIÊNCIA DA COMPUTAÇÃO

DEFESA DE TESE DE DOUTORADO Nº 028/2014

Aos 03 dias de dezembro de 2014 realizou-se a sessão pública de Defesa da Tese de Doutorado intitulada **"Classificação Automática de Espécies de Pássaros usando Estratégias Superficiais e Profundas"** apresentada pela aluna **Andréia Marini** como requisito parcial para a obtenção do título de Doutor em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Alessandro L. Koerich
PUCPR (Orientador) _____ APROV
(assinatura) (aprov/reprov.)

Prof. Dr. Alceu de Souza Brito Junior
PUCPR _____ APROV.

Prof. Dr. Luiz Eduardo S. de Oliveira
UFPR _____ APROV

Prof. Dr. Carlos Nascimento Silla Junior
UTFPR _____ APROV

Prof. Dr. Alessandro Zimmer
UFPR _____ APROV

Conforme as normas regimentais do PPGIa e da PUCPR, o trabalho apresentado foi considerado APROVADO (*aprovado/reprovado*), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof.ª Dr.ª Andreia Malucelli
Coordenadora do Programa de Pós-Graduação em Informática

Programa de Pós-Graduação em Informática
PPGIa
PUCPR

Agradecimentos

Agradeço ao Prof. Dr. Alessandro Lameiras Koerich, pela orientação acadêmica e o constante apoio ao desenvolvimento deste trabalho. Felizmente, pude contar com sua ajuda durante o mestrado e outra vez durante todo o período de doutoramento. Aos demais professores do Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná, em especial aos professores do grupo de pesquisa Descoberta do Conhecimento e Aprendizagem de Máquina, Prof. Dr. Júlio César Nievola e Prof. Dr. Emerson Cabrera Paraiso, pelas discussões, esclarecimentos e suporte tecnológico. Ao professor Dr. Jaques Facon, pelo apoio na fase inicial deste trabalho.

Agradeço à Pontifícia Universidade Católica do Paraná, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Fundação Araucária pelo apoio financeiro que viabilizou a execução deste trabalho.

Agradeço aos meus colegas pelas conversas, pela atenção, pelos conselhos, pelos elogios e as críticas, e, principalmente, por sempre se importarem comigo. São alguns deles (em ordem alfabética) Adilson, Alef, Anderson, Andre, Aracélia, Cheila, Denise, Edenilson, Elias, Fabíola, Franciele, Irapuru, Jean, Jhonatan, Leila, Luciana, Mariza, Priscila, Ronan, Rosana, Tania e Viviane. Agradeço a colaboração do colega Luiz Gustavo Hafemann para integração das *Convolutional Neural Networks* ao trabalho.

Finalmente, agradeço a minha família, meus pais, meu irmão e sua esposa, pelo apoio incondicional que tornou possível a realização desta atividade, com dedicação exclusiva. Também agradeço ao meu afilhado Mateus, presente da vida, amor inestimável. Ao seu lado entendo a doçura de ser criança outra vez.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

BoK -	*Bag-of-keypoints*
CNN -	*Convolutional Neural Network*
CUB 200 -	*Caltech-UCSD Birds 200*
CUB 200 2011 -	*Caltech-UCSD Birds 200 2011*
DoG -	*Difference of Gaussian*
FFT -	*Fast Fourier Transform*
FGVC -	*Fine-Grained Visual Categorization*
HSV -	*Hue, Saturation, Value*
LBP -	*Local Binary Pattern*
MARSYAS -	*Music Analysis, Retrieval and Synthesis for Audio Signals*
MFCC -	*Mel-Frequency Cepstral Coefficients*
RBF -	*Radial Basis Function*
RGB -	*Red, Blue, Green*
SIFT -	*Scale-Invariant Feature Transform*
SVM -	*Support Vector Machine*

Resumo

A pesquisa referente a identificação visual, aplicada a conjuntos de dados de imagens sem restrições, apresenta muitos desafios, especialmente quando relacionadas a problemas de granularidade fina. Este trabalho utiliza a identificação de espécies de pássaros por meio de imagens (em particular os conjuntos de dados CUB 200 e CUB 200 2011) para contribuir e esclarecer aspectos referentes ao emprego de características visuais. Adicionalmente, é abordado o tema da identificação de espécies pelo paradigma de aprendizagem profunda. Dessa forma, é possível estabelecer comparação entre as estratégias superficiais e profundas para a identificação visual de granularidade fina. Além disso, o trabalho propõe dois novos métodos: um para a segmentação de imagens, baseado nas cores presentes na imagem e outro para a fusão de informação audiovisual na identificação de espécies de pássaros. Para realizar a fusão audiovisual foi necessário construir um conjunto de dados de áudio relacionado ao conjunto de dados de imagens. O trabalho evidencia outras contribuições pontuais quanto a abordagens de características visuais, suas influências e restrições em relação ao processo de identificação, ou de baixo nível de categorização, o qual exige o tratamento de diferenças sutis entre as classes. Finalmente, destaca-se que a identificação visual de espécies é um tema recente que possibilita a cooperação entre as comunidades científica, tecnológica, ecológica e a sociedade.

Palavras-chave: *1) Identificação visual, 2) Granularidade fina, 3) Áudio, 4) Estratégias superficiais e profundas.*

Abstract

There are many challenges related to the visual identification on unconstrained image datasets, particularly in fine-grained problems. This work contributes in several aspects related to visual features taking into account the bird species identification problem (evaluated on CUB200 and CUB200-2011). In addition, such a problem is tackle using the deep learning paradigm to allow a comparison between shallow and deep strategies in visual identification problems. Moreover, two novel methods are also proposed: a method for color image segmentation based and a method for bird species identification that relies on the fusion of visual data extracted from unconstrained bird images and acoustics data extracted from bird vocalizations. The evaluation of such audiovisual fusion approach was carried out on a self-made dataset that relates audio and image data. This work also contributes in other aspects related to the influence and the restrictions of visual features in the identification process and low-level categorization. Finally, an important remark is that visual identification is a recent research subject that foster the cooperation among scientific, technological and ecological communities and the society.

Keywords: *1) Visual Identification, 2) Fine Grained, 3) Audio, 4) Deep and Shallow Strategies.*

Capítulo 1

Introdução

Uma tendência para resolver problemas de reconhecimento visual é transferir para um sistema computacional a capacidade humana de lidar efetivamente com a complexa tarefa de distinguir um objeto ou uma categoria de objetos, mesmo quando ocorrem mudanças na escala, localização, iluminação, oclusão de partes, diferentes cenários e pontos de vista. Após décadas de pesquisas, o reconhecimento visual ainda é considerado uma tarefa difícil para computadores. O reconhecimento visual é apresentado na literatura de várias maneiras, tais como: a categorização, classificação ou identificação (abordagens que buscam o nome do objeto pertencente à imagem); localização (abordagens que buscam encontrar o objeto na imagem); segmentação (abordagens que buscam delinear o contorno do objeto e separá-lo do restante da imagem), e recuperação de objetos semelhantes (abordagens que buscam encontrar objetos semelhantes em uma grande quantidade de imagens). O paradigma dominante para a construção de sistemas de identificação de imagens tem sido proporcionado pela Aprendizagem de Máquina. O pressuposto de que a Aprendizagem de Máquina pode resolver este tipo de problema reside na capacidade desta de lidar com a complexidade inerente, aprendendo a partir de dados rotulados. Desta forma, cabe aos pesquisadores de visão computacional a concepção de modelos e representações dos recursos utilizados. Entretanto, uma quantidade substancial do trabalho pesado é feito por algoritmos de aprendizagem em relação aos dados coletados.

Dependendo de como um sistema de classificação de imagens é organizado, ele pode estar preparado para classificar os objetos que pertencem ao mesmo nível básico de categoria ou a baixo nível de categorização. No primeiro caso, para resolver o problema de classificação, o sistema deve identificar um objeto de interesse na imagem, por exemplo, uma mesa, uma cadeira, uma planta, um pássaro, um veículo, uma pessoa. Em geral, nesses casos existem diferenças significativas entre os objetos, que podem gerar características extremamente discriminantes. Ao contrário da categorização de nível básico, a

subcategorização, ou classificação de baixo nível pode ser entendida como um problema de granularidade fina, ou seja, problemas que exigem a identificação entre objetos fortemente semelhantes, muitas vezes diferenciados apenas por detalhes sutis. Neste caso, exemplos de resposta do sistema de classificação ao objeto de interesse, cadeira: cadeira da sala ou da cozinha; da planta: tóxica ou não tóxica; do pássaro: a que espécie pertence; e assim por diante. De acordo com Yao et al. (2012), este tipo de aplicação poderá se tornar uma tarefa importante para a área de visão computacional e útil em muitas aplicações reais.

1.1 Definição do problema

Um exemplo muito conhecido de categorização visual de baixo nível é o reconhecimento de espécies ou raças. Para exemplificar um problema de granularidade fina, será empregada a identificação de espécies de pássaros com duas classes. Consideramos dois pássaros da família Auklet: *Crested auklet* e *Parakeet auklet*, conforme ilustrados na Figura 1.1. Caso sejam utilizadas características simples, como um histograma (frequência de cores na imagem), torna-se praticamente impossível distinguir entre tais espécies. Mas, se pudermos identificar as principais diferenças visuais associadas às espécies e representadas na imagem, a tarefa se torna plausível. O pássaro *Crested auklet* possui o peito mais escuro e uma penugem na cabeça, enquanto o *Parakeet auklet* não possui penugem na cabeça e seu peito é mais claro. Nas condições em que a imagem foi adquirida, contudo, essas diferenças visuais podem não ficar evidentes. Deng et al. (2013) enfatiza que, com informações limitadas, a seleção automática de características discriminantes se torna difícil, pois, um grande número de características irrelevantes pode causar um alto *overfitting*, ou seja, o modelo não consegue obter uma boa generalização devido ao superajustamento dos dados. A Figura 1.2 apresenta outra particularidade de problemas de granularidade fina: a grande variabilidade intraclasse. Neste caso, todas as imagens são da espécie *Anna hummingbird*, porém, a depender do gênero, da idade e da maneira como a aquisição da imagem foi realizada, a variabilidade intraclasse pode ser maior ou menor.

Neste trabalho, adotaremos uma abordagem a fim de identificar as diferentes espécies de pássaros por meio de imagens. Desta forma, o problema de classificação pode ser definido como: dada uma imagem de um pássaro, identificar dentre um número fixo, porém com grandes possibilidades, a qual espécie ele pertence. O desafio da tarefa de classificação se deve à variação do fundo, iluminação e pose, já que a maioria das imagens são coletadas no habitat natural dos pássaros. Nestas imagens não é possível controlar rotação, escala e ângulo de visão no momento da aquisição.

Figura 1.1: À esquerda *Crested Auklet* no centro *Parakeet Auklet*. À direita superior *Great Grey Shrikes* e inferior *Loggerhead Shrikes* são alguns exemplos da grande similaridade entre as espécies (classes) em um problema de granularidade fina. (Fonte: Welinder et al. 2010).

Figura 1.2: Visão geral da variabilidade das amostras pertencentes à espécie *Anna Hummingbird*. Esse tipo de situação é comumente encontrado em classes de problemas de granularidade fina. (Fonte: Welinder et al. 2010).

Do ponto de vista prático, esse impasse é comum até mesmo entre profissionais e especialistas humanos. Um exemplo disso é um artigo publicado recentemente no website eBird[1]. Por meio do artigo, é discutida a identificação de duas espécies no período de migração. As sugestões oferecidas para identificar *Warblers bay-breasted* e *Blackpoll* são essencialmente de verificação de cores, pois a plumagem e o formato das duas espécies são muito semelhantes. O desafio, no contexto, é como proceder com pássaros que podem não mostrar as características distintivas mencionadas. Nesses casos, a sugestão é olhar atentamente para uma combinação de várias características diferentes e buscar um consenso para apontar para uma ou outra espécie.

A partir da exposição do problema, pretende-se por meio desta tese responder a algumas questões primordiais em relação a problemas de identificação visual de granularidade fina usando como estudo de caso a identificação de espécie de pássaros:

1. Informações geradas por algoritmos do Estado da Arte de extração de características e aprendizagem supervisionada, em imagens sem restrição de entrada, podem levar a uma abordagem viável?

2. A partir da hipótese de que o ser humano faz uso de características visuais para identificar espécies, o uso de características visuais que diferenciam espécies podem auxiliar na identificação de espécies de pássaros de maneira computacional, baseando-se em Aprendizagem de Máquina?

3. A considerar que a identificação de espécies de pássaros pode ser feita a partir de imagens ou sons emitidos pelos pássaros, um método de fusão de informações visuais e acústicas pode auxiliar e melhorar os resultados de um sistema de identificação automático?

4. Considerando que a utilização de características isoladas, ou a possibilidade de existir complementariedade entre diferentes características pode não ser suficiente para identificar espécies de pássaros, qual abordagem seria mais adequada para representar problemas de granularidade fina e os impactos desta abordagem nas taxas de identificação visual automática?

1.2 Objetivos

O objetivo geral deste trabalho é identificar espécies de pássaros por meio de aspectos relativos à identificação visual automática, no contexto de sistemas de classifica-

[1] www.ebird.org

ção supervisionada para problemas de granularidade fina, considerando duas abordagens principais de representação: superficial e profunda. Para este propósito, os objetivos específicos respectivos são detalhados a seguir:

- Lidar com um problema de granularidade fina em imagem digital, como ao identificar espécies de pássaros, desafio tanto para os seres humanos quanto para algoritmos computacionais que pretendam realizar essa tarefa de forma ágil e automática;
- Investigar uma tarefa de classificação desafiadora devido à variação do fundo e iluminação quando a maioria das imagens foi coletada em habitat natural. Nessas imagens não é possível controlar rotação, escala e ângulo de visão, no momento da aquisição de imagens;
- Aplicar abordagens diferenciadas de extração de características e de classificação ao identificar automaticamente espécies de pássaros. Particularmente, discutir estratégias superficiais e profundas.
- Aumentar a precisão da identificação, que poderá servir como segunda opinião ao especialista humano ou à implementação de aplicativos relacionados;
- Realizar fusão de informações audiovisuais. Aproveitar-se da ideia de que a identificação espécies de pássaros pode ocorrer por meio de imagens ou áudios, agrupá-los, para verificar o comportamento do sistema de identificação automática;
- Descrever o comportamento do problema de identificação automática de espécies de pássaros em casos específicos, computacionalmente efetivos, para lidar com a complexidade e a variabilidade dos dados, considerando também a avaliação do desempenho dos algoritmos sobre os conjuntos de dados contemplados e a definição de um sistema de referência para tais casos.
- Avaliar qual o tipo de representação, superficial ou profunda, consegue fornecer melhor representação em relação à coleta de características. Especificamente, avaliar se a abordagem de aprendizagem de características pode ser melhor do que as características especializadas.

1.3 Justificativas

Para esta tese, duas abordagens centrais são justificadas: motivos que tornam problemas de granularidade fina peculiares, e a escolha da identificação de espécie de pássaros para representá-los. As subseções a seguir discutem os dois casos.

1.3.1 Problemas de Granularidade fina

O reconhecimento visual tem sido uma área de pesquisa ativa em visão computacional, de modo que as abordagens atuais fazem uso de algoritmos de Aprendizagem de Máquina. De forma geral, para melhorar o desempenho de um sistema, é necessário coletar maiores conjuntos de dados de treinamento para superar a variabilidade e aprender a reconhecê-los. Assim, torna-se importante construir modelos mais poderosos, extrair características mais discriminantes e usar técnicas para melhorar e para prevenir *overfitting*. Atualmente, existem muitas alternativas para reconhecer objetos em nível básico, em que é preciso distinguir categorias como, por exemplo: cadeiras, veículos, animais, eletrodomésticos, entre outros. Contudo, o reconhecimento de objetos em baixo nível ainda é um campo de pesquisa que permanece com muitos desafios. A resolução de problemas de granularidade fina com múltiplas classes e poucos exemplos associados à cada classe é uma questão desafiadora para a aprendizagem de máquina, o que pode ser sintetizado da seguinte forma: dada uma imagem de entrada com um objeto de interesse em primeiro plano predizer a classe do objeto pertencente à imagem.

Ao contrário da categorização de nível básico, a categorização de baixo nível, que pode ser encontrada em problemas de granularidade fina, muitas vezes se apresenta como um desafio para seres humanos[2], mesmo altamente experientes, tendo em vista que, existem menos recursos visuais discriminantes em relação ao reconhecimento no nível básico. Em geral, para esse tipo de problema é difícil obter uma quantidade razoável de dados rotulados, visto que, na maioria dos casos, é necessário um especialista de conhecimento profundo do domínio específico.

Ao buscar estabelecer as diferenças entre problemas de granularidade fina e reconhecimento de objetos, podemos afirmar que, no primeiro caso, distinguir entre as classes pode ser mais difícil com apenas características visuais sutis disponíveis. No segundo, é comum uma abundância de características visuais úteis. Com informações limitadas, a seleção automática de características discriminantes se torna complexa, já que, o grande número de características irrelevantes pode causar problemas ao algoritmo de aprendizagem (Deng, Krause, e Fei-Fei, 2013).

Avanços na tecnologia e a revolução da informação trouxeram um crescimento exponencial de dados de imagem não processados. Segundo Fei-Fei (2013), fundamentado por dados da empresa Cisco[3], a estimativa de consumo global de tráfego de dados na Internet por mês foi de 21 Exabytes em 2011, será de 48 Exabytes em 2014 e de 83

[2]Reconhecimento visual é uma das funcionalidades fundamentais dos seres humanos, é tão importante, que a natureza dedicou mais de 50% dos neurônios no cérebro para esse fim (Deng, 2012).

[3]http://www.cisco.com/

Exabytes em 2016. Sendo neste último cenário 86% tráfego de conteúdo visual. As estimativas indicam que empresas como o You Tube[4] deverão gerenciar 72 horas de vídeos por minuto e como o Facebook[5], 300 milhões de imagens por dia. É fato que dispositivos móveis estão cada vez mais presentes na vida humana. Por exemplo, novos modelos de celulares são produzidos e lançados a todo momento. Juntamente com esses novos modelos, as câmeras fotográficas estão cada vez mais sofisticadas e com maior definição de imagens. Desta forma, um grande número de pessoas pode gerar inúmeras imagens coletadas sem restrições, que devem ser gerenciadas, recuperadas e utilizadas de diferentes formas.

Nos últimos anos, está disponível uma grande quantidade de conjuntos de dados, de imagens, dedicadas a assuntos específicos, tais como: imagens de plantas, animais, objetos domésticos, veículos etc. Para implementar um sistema de classificação automático viável, os conjuntos de imagens disponíveis, em especial imagens rotuladas, são um componente crítico para a aprendizagem de máquina, considerando-se que, o objetivo de um algoritmo de aprendizagem é o de produzir um modelo generalista a partir dos dados existentes. Em muitos casos, são necessárias centenas de milhares ou milhões de imagens de alta resolução já rotuladas e com diversas informações adicionais. Um exemplo desse tipo de conjunto de dados a IMAGEnet[6], que consiste de mais de 15 milhões de imagens de alta resolução rotuladas em mais de 20000 categorias. De forma geral, enfatiza-se que existem muitas alternativas para reconhecer objetos em nível básico, mas nem tantas para reconhecimento de baixo nível. A construção do conjunto de dados para o reconhecimento visual de baixo nível se apresenta aos pesquisadores como um problema e, em alguns casos, apenas um especialista poderá construí-la.

O desenvolvimento de um sistema de visão computacional, no contexto de granularidade fina, passa por um desafio adicional relacionado à variabilidade nas imagens de um conjunto de dados. Ao exemplificar por meio do problema da identificação de espécies, dada uma imagem de um pássaro qualquer, o sistema automático deverá identificar a que espécie ele pertence. Podemos observar a grande quantidade de espécies, aproximadamente 10.000 espécies de pássaros, que podem aparecer em diferentes cenários (seu habitat natural, como florestas ou parques, entre grupos de outros animais etc.), apresentar diferentes poses, tamanhos e ângulos de visão. As imagens podem apresentar variações de iluminação e partes do pássaro podem ser omitidas por outros elementos do cenário. Além disso, podem ocorrer variação de cores, rotações, escolha de diferentes

[4]https://www.youtube.com/
[5]https://www.facebook.com/
[6]http://image-net.org/

escalas. A próxima seção descreve detalhadamente o conjunto de dados escolhido para tratar granularidade fina por meio da identificação de espécies de pássaros.

O resultado da tarefa de classificação deste tipo de problema pode trazer informações relevantes para um usuário e ser útil na forma de aplicações reais. Problemas de granularidade fina se tornaram populares ao longo dos últimos anos. Exemplos de categorização visual de baixo nível podem ser observados em sistemas de identificação de espécies, raças, insetos, plantas, alimentos, entre outros.

1.3.2 A importância da identificação de espécies de pássaros

Avanços na tecnologia têm permitido novas abordagens de coleta de dados referentes ao meio ambiente em um importante domínio de aplicação: o monitoramento de ecossistemas (Kasten, McKinley, e Gage, 2010). Nossa compreensão dos sistemas ecológicos ainda é limitada. Recursos como satélites, redes de sensores, técnicas de reconhecimento de padrões e visão computacional têm fornecido ferramentas úteis para a aquisição de dados ambientais em larga escala. No entanto, de acordo com Acevedo et al. (2009), a coleta de dados da biodiversidade, especialmente para a fauna, ainda é limitada devido à necessidade de identificar espécies por seres humanos.

A avaliação da biodiversidade é um desafio para ecologistas, biólogos e profissionais que pretendem descrever ou quantificar biodiversidade em escalas ecologicamente relevantes e fornecer sínteses oportunas e interpretações que possam permitir decisões responsáveis que reduzem os riscos para espécies ameaçadas de extinção, populações e habitats. Glotin et al. (2013) enfatiza que este tipo de tarefa é essencial para um desenvolvimento sustentável e pertinente ao contexto atual, quando muitos fatores gerados por demandas econômicas podem ocasionar mudanças ambientais globais da flora e da fauna. Entretanto, tais informações muitas vezes estão parcialmente disponíveis para professores, cientistas e cidadãos e, de modo frequente, incompletas para ecossistemas com grande diversidade.

De forma geral, vários impasses que permanecem insolúveis, oriundos da necessidade de conservar ecossistemas, podem ser tratados com a aplicação de aprendizado de máquina. Alguns exemplos listados por Glotin et al. (2013): o uso de inferência Bayesiana para inferir as velocidades de aves migratórias, a partir de dados de radares meteorológicos e modelagem de distribuição de espécies; uso de conjuntos de dados, com registros de observações da presença ou ausência de espécies em determinados locais (podem ser desenvolvidos modelos que possam prever a presença ou ausência em outros lugares). Além disso, os interesses em conhecer o comportamento de espécies podem ser explorados na

migração de aves, na quantidade de horas de voo de mariposas, no processo de migração do salmão, na propagação de espécies invasoras, na sobrevivência de espécies ameaçadas de extinção, entre outros destacados por Glotin et al. (2013).

Particularmente, identificar a espécie a que um pássaro pertence desperta interesse de diferentes grupos de admiradores, pela beleza das aves e de seu canto e também de especialistas, pela sua importância ecológica. Identificar aves é uma atividade complexa, bem conhecida pelos ornitólogos, e considerada uma tarefa científica desde a antiguidade. Os ornitólogos estudam as aves; sua existência na natureza, sua biologia, seus cantos, sua distribuição e seu impacto ecológico. A classificação de espécies de pássaros, geralmente, é feita por especialistas em ornitologia com base em um sistema proposto por Linnaeus: Unido, Filo, Classe, Ordem, Família, e Espécie[7].

Existem algumas razões práticas para observar, estudar ou monitorar pássaros. Cientistas costumam usar pássaros para estudar e compreender ecossistemas, por serem numerosos, sensíveis às mudanças ambientais, mais fáceis de controlar do que outras espécies, estão por toda parte e são relativamente fáceis de serem vistos. Diversas aplicações do mundo real podem utilizar as aves como coadjuvantes, tais como: monitoramento de poluição ambiental (Lovett, 2012), avaliação da qualidade do meio ambiente (Bardeli, Wolff, Kurth, Koch, Tauchert, e Frommolt, 2010) e indicadores de sustentabilidade[8]. De acordo com Bardeli et al. (2010), pássaros são um eficaz indicador para alterações na biodiversidade, porque são distribuídos sobre uma ampla gama de paisagens, são fáceis de detectar em comparação com outros grupos de animais e temos um bom conhecimento a respeito da maioria das espécies. Em muitos casos, é necessário estimar o número de pássaros em áreas ecologicamente sensíveis, como por exemplo, reservas naturais ou em áreas de difícil acesso. Portanto, a adoção de métodos automatizados para identificar espécies de aves é uma forma interessante para avaliar a quantidade e diversidade das aves que aparecem em uma região e podem auxiliar em várias aplicações práticas. Desta forma, as razões práticas mencionadas anteriormente justificam o estudo de mecanismos para a identificação de espécies de pássaros.

1.4 Principais contribuições

Considerando a produção científica relativa a problemas de granularidade fina, esta pesquisa tem como objetivo apresentar contribuições pontuais que abordam aspectos referentes à identificação visual. Tendo em vista o aspecto temporal desta pesquisa, a

[7] http://www.birdsinbackyards.net
[8] http://www.birding.com

contemporaneidade e a decorrente diversificação de métodos aplicados para a identificação de espécies de pássaros, enfatiza-se a inovação do tema em face ao Estado da Arte, uma vez que a maioria dos trabalhos abordados foi desenvolvida em paralelo a essa pesquisa no período de 2012 a 2014.

A perspectiva recente do tema estabelece novas abordagens, suas regras são inovadoras, ainda estão em adaptação e não são totalmente compreendidas. Essa condição atual permite ressaltar o fato de que o conhecimento científico poderá ser constituído a partir de contribuições pontuais, suas influências e restrições. Na sequência, é apresentada, a fim de se perceberem as contribuições científicas desta pesquisa de tese, uma lista das principais contribuições:

- Novo método de segmentação de imagens baseado nas cores presentes na imagem;
- Novo método de fusão de informação audiovisual na identificação de espécies de pássaros;
- Construção de um conjunto de áudio relacionado às espécies pertencentes ao conjunto CUB 200 2011. O conjunto CUB 50 Songs possui as gravações de áudio de 50 espécies. As amostras foram manualmente segmentadas, gerando um novo conjunto de áudios. Ambos os conjuntos podem ser explorados em novas pesquisas;
- Definição de um sistema de referência para a aplicação do operador de texturas nos conjuntos CUB 200 e CUB 200 2011;
- Comparação do tema da identificação de espécies por meio dos paradigmas de aprendizagem superficial e profunda;
- Combinação de classificadores, em termos abstratos e sem apoio probabilístico, considerando classificadores que adotam abordagens superficiais e profundas;
- A verificação de que existe complementariedade entre diferentes características visuais, mesmo que a variação de desempenho entre elas seja alta.

A partir das contribuições descritas foi possível derivar outras contribuições pontuais:

- Estudos detalhados de abordagens de identificação visual baseadas exclusivamente em cores e texturas, já publicadas no formato de artigo científico em conferências nacionais e internacionais da área;

- A conclusão de que as características visuais percebidas pelos seres humanos podem auxiliar no processo de identificação; contudo, não são suficientes;
- A aplicação de mecanismos de rejeição com variações no compromisso erro e rejeição se apresenta como uma alternativa viável para lidar com um problema tão peculiar. Além disso, torna possível a fusão de informação de diferentes classificadores;
- A combinação de classificadores pode auxiliar no aumento da taxa de acerto de um sistema automático de identificação de espécies. Embora esse aumento não seja significativo, a complementariedade entre as características é fortemente evidenciada;
- A análise de erros para compreender a similaridades entre as imagens incorretamente classificadas e a organização dessas imagens em uma estrutura em que seja possível a interpretação dos resultados. A investigação é feita a partir do grupo de imagens incorretamente classificadas, resultantes da aplicação de todos os classificadores utilizados no decorrer da pesquisa;
- A constatação de que a organização taxonômica: Ordem, Família, Gênero e Espécie reflete nos erros do classificador. As peculiaridades de uma espécie interferem diretamente nas taxas de acerto do classificador.

Por meio deste trabalho, a geração de um saber científico sobre o uso dessas características visuais pode, na práxis, propiciar melhor desempenho no estabelecimento de sistemas automáticos de identificação de espécies. Além disso, é possível estabelecer cooperação entre a comunidade científica, tecnológica, ecológica e a sociedade.

Contribuições na esfera social podem ser alcançadas com a aplicação dos conhecimentos obtidos por meio desta pesquisa na construção de ferramentas de identificação automática e multimídia para auxiliar na identificação de espécies que, geralmente, é desafio árduo para o público em geral e, muitas vezes, difícil até mesmo para profissionais, porque exige a manipulação de nomes taxonômicos. Aplicações de sistemas multimídia podem gerar soluções auxiliares para facilitar a manipulação de nomes taxonômicos. Essas soluções são de interesse de profissionais para a identificação de espécies em campo, de educadores como ferramenta educacional, ou de observadores e admiradores da avifauna por meio de um aplicativo de entretenimento, entre outros.

Outras contribuições do ponto de vista ecológico podem ser vinculadas pelo interesse de agências reguladoras no monitoramento de tempo real em ecossistemas. Informações precisas e atualizadas em relação à avifauna possibilita que as agências reguladoras possam avaliar impactos ambientais das atividades humanas e proporcionar medidas mais

rápidas e efetivas. Certamente, a comunicação entre as comunidades facilita o uso de informação e conhecimento científico na implantação de políticas e programas que possam beneficiar a sociedade.

1.5 Organização do trabalho

*E*ste trabalho está organizado em sete capítulos, conforme descrição a seguir: *Capítulo 1* - enuncia as principais motivações, desafios e justificativas para contextualizar os temas a serem abordados nesta tese; *Capítulo 2* - aborda o referencial teórico, sob o enfoque das principais técnicas relacionadas aos temas centrais deste trabalho; *Capítulo 3* - trata de uma detalhada revisão bibliográfica com relação aos temas abordados nesta tese; *Capítulo 4* - detalha os métodos estabelecidos para verificar as proposições iniciais, alicerçados na investigação bibliográfica; *Capítulo 5* - apresenta um novo método de fusão audiovisual na identificação de espécies de pássaros; *Capítulo 6* - evidencia os resultados obtidos, bem como comparações e análises realizadas; *Capítulo 7* - expressa as conclusões, as principais contribuições desta tese, bem como sugestões de continuidade e trabalhos futuros.

1.6 Considerações finais

No presente capítulo, enunciamos a relevância da investigação de novas abordagens para identificar espécies de pássaros, estabelecendo conexões ecológicas e tecnológicas entre Visão Computacional, reconhecimento sonoro, Aprendizagem de Máquina, entre outras áreas analisadas. Adicionalmente, exemplificamos a complexidade e os problemas da identificação visual de granularidade fina. Além disso, esclarecemos a motivação para esta pesquisa, seu escopo, objetivos e contribuições.

Capítulo 2

Fundamentação Teórica

Neste capítulo é apresentada uma visão geral sobre os principais elementos para a construção de um sistema de identificação automático de espécies. Primeiro, são fornecidas algumas definições necessárias para estabelecer o referencial teórico que fundamenta este trabalho. Em seguida, são descritas as principais práticas adotadas pelas metodologias atuais para a identificação de espécies de pássaros por meio de imagens.

2.1 Conjunto de dados Caltech CUB 200

Em particular sobre a identificação de espécies de pássaros muitos autores têm relatado resultados utilizando o conjunto de dados Caltech-UCSD Birds 200 (CUB 200). CUB 200 é um conjunto de imagens proposto por Welinder et al. (2010) com 200 espécies de pássaros (principalmente norte-americanos) no qual são disponibilizadas 6.033 imagens rotuladas e diversos atributos associados. Tal base permite o estudo caracterizado de dados, o que não é possível com outros conjuntos mais populares que se concentram no nível básico de informações. Todas as anotações realizadas na imagem foram obtidas por meio de usuários da interface *Mechanical Turk*[1]. Foram coletados dois tipos de anotações, a segmentação aproximada e a caixa delimitadora. A segmentação aproximada foi obtida do desenho do usuário com pincel grosso para tocar todos os *pixels* de contorno do pássaro em primeiro plano. A caixa delimitadora foi deduzida a partir das informações do desenho do usuário. A Figura 2.1 apresenta exemplos de anotações efetuadas através de uma interface visual e critérios pré-estabelecidos. Em outra tarefa, os usuários são convidados a fornecer anotações de atributos. São coletados 25 atributos visuais quando o usuário observa uma imagem. Durante o processo eles são questionados em relação ao nível de confiança de sua rotulação em três alternativas: certeza (definitivamente);

[1]Disponível em: https://www.mturk.com/

certo (provavelmente), e adivinhação. Estão disponíveis arquivos no formato Matlab com diversas anotações referentes a cada imagem, resultantes de coletas com 1577 usuários. A Figura 2.2 apresenta uma visão geral de amostras pertencentes ao conjunto de dados CUB 200 e CUB 200 2011.

Figura 2.1: Exemplos de anotações realizadas em imagens do conjunto de dados Caltech-UCSD Birds 200 (CUB 200). Para cada imagem é atribuída a informação de segmentação aproximada (contorno na cor verde) em seguida é estimada a caixa delimitadora e o *ground truth* (Welinder et al. 2010).

Figura 2.2: Visão geral de amostras pertencentes ao conjunto de dados CUB 200 e CUB 200 2011 (Welinder et al. 2010).

O conjunto de dados Caltech-UCSD Birds CUB 200-2011 (Wah et al. 2011) é uma versão atualizada do conjunto CUB 200 (Welinder et al. 2010). Na nova versão são disponibilizadas 11.788 imagens, sendo em média 30 imagens para cada classe, para o conjunto de treinamento, e o restante para o conjunto de teste. Todas as informações disponi-

bilizadas anteriormente são mantidas e atualizadas para novas imagens. Em relação à versão anterior, a novidade é a adição de informação referente à localização de 15 partes do corpo e da cabeça do pássaro. A Tabela 2.1 exemplifica as partes coletadas indicadas pelos usuários. A Tabela 2.2 apresenta os atributos associados a cada parte. O *ground truth* da localização de partes foi obtido da média da identificação indicadas por cinco usuários do *Mechanical Turk*.

Tabela 2.1: Coleções de partes coletadas para o conjunto de dados CUB 200 2011 (Wah et al. 2011).

Partes da Cabeça	*beak, nape, throat, left eye, right eye, crow e fore-head*
Partes do Corpo	*breast, belly, left leg, right leg, back, left wing, right wing e tail*

Tabela 2.2: Atributos associados às partes coletadas para o conjunto de dados CUB 200 2011 (Wah et al. 2011).

has bill shape cone	has wing color brown	has belly pattern solid
has upperparts color orange	has breast pattern striped	has breast color purple
has size small	has primary color yellow	has crown color white

2.2 Estratégias superficiais e profundas

Uma escolha importante para a construção de sistemas de classificação, reconhecimento e identificação de objetos de interesse em imagens disponíveis no Estado da Arte são características visuais para a representação de uma imagem (Chatfield et al., 2014). Nos últimos anos, uma questão crucial para as áreas de reconhecimento de padrões e visão computacional é como extrair características visuais que sejam robustas as mudanças que uma imagem pode sofrer, por exemplo, escala, pontos de vista, condições de iluminação e cenários. O que se observa é que algumas técnicas estão em uso há mais de uma década. É o caso de técnicas como: *Bag-of-Visual-Words*, *Fisher Vector*, SIFT, entre outros descritores visuais (Chatfield et al., 2014).

Observando as práticas atuais para implementar um sistema de identificação de imagens, é necessário que a mesma seja processada e reduzida a um conjunto de descritores chamados de vetor de características. Este processo de grande relevância é conhecido como extração de características. A grande maioria dos métodos de extração de características buscam uma maneira de expressar numericamente a maior quantidade de atributos inerentes ao objeto de interesse pertencente à imagem, usando a menor quantidade possível de informação, para em seguida ser tratada por algoritmos de aprendizagem de máquina.

Recentemente, na área de visão computacional, esta abordagem de extração de características foi substancialmente superada pela introdução de Redes Neurais Convolucionais. Estas redes têm uma estrutura mais sofisticada do que as representações padrão, pois compreendem várias camadas de extratores de características não-lineares, o que permite defini-la como profunda (Chatfield et al., 2014) (em contraste a representação clássica referida no parágrafo anterior que pode ser definida como superficial). A Figura 2.3 organiza as principais técnicas para as abordagens superficiais e profundas no contexto da Aprendizagem de Máquina. Cabe enfatizar que esta tese utiliza duas abordagens diferentes. Uma abordagem superficial supervisionada por meio do SVM e outra profunda, por meio das redes neurais convolucionais. Isto possibilita a comparação entre ambas.

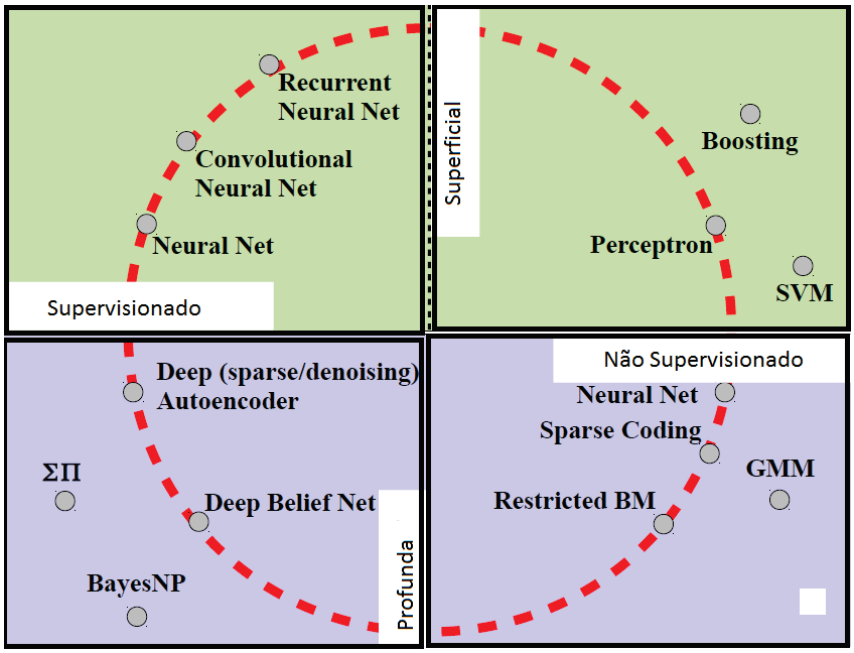


Figura 2.3: Visão geral das principais abordagens superficiais e profundas relacionadas a aprendizagem de máquina (Ranzato, 2013).

Na literatura existem muitas abordagens para a extração de características, algumas mais abrangentes e outras fortemente direcionadas a domínios específicos. Tais métodos foram recentemente referenciados como métodos de representação superficial ou estratégias de abordagens superficiais. Esta tese utiliza importantes algoritmos de extração de características por meio de cores, texturas e pontos de interesse. As subseções 2.3.1, 2.3.2 e 2.3.3 descrevem como são extraídas as características visuais que pertencem à estratégia de abordagem superficial. Em seguida, a subseção 2.5 descreve a abordagem do problema de granularidade fina por meio de representação profunda, caso em

que a extração de características não é relevante. O fato de não realizar a etapa de extração de características e explorar os dados disponíveis, permitindo uma aprendizagem de características, torna-se cada vez mais popular e eficaz para o reconhecimento visual. Em particular, este trabalho utilizará as redes neurais convolucionais, um método de aprendizagem profunda que envia *pixels* através de várias camadas, transformando-os em características que são utilizadas por uma rede neural para identificação de espécies de pássaros.

2.3 Extratores de características

Uma imagem digital possui uma descrição matemática de uma matriz onde cada elemento é denominado *pixel*. Segundo Gonzalez e Woods (2000) o termo imagem digital refere-se a uma função de intensidade luminosa bidimensional, denotada por $f(x, y)$, em que o valor ou amplitude de f em coordenadas (x, y) resulta na intensidade da imagem em ponto específico, ou *pixel*. Como a luz é uma forma de energia, $f(x, y)$ deve ser positiva e infinita. Desta forma, as imagens que as pessoas percebem em atividades visuais consistem de luz refletida dos objetos. Embora todas as informações encontrem-se dentro deste conjunto de *pixels*, em muitos casos a imagem inteira possui muita informação redundante ou que não é útil para descrever seu conteúdo. Neste caso, os extratores de características são utilizados para obter especificamente o tipo de característica de interesse. As subseções descrevem os extratores utilizados para obter as características de cor, textura e pontos de interesse de uma imagem digital.

2.3.1 Características de cor

A habilidade de reconhecer padrões é uma característica dos seres humanos. O grande desafio é desenvolver máquinas tão eficientes quanto as máquinas humanas (Duda, Hart, e Stork, 2000). Aproveitando a ideia de que os seres humanos conseguem perceber diferentes cores, novas abordagens baseadas em cores pertencentes em uma imagem vem sendo utilizadas para a implementação de sistemas de classificação. O reconhecimento de espécies de pássaros efetiva-se, normalmente, pela observação, utilizando características visuais dos pássaros, tais como as cores, os tamanhos, os formatos, etc. Neste contexto, a relação com esse trabalho se estabelece principalmente pela abordagem do reconhecimento baseado em cores. As características de cores estão sempre disponíveis, distribuídas uniformemente entre os indivíduos da mesma espécie, permitindo em muitos casos, a identificação da espécie em questão. A identificação de espécies de pássaros a

partir de imagens digitais poderá, igualmente, utilizar uma abordagem baseada em cores. As cores pertencentes a uma imagem podem ser representadas de formas diferentes, dependendo do espaço de cor a que pertencem (por exemplo RGB[2] e HSV[3]).

Antes que a extração de características seja realizada, uma mudança na representação de cores da imagem poderá revelar, em alguns casos, informações antes desconhecidas. Os efeitos resultantes dessas transformações serão investigados no decorrer desta tese. A característica cor representa um atributo forte, por tratar-se dos argumentos utilizados pelos especialistas humanos para identificar a qual espécie um pássaro pertence. O impacto da característica cor no processo de segmentação, consequentemente, na performance dos classificadores quando aplicados à classificação de espécies de pássaros, é discutido em Marini et al. (2013) e em particular nos capítulos 4 e 6 deste trabalho.

2.3.1.1 Histograma de cores

Em um histograma colorido cada *pixel* é associado a uma cor específica, uma cor poderá ser descrita em diferentes espaços de cores. Formalmente, um histograma colorido pode ser construído por meio de uma imagem RGB pela função $f(x, y)$ com $x = 0, 1, ..., N - 1$ e $y = 0, 1, ..., M - 1$ possuindo valores discretos $i = 0, 1, ..., G - 1$, onde G é o valor possível de intensidade, N, M são as dimensões em x e y da imagem. Em cada canal de cor na imagem considera-se uma função $f_c(x, y)$ onde c indica o canal de cor. Segundo Gonzalez e Woods (2000) o histograma de cores é uma função discreta, mostrando o número de *pixels* da imagem que possuem uma determinada cor. Conforme a Equação 2.1, r_k é a representação do k-ésimo elemento de cor, n_k é o número de *pixels* da mesma cor na imagem, n o número total de *pixels* na imagem e k é o número de valores de *pixel* na imagem. Finalmente p_k é o resultado da função discreta representada por r_k.

$$p_k(r_k) = \frac{n_k}{n} \tag{2.1}$$

[2] O modelo RGB (do inglês: *Red, Blue, Green*) é baseado em um sistema de coordenadas cartesianas. É um modelo de cor baseado na síntese aditiva, com o qual é possível representar uma cor com diferentes adições das três cores primárias. Os componentes do modelo RGB podem variar entre 0 e 255, definindo-se um cubo de cor, onde o valor (0,0,0) corresponde ao preto, e o valor (255,255,255) corresponde ao branco. Trata-se um modelo muito utilizado sendo aplicado, por exemplo, em câmeras e monitores de vídeo. Tal modelo apresenta uma grande correlação entre seus canais. A alteração em um canal reflete na imagem inteira (Gonzalez e Woods 2000).

[3] O modelo HSV (do inglês: *Hue, Saturation, Value*) representa uma cor em torno dos valores da matiz, saturação e valor. Quando o valor de S diminui sem alterar o valor de V, isto corresponde a adicionar branco na cor inicial. Quando o valor de V diminui sem alterar o valor de S, isto corresponde a adicionar preto na cor inicial. Quando ambos os valores de S e V diminuem, vários tons da cor inicial são compostos (Gonzalez e Woods 2000).

2.3.2 Características de textura

Textura pode ser entendida como característica intrínseca de uma imagem. Apesar de não haver um conceito formal para textura entende-se como um padrão que recobre um objeto que remete à ideia de regularidade (Mäenpää, Pietikäinen, Maenpaa, e Pietikainen, 2004), (Mäenpää e Pietikäinen, 2005). Diferentes métodos de extração e representação de texturas podem ser utilizados, os quais podem ser classificados em estatísticos (distribuições espaciais de níveis de cinza), geométricos (propriedades de elementos de textura ou primitivas) e de processamento de sinais (análise da frequência da imagem para classificar a textura) (Tuceryan e Jain, 1993). O descritor utilizado nesta tese é o Padrão Binário Local (do inglês - *Local Binary Pattern* LBP). Este método é utilizado para a extração de características globais, tendo sido descrito pela primeira vez por (Ojala, Pietikäinen, e Harwood, 1996). Devido ao poder discriminante e à sua simplicidade computacional, este operador de texturas tem sido largamente utilizado e se apresenta com sucesso em diferentes domínios de aplicação, tais como inspeção visual (Mäenpää, Turtinen, e Pietikäinen, 2003), reconhecimento facial (Choi, Plataniotis, e Ro, 2010), reconhecimento visual de objetos (Zhu, Bichot, e Chen, 2010), (Zhu, Bichot, e Chen, 2011) e reconhecimento de gêneros musicais (Costa, Oliveira, Koerich, Gouyon, e Martins, 2012).

2.3.2.1 Padrões Binários Locais

O operador binário local representa uma medida geral para microtextura unificando modelos estatísticos e estruturais de análise de textura e sendo invariante a alterações de intensidade e de baixo custo computacional, isso permite aplicá-lo em problemas de processamento em tempo real. Em imagens coloridas podem ser aplicados em problemas com grande variação de iluminação, variação de poses ou diferentes resoluções de imagens (Ojala, Pietikäinen, e Mäenpää, 2002).

A extração da textura da imagem é realizada através de um operador e organizada em um histograma de ocorrência desses padrões. O operador permite a detecção dos padrões binários em uma vizinhança circular, em qualquer espaço angular, e independente de resolução da imagem. O LBP associado a um *pixel* de uma imagem é calculado por P vizinhos simétricos usando uma circunferência de raio R, conforme a Figura 2.4. Em seguida os *pixels* de uma imagem são rotulados pelo operador de textura gerando um mapa LBP, limiarizando os vizinhos de cada *pixel* da imagem, e atribuindo um código binário para cada um destes *pixels* (Ojala, Pietikäinen, e Mäenpää, 2002). O operador LBP pode ser definido por meio de uma textura T em uma vizinhança de uma imagem monocromá-

tica (distribuição em níveis de cinza de P (P> 1). Definido a partir da intensidade g_p de cada amostra, de forma que $0 \leq p < P$ e da intensidade do *pixel* g_c. O nível de cinza g_c corresponde ao valor do *pixel* central, e os valores g_p ($p = 0, ..., P-1$) aos valores de cinza dos P *pixels* vizinhos em uma circunferência de raio $R(R>0)$. Caso a posição do *pixel* central g_c sejam $(0, 0)$, as coordenadas de g_p serão obtidas por $(-Rsin(2\pi/P), Rcos(2\pi/P))$ (Ojala, Pietikäinen, e Mäenpää, 2002).

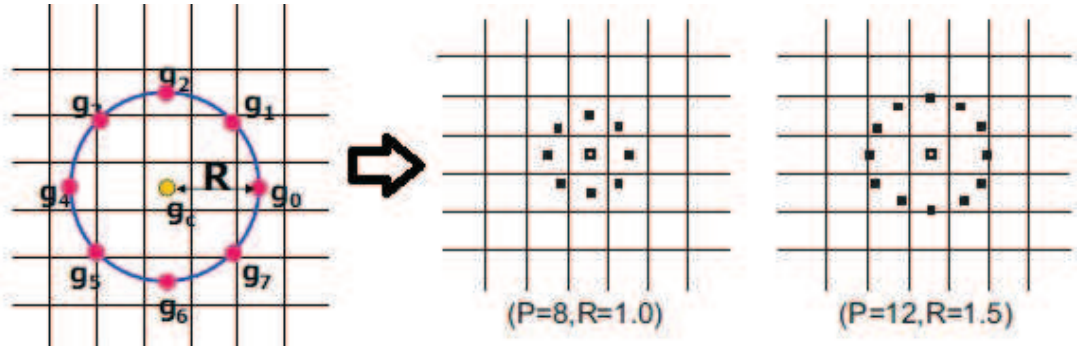


Figura 2.4: Conjunto de vizinhos igualmente espaçados em uma superfície circular para diferentes P e R (Ojala, Pietikäinen, e Mäenpää, 2002).

O valor de textura da imagem é obtido subtraindo o valor do *pixel* central g_c dos valores de cada *pixel* da sua vizinhança g_p (p=0,...,P-1) conforme a Equação 2.2:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c) \tag{2.2}$$

Assim, assumimos que a diferença de $g_p - g_c$ seja independente de g_c de forma que pode ser representada pela Equação 2.3.

$$T \approx t(g_c)t(g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c) \tag{2.3}$$

O operador descrito é sensível a variações na escala de cinza, então para obter invariância de textura independente de alterações na escala de cinza de uma imagem consideramos apenas os sinais dos resultados ao invés de seu valor exato, como na Equação 2.4. Onde S(x) = 1 se x for não negativo, e 0 caso contrário.

$$T = t(s(g_0 - g_c, s(g_1 - g_c), ..., s(g_{p-1} - g_c)) \tag{2.4}$$

Em seguida, atribuindo a potência 2^p para cada sinal $s(x)$, obtém-se um valor único que identifica cada estrutura da textura obtendo o operador $LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p$.

O operador $LBP_{P,R}$ gera valores correspondentes aos 2^P diferentes padrões que

podem ser estabelecidos por P *pixels* vizinhos. Quando uma imagem sofre uma rotação, os P *pixels* irão transitar no perímetro do círculo ao redor de g_c, como g_0 sempre é associado ao *pixel* à direita de g_c uma rotação dos valores resultará (assinalado um identificador único para cada padrão, definido através da Equação 2.5) em uma mudança no valor do operador $LBP_{P,R}$. Desta forma $ROR(x, i)$ fará uma rotação nos bits associados a x, i vezes, buscando o menor número inteiro entre os bits, ou seja, que tenha a maior quantidade de zeros nos bits mais significativos.

$$LBP^{ri}_{P,R} = minROR(LBP_{P,R}i)|i = 0, 1, ..., P - 1 \tag{2.5}$$

Esse trabalho utiliza os operadores $LBP^{u2}_{P,R}$, $LBP^{ri}_{P,R}$, $LBP^{riu2}_{P,R}$. No primeiro, a medida *u2* corresponde ao número de transições espaciais, bit a bit 0 ou 1 de alterações entre sucessivos bits na representação circular LBP. São designados padrões que tenham valor U de no máximo 2 como "uniforme"e usadas as estatísticas de ocorrência de diferentes padrões uniformes para discriminação da textura (Ojala, Mäenpää, e Pietikäinen, 2001). No segundo, o operador é invariante à rotação e igualmente invariante a transformações monotônicas da escala de cinza, portanto, a textura de cada *pixel* pode ser definido, pelas Equações 2.6 e 2.7. Finalmente, o operador LBP é uma excelente medida espacial e estrutural para texturas de uma imagem. O operador LBP pode ser aplicado em imagens coloridas, à adaptação mais usual é o processamento dos canais de cores separados. Em seguida os valores obtidos em cada canal podem ser utilizados individualmente ou concatenados para compor um vetor de características de texturas coloridas.

$$LBP^{riu2}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{se } U(LBP_{P,R}) \leq 2 \\ P + 1 & \text{caso contrário,} \end{cases} \tag{2.6}$$

Onde

$$\begin{aligned} U(LBP_{P,R}) = &|s(g_{P-1} - g_c) - s(g_0 - g_c)| \\ &+ \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \end{aligned} \tag{2.7}$$

2.3.3 Características de forma

O método utilizado para extrair as características estruturais é o *Scale Invariant Feature Transform* (SIFT) (Lowe, 1999) e (Lowe, 2004). O algoritmo SIFT detecta e extrai descritores locais em uma imagem. Estes devem ser encontrados em quantidades para cobrir densamente o objeto de interesse e possibilitar o reconhecimento. Este método transforma uma imagem numa coleção de vetores de características, as quais apresentam

invariância em relação à translação, rotação, e escala de uma imagem, e parcialmente invariantes a mudanças de iluminação. Um vetor descritor gerado a partir da técnica SIFT armazena informações extraídas de gradientes locais (magnitude e orientação)[4].

2.3.3.1 Algoritmo SIFT

A geração dos descritores SIFT[5] baseia-se em quatro etapas principais, citando Lowe (2004): 1) detecção dos extremos; 2) localização dos pontos-chave; 3) atribuição da orientação; 4) construção do descritor.

A primeira etapa é a **detecção de extremos** que consiste em encontrar máximos e mínimos locais por meio de uma filtragem em cascata com a função DoG (*Difference of Gaussian*), buscando as características que sejam mais estáveis no espaço escalar. A primeira fase de detecção dos pontos-chave é identificar locais e escalas que podem ser atribuídos repetidamente em relação a diferentes pontos de vista de um mesmo objeto. Dada uma imagem de entrada $I(x, y)$ a representação $L(x, y, \sigma)$ é gerada a partir da convolução ($*$) de uma função Gaussiana (uma imagem borrada é gerada a partir da aplicação da função Gaussiana sobre uma imagem original), $G(x, y, \sigma)$, com a imagem de entrada. Sendo este filtro variável à escala através do parâmetro σ, conforme a Equação 2.8:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2.8}$$

Conforme e Equação 2.9, a DoG é obtida pela subtração das imagens filtradas com escalas variadas por uma constante k, ou seja, é a diferença entre as imagens borradas por um filtro gaussiano nas escalas σ e $k\sigma$. Desta forma, é possível detectar variações de intensidades, como por exemplo, bordas. Cada conjunto, composto pelas imagens da filtragem Gaussiana e pelas resultantes das diferenças (DoG) entre elas, forma uma oitava, como exemplifica a Figura 2.5. Após, a imagem filtrada com σ tem seu tamanho reduzido à metade e serve de entrada para a próxima geração de oitavas, formando uma pirâmide. O número de oitavas não é fixo. Depois da geração da pirâmide, usa-se a imagem do intervalo DoG, e compara cada *pixel* com seus vizinhos adjacentes. A Figura 2.6 apresenta a detecção de extremos no espaço de escala. A próxima etapa é definir a

[4] A implementação utilizada nesta tese é disponibilizada pela biblioteca VLFEAT (Vedaldi e Fulkerson 2010) compatível como trabalho de ((Lowe, 2004)).

[5] Uma variação do algoritmo SIFT é o Dense SIFT. Trata-se de uma mudança que torna o algoritmo rápido para computar conjuntos densos de descritores SIFT Vedaldi e Fulkerson (2010).

localização dos pontos-chave e fazer o descarte de pontos instáveis.

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{2.9}$$

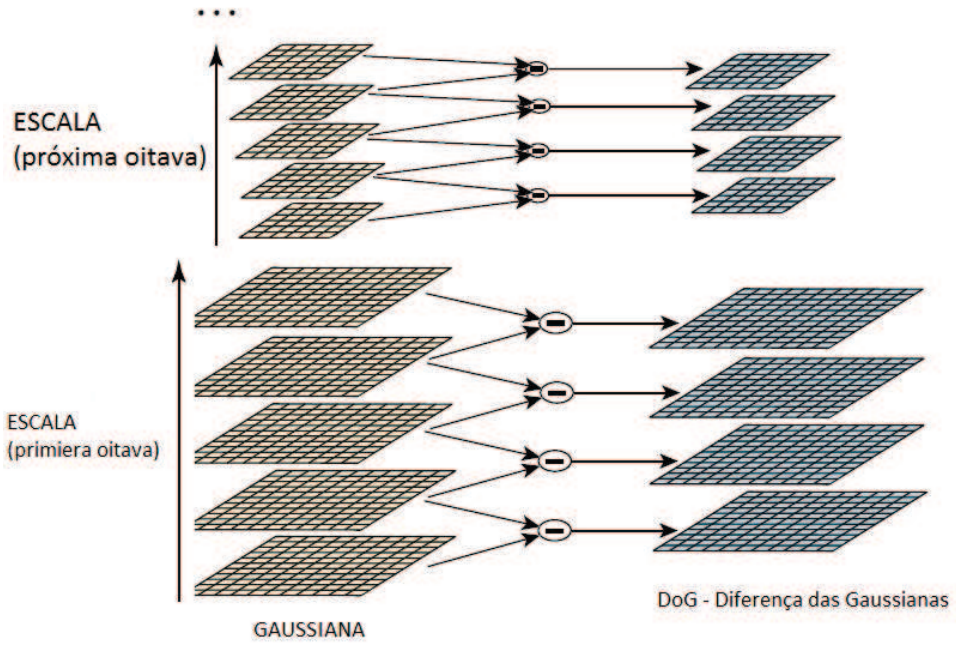


Figura 2.5: Representação da obtenção das DoG para as oitavas de uma imagem (Lowe 2004).

A segunda etapa é a **localização dos pontos-chave.** Nesta etapa, realizam-se as localizações precisas dos pontos-chave. O método aplicado consiste em ajustar uma função quadrática 3D do ponto de amostragem local, de modo a determinar uma localização interpolada do máximo da amostragem. Isto é feito utilizando uma expansão por série de Taylor da DoG aplicada à imagem, ou seja $D(x, y, \sigma)$, deslocada de modo que a origem desta expansão esteja localizada no ponto de amostragem. Conforme a Equação 2.10 é possível estimar a localização do extremo, representado por $\hat{X}$. Os pontos detectados na etapa anterior como extremos são candidatos a pontos-chave, dos quais se deseja calcular a localização exata. Nesta etapa é possível rejeitar pontos de baixo contraste (logo, sensíveis a ruídos) ou aqueles mal localizados ao longo de uma aresta. Segundo Lowe (2004) recomenda-se definir algum limiar para descartar os pontos de baixo contraste e aumentar a estabilidade, de forma que $(|D(\hat{X})| < \mathit{limiar})$.

$$\hat{X} = -\frac{\delta^2 D^{-1}}{\delta x^2} \frac{\delta D}{\delta x} \tag{2.10}$$

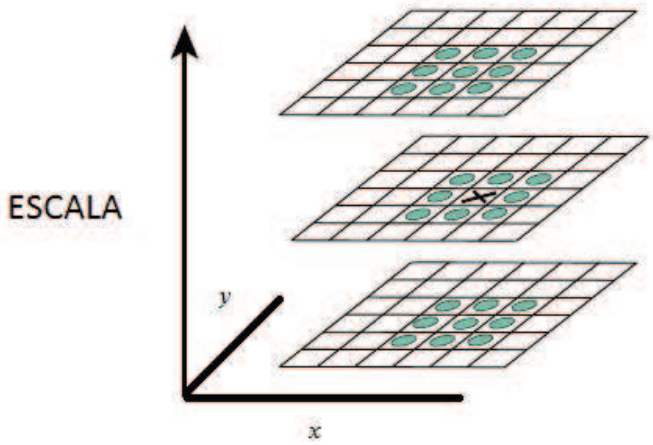
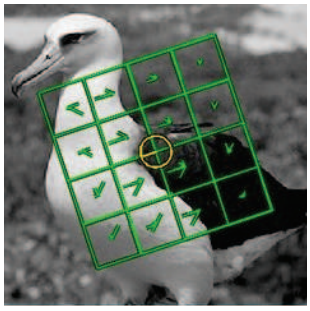


Figura 2.6: Detecção de extremos no espaço de escala (Lowe 2004).

A Figura 2.7 apresenta a imagem original de um pássaro, seguida da imagem convertida para a escala de cinza e a fixação de orientação para extração dos pontos-chave.

(a) Imagem original.

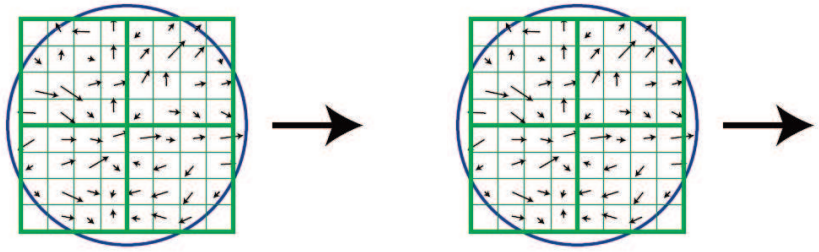


(b) *Quadro* de orientações.

Figura 2.7: Exemplo de *quadro* que fixa as orientações para o algoritmo SIFT.

A terceira etapa, **atribuição da orientação**, é baseada em propriedades locais de cada ponto-chave, o que possibilita alcançar invariância à rotação da imagem. Nesse momento, seleciona-se a imagem filtrada que tenha escala Gaussiana mais próxima e de oitava referente ao ponto em análise. Isto garante que os cálculos sejam realizados em escala invariante. Constrói-se então um histograma de orientações para os *pixels* da região do ponto-chave. Lowe (2004) indica 36 valores para σ, abrangendo 360 graus de orientações, divididos em 10 intervalos. Aos pontos vizinhos são dados três pesos, sendo o primeiro, conforme uma função de distância normalizada entre a orientação dos *pixels* e a orientação do ponto-chave; o segundo, com base na magnitude $m(x, y)$ e o terceiro,

usando uma janela gaussiana circular com o valor de σ 1,5 vezes maior do que a escala do ponto-chave. Com esses pesos o histograma é atualizado. Os máximos no histograma de orientação representam as direções dominantes dos gradientes locais. Além do pico do histograma, também são usados, para definir a orientação, outros picos com valor acima de 80% em relação ao maior. No final, ainda se aplica um ajuste parabólico aos três valores mais próximos de cada pico, a fim de interpolar a posição do máximo. Desse modo, a próxima etapa é construir o descritor.



(a) Gradientes de uma imagem. (b) Descritor de ponto-chave.

Figura 2.8: Detalhes da construção do descritor. O histograma tem 8 valores de orientação, cada um criado ao longo de uma janela de apoio de 4x4 *pixels*. O vetor de características resultante tem 128 elementos com uma janela de apoio de 16x16 *pixels* (Lowe 2004).

A quarta etapa é a **construção do descritor local**. Segundo Lowe (2004), para construir o descritor local é necessário: selecionar a imagem filtrada e a oitava referente ao ponto-chave; rotacionar as coordenadas do descritor e as orientações dos gradientes em relação à orientação do ponto chave; aplicar uma ponderação Gaussiana (igual ou a metade da janela do descritor) para atribuir um peso à magnitude de cada ponto vizinho, enfatizando gradientes mais afastados; definir $k\,x\,k$ *pixels* em cada $n\,x\,n$ regiões em torno do ponto-chave; implementar em cada região um histograma de 8 direções, com base nas magnitudes dos *pixels*. Para o caso de um conjunto 4 x 4 de histogramas com 8 células de acumuladoras o vetor de características gerado é 4 x 4 x 8 = 128 valores para cada ponto-chave; construir o descritor local a partir do vetor que armazena os valores do histograma; calcular a normalização do vetor para a unidade desejada.

2.3.3.2 Vocabulários e histogramas

O trabalho de Csurka et al. (2004) apresenta uma abordagem simples e eficiente para a categorização visual genérica, utilizando vetores de características construídos a partir de descritores de partes de imagens (condição que pode ser obtida com a utilização do algorítimo SIFT descrito na seção 2.3.3). Esta abordagem tem sido avaliada em

diferentes conjuntos de dados e se apresenta como um método para produzir bons resultados, mesmo sem explorar as informações geométricas. O método de *bag of keypoints* (vetor de pontos chave) é baseado na quantização vetorial de descritores. A abordagem vetor de pontos chave é análoga ao método de aprendizagem utilizando a representação *bag-of-words* para a categorização de texto. Um vetor de pontos-chave corresponde a um histograma do número de ocorrências de determinados padrões em uma imagem. As principais vantagens deste método são a simplicidade, eficiência computacional, invariância a transformações, oclusão, iluminação e as variações intraclasses.

A quantização dos descritores agrupados em um vetor serão os pontos-chave visuais associados à imagem. O vocabulário é construído a partir de partes de uma imagem, levando em conta a precisão de correspondência de cada parte e o poder de expressão avaliados sobre o restante do conjunto Sivic e Zisserman (2003). A quantização vetorial é realizada pelo algorítimo K-*means*. Em particular uma variação proposta por Elkan (2003) que acelera o cálculo em encontra os mesmos agrupamentos do algoritmo original.

2.4 Máquinas de vetores de suporte

Máquinas de vetores de suporte (do inglês: *Support Vector Machines* - SVM) se apresentam como uma abordagem amplamente utilizada para treinamento de classificadores. Foram propostas e fundamentadas na teoria de aprendizado estatístico, que estabelece uma série de princípios que devem ser seguidos para se obter classificadores com boa capacidade de generalização (Vapnik, 2000). A ideia central das SVMs é encontrar o hiperplano ótimo para separar um conjunto de dados, enquanto há, teoricamente, infinitos hiperplanos para separar o conjunto de dados. Um hiperplano é escolhido, de modo que a distância até o ponto de dados mais próximo de ambas as classes é maximizada. Os pontos que medem o hiperplano são os vetores de suporte. O SVM é um classificador fundamentado na teoria da maximização marginal a fim de minimizar o erro de generalização. Inicialmente, o hiperplano separava os dados linearmente em duas classes: positivas e negativas. Em implementações seguintes a abordagem foi reformulada para ser aplicada em situações mais genéricas, considerando casos de problemas que não sejam linearmente separáveis e problemas com múltiplas classes.

Para exemplificar o funcionamento do classificador SVM [6] considere um conjunto de dados $D = \{x_1, x_2, ...x_l\}$ (onde l representa o número de observações). Em um problema de classificação binária, onde cada instância de D pertence a uma classe positiva ou negativa, de modo que $\{x_i, y_i\}, i = 1, ..., l$, tal que $y_i \in \{+1, -1\}$ e $x_i \in R^d$.

[6] Adaptado de Chang e Lin (2013) e Fan et al. (2008).

A separação das instâncias pertencentes as classes positivas e negativas é realizada por meio do hiperplano. A Equação 2.11 apresenta como são definidos os pontos para o hiperplano. Desta forma, w é o vetor perpendicular ao hiperplano e $\frac{|b|}{\|w\|}$ estabelece a distância entre o hiperplano e o ponto de origem. A Equação 2.12 indica que uma margem define as fronteiras para cada classe em ambos os lados do hiperplano.

$$w.x + b = 0 \tag{2.11}$$

$$y_i(w.x + b) \geq 1 \tag{2.12}$$

Os exemplos do conjunto de dados de treinamento que definem o hiperplano, satisfazem a Equação 2.12, estão localizados sobre as fronteiras das classes e chamados de vetores de suporte. O objetivo do classificador é maximizar a margem por meio do posicionamento do hiperplano[7]. É usual representar o algorítimo de construção do classificador SVM por multiplicadores de Lagrange, neste caso α_i. A posição ótima do hiperplano pode ser obtida por maximização, conforme a Equação 2.13, sujeita as restrições das Equações 2.14 e 2.15.

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i . x_j \tag{2.13}$$

$$w = \sum_i \alpha_i y_i x_i \tag{2.14}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.15}$$

O lado do hiperplano onde encontra-se um novo exemplo a ser classificado pode ser calculado pela Equação 2.16. Entretanto, nem sempre os dados podem ser separados perfeitamente, quando isso ocorre uma variável de custo C é introduzida para penalizar as classificações incorretas do conjunto de treinamento. Essa penalidade estabelece uma restrição em relação aos multiplicadores de Lagrange, de maneira que $0 \leq \alpha_i \leq C$. A Função kernel localizada no núcleo do classificador SVM tem a função de mapear as características em um espaço de dimensão maior. Por exemplo, na Equação 2.16 existe a função kernel $K(x, x_i)$, onde x representa o vetor de entrada, x_i os exemplos de

[7]Por meio de Geometria é possível mostrar que a largura da margem é $\frac{|2|}{\|w\|}$ e que a margem do hiperplano pode ser maximizada minimizando $\|w\|^2$ sujeito a restrição imposta pela Equação 2.12.

treinamento e y_i a classe associada a cada exemplo.

$$f(x) = sgn(\sum_i \alpha_i y_i K(x.x_i) + b) \tag{2.16}$$

A função kernel utilizada $K(x, x_i)$ é conhecida como função kernel linear, pois encontra o produto dos vetores de entrada em um espaço linear. Além da separação linear a implementação do classificador SVM pode contemplar a classificação em dados não linearmente separáveis. As funções de kernel não lineares comumente utilizadas são a *Radial Basis Function* (RBF)e a Polynomial.

As SVMs são eficazes na classificação de conjunto de dados que são linearmente separáveis. Para os vetores com grande número de características, um SVM linear pode ser utilizado. Fundamentado pelo teorema de Cover[8] que afirma que um conjunto de dados não linear no espaço de entrada, pode ser transformado em um espaço de características com alta probabilidade dos objetos serem linearmente separáveis. Mapeia-se o vetor para um espaço de maior dimensão e aplica-se um SVM linear sobre esse espaço. Assim, é possível encontrar o hiperplano com maior margem de separação e boa generalização.

Os valores dos rótulos e as probabilidades estimadas podem ser armazenadas. O LIBSVM (Chang e Lin, 2013) utiliza o método proposto por Platt[9] para converter essa saída em uma probabilidade. Neste método, uma função sigmóide estimada empiricamente é utilizada para mapear as saídas descalibradas do SVM em probabilidades, retornando um valor entre 0 e 1 para cada resultado do classificador.

2.5 Aprendizagem profunda

Questões relativas aos aspectos da aprendizagem superficial e aprendizagem profunda são, atualmente, de grande interesse da comunidade científica que atua em visão computacional, processamento de sinais e linguagem natural. No enfoque superficial, como já discutido, a ideia central é encontrar o conjunto de características mais discriminante para ser submetido à avaliação de um algoritmo de Aprendizagem de Máquina. Entretanto, a ênfase na aprendizagem profunda é a possibilidade de ocorrer a aprendizagem de características nos dados disponíveis. De forma geral, o termo aprendizagem profunda refere-se a métodos de Aprendizagem de Máquina que possuem uma arquitetura com múl-

[8] COVER, 1965 *apud* HAYKIN, 2001
HAYKIN, S. Neural Networks: A Comprehensive Foundation. 2 ed, Prentice-Hall, Upper Saddle River, NJ, USA, 1999.

[9] *Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods*. De autoria de John C. Platt. Parte do livro: Advances in Large Margin Classifiers. Páginas: 61–74. MIT Press, 1999.

tiplas camadas e utilizam efetivamente esse recurso. Segundo Bengio (2009) arquiteturas profundas são compostas de vários níveis de operações não-lineares, por exemplo, em redes neurais e modelos escondidos de Markov. Desta forma, a profundidade de uma arquitetura refere-se ao número de operações não lineares atribuídas a ela. A inspiração para a organização destes modelos é o conhecimento atual da anatomia do cérebro do humano e a constatação de que o cérebro parece considerar muitas camadas para o sistema visual humano. Implementar arquiteturas profundas é uma tarefa de otimização difícil, mas alguns modelos de aprendizagem profunda vem sendo utilizados com sucesso em muitos domínios, incluindo várias tarefas de classificação de imagens. Alguns exemplos são os trabalhos de Krizhevsky et al. (2012), Sermanet et al. (2012), Ciresan et al. (2012).

2.5.1 Redes neurais para arquiteturas profundas

Uma rede neural de múltiplas camadas, tipicamente utilizada na aprendizagem supervisionada para fazer previsão ou classificação, por meio de uma série de camadas de neurônios, com operações não lineares, pode ser entendida como uma arquitetura profunda. O neurônio artificial é a unidade básica das redes neurais artificias. Um neurônio artificial pode ser representado pela Equação 2.17, onde *sigm* é a função sigmoide que pode ser descrita por $sigm(x) = \frac{1}{1+e^{-x}}$, sendo x_i a entrada i, w_i são os pesos associados às entradas e b valores de *bias*.

$$f(x) = sigm(\sum_{i=1}^{n} x_i w_i + b) \tag{2.17}$$

Uma rede neural de múltiplas camadas, segundo Bengio (2009), pode ser definida quando uma camada k, normalmente conectada com todas as outras camadas adjacentes ($k-1$ e $k+1$), produz um vetor de saída h^k, aproveitando a saída h^{k-1} nas próximas camadas. A saída das camadas é calculada por uma função de ativação para todos os neurônios da camada, conforme a Equação 2.18. Iniciando com uma entrada de $x = h^0$ e vetor de *bias* b^k para cada neurônio da camada e matriz de pesos w^k para cada par de neurônios da camada k e $k-1$. A função tangente hiperbólica ($tanh$) aplicada a cada um dos neurônios pode ser substituída[10] por $sigm(u) = 1/(1+e^{-u}) = \frac{1}{2}(tanh(u)+1)$.

$$h^k = tanh(b^k + w^k h^{k-1}) \tag{2.18}$$

A camada de saída h^k é utilizada para predizer e combinar com o objetivo da saída y juntamente com a função de perda $L(h^l, y)$ tipicamente convexo em $b^l + w^l h^{l-1}$. Na

[10]Outra função seria a Logística (ou *identidade*) representada pela função: $f(x) = 1/(1+e^{-x})$.

saída da camada utiliza-se uma função de ativação, por exemplo, *softmax*. A Equação 2.19 apresenta a definição da função *softmax*. Essa fase, de aplicação da função de ativação, que também pode ser chamada de propagação para frente, consiste na ativação dos neurônios, iniciando na primeira camada, depois das entradas até a camada de saída.

$$h_i^l = \frac{e^{b_l^i + w_l^i h^{l-1}}}{\sum j e^{b_j^l + w_j^l h^{l-1}}} \tag{2.19}$$

A saída *softmax* h_i^l pode ser usada para estimar $P(Y = i|x)$, com a interpretação de que Y é a classe associada com o padrão de entrada x. Neste caso, muitas vezes usa-se $L(h^L, y) = -log\, P(Y = y|x) = -log\, w_l^y$ como uma função de perda, ou função de erro, cujo valor esperado para (x, y) deverá ser minimizado. O objetivo do treinamento é minimizar a função de erro. A fase chamada de retropropagação consiste em calcular a função de erro em relação aos parâmetros do modelo (pesos e *bias*), propagando o erro das camadas de saída, para as camadas iniciais, uma camada de cada vez. Alguns algoritmos podem ser utilizados para minimizar a função de erro, como é o caso do método estocástico descida do gradiente (Lecun et al., 1998).

2.5.2 Aprendizagem de características

A utilização de uma abordagem profunda permite que sejam aprendidas hierarquias de vários níveis de características. Para a aprendizagem de características em nível de *pixel* podemos fazer uso de uma rede neural convolucional. Basicamente, a estrutura consiste em alternar camadas de convolução e de agrupamento, seguidas de camadas locais e camadas totalmente conectadas antes da camada de saída. A seguir, descrevemos os tipos de camadas mais utilizadas e um breve resumo de suas funções (Krizhevsky, 2014):

- **Camada de convolução**: permite a aplicação de conjunto de filtros na imagem de entrada. Esses filtros também chamados de mapas de características, são aplicados em toda a entrada da camada. Para cada filtro, cada neurônio está apenas ligado a um subconjunto de neurônios na camada anterior. No caso de imagens, os filtros podem definir uma pequena área de *pixels* (por exemplo, 3x3 ou 5x5 *pixels*), e cada neurônio é ligado aos neurônios mais próximos na camada seguinte. Os pesos são compartilhados entre neurônios, levando os filtros a aprender padrões frequentes que ocorrem em qualquer parte da imagem.

- **Camada de agrupamento**: implementa uma função de redução da resolução não linear de redução, a fim de reduzir a dimensionalidade e capturar pequenas variações.

Uma camada de agrupamento pelo valor de máximo possui uma pequena área (por exemplo, 3x3, 5x5) que reduz o número de neurônios, utilizando somente o valor mais alto da entrada. Essa camada normalmente é utilizada após uma camada de convolução.

- **Camada conectada localmente**: essa camada é muito semelhante às camadas convolucionais, mas apenas conecta os neurônios à próxima camada, sem compartilhar os pesos. A motivação para este tipo de camada é conseguir algo semelhante a um dos benefícios de camadas convolucionais: reduzir o número de conexões de entrada para cada neurônio.

- **Camada totalmente conectada**: a função desta camada é simplesmente multiplicar a sua entrada por uma matriz de pesos. Normalmente é definida com o número de saídas igual ao número de classes do problema.

Algumas camadas podem ser configuradas internamente com parâmetros específicos. Esses parâmetros podem interferir diretamente no resultado final. Algumas das principais opções foram descritas por Krizhevsky et al. (2012):

- **Neurônio**: a forma padrão para modelar a saída de um neurônio f como uma função de entrada x dada por $f(x) = tanh(x)$ ou $f(x) = (1 + e^x)^{-1}$. Os neurônios - Unidades Lineares Retificadas (Relus) utilizados com as redes neurais convolucionais podem ser mais rápidos do que seus equivalentes com função ($tanh$). Em termos de tempo de treinamento a função não-linear Relu representada por $f(x) = max(0, x)$ é muito mais rápida.

- **Momento de abandono**: a técnica chamada de *abandono* (*dropout*), consiste em definir zero na saída de cada neurônio com probabilidade igual a 0,5. Os neurônios que não fazem parte desta condição não participam da propagação. Esta técnica reduz a adaptação de neurônios, pois, a presença de um determinado neurônio na rede não é garantida. Desta forma, é necessário aprender características mais robustas. Essa técnica, praticamente dobra o número de iterações necessárias para a rede convergir.

- **Aumento do número de exemplos para o modelo por meio da extração de subimagens**: é a possibilidade de ampliar artificialmente o conjunto de dados, em tempo de execução, utilizando translações da imagem original, preservando rótulos em busca da diminuição de *overffiting*.

A Figura 2.9 apresenta uma arquitetura convolucional para o reconhecimento de dígitos em imagens de endereços residenciais. É possível perceber a organização das camadas, os estágios de aprendizagem e de classificação.

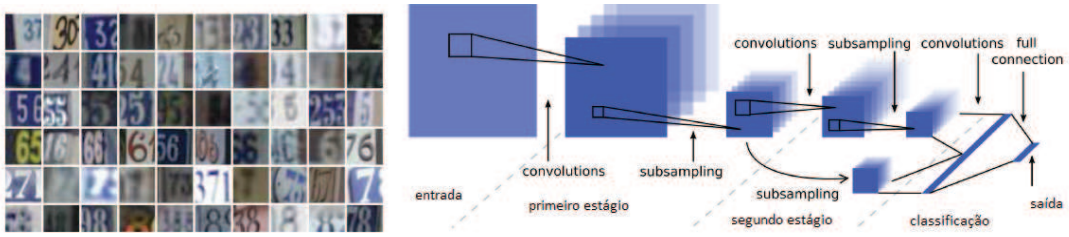


Figura 2.9: Imagens de números residenciais recortadas em 32x32. Cada amostra é rotulada por um digito de (0 até 9). A arquitetura da rede neural convolucional aplicada ao reconhecimento de dígitos que consegue obter uma taxa de acerto de 95,10% (Sermanet, Chintala, e Lecun, 2012).

2.6 Considerações finais

Neste capítulo foram abordadas as principais estratégias e técnicas que compõem os métodos proposto nesta tese. Eles foram selecionados de acordo com critérios que levam em conta a quantidade e qualidade das informações geradas e a semelhança com a presente proposta em relação aos objetivos a serem alcançados. O próximo capítulo apresenta uma abrangente revisão bibliográfica em relação a problemas de granularidade fina e identificação de espécies de pássaros.

Capítulo 3

Estado da Arte

Problemas de granularidade fina tornaram-se objeto de estudo há pouco tempo na área de visão computacional. De forma geral, os trabalhos recentes concentram-se em algumas abordagens que podem organizar a apresentação do Estado da Arte destes trabalhos. Este capítulo está organizado da seguinte forma: a) Definição de um sistema de referência para problemas de granularidade fina, especificamente na identificação de espécies; b) Métodos que consideram interação humana para solução de problemas de granularidade fina; c) Métodos que utilizam abordagens de segmentação nas imagens; d) A forte tendência em resolver problemas de granularidade fina por meio de métodos que consideram partes do objeto de interesse. e) Demais abordagens adotadas. O capítulo é finalizado com uma análise referente aos trabalhos descritos.

3.1 Sistema de referência

Tomando como ponto de partida a introdução do conjunto Caltech-UCSD Birds 200 (CUB 200) (Welinder et al., 2010) e sua atualização (Wah et al., 2011) descritos no capítulo 2, esta seção detalha alguns experimentos que servem como sistema de referência para problemas de granularidade fina e também para a identificação de espécies de pássaros no conjunto de 200 espécies de pássaros tipicamente americanos. Os experimentos realizados para a primeira versão do conjunto de dados CUB 200 ((Welinder et al., 2010)) consideram características simples como o tamanho da imagem em *pixels*: altura e largura e um histograma de cores com 10 intervalos por canal de cores; e também a análise de 128 componentes principais (*Principal Component Analysis* - PCA). Inicialmente, o classificador escolhido para definição do sistema de referência foi o vizinhos mais próximos (*Nearest Neighbor* - NN) que obteve um desempenho de 0,6% de taxa de acerto com a utilização da característica tamanho da imagem. Com a utilização de um histograma de cores da

imagem o resultado obtido foi de 1,7%. Outro experimento sugerido como sistema de referência foi detalhado no trabalho de Branson et al. (2010) o qual foi o primeiro artigo a utilizar o conjunto de dados. Neste experimento são utilizadas características extraídas por meio do algoritmo SIFT[1] em conjunto com a utilização do classificador SVM. Outro aspecto importante é que considera 15 imagens por classe no conjunto de treinamento e o restante das imagens para o conjunto de testes. Essa abordagem obtém uma taxa de acerto de 19% e inspira muitos outros trabalhos posteriores. Com as alterações realizadas na nova versão do conjunto de dados CUB 200 o sistema de referência foi redefinido em novos experimentos que consideram a utilização de partes do pássaro. No primeiro caso, o desempenho obtido foi de 10,26% pelo classificador SVM, considera a imagem inteira, informações referentes a partes do pássaro, utilizando no máximo 52 imagens por classes no conjunto de treinamento. No segundo caso, o *ground truth* referente a informações de partes foi incluído e a taxa de acerto obtida foi de 17,3%.

3.2 Métodos que consideram interação humana

Um trabalho recente proposto por Branson et al. (2014) apresenta um sistema de reconhecimento visual para problemas de granularidade fina. O sistema é híbrido, composto de habilidades humanas e de máquinas trabalhando em conjunto. Dessa forma, são combinados os pontos fortes e complementares de algoritmos de visão computacional, e usuários humanos não especialistas no domínio. Este trabalho é fundamentado na seguinte ideia: a categorização visual refinada é difícil e possui diferentes pontos fortes e fracos tanto para os seres humanos quanto para máquinas. Os seres humanos são capazes de detectar e categorizar objetos, mesmo quando eles não reconhecem completamente, ou não conhecem exatamente o objeto de interesse. Eles podem localizar formas geométricas, partes do objeto, ou reconhecer cores e materiais. Entretanto, podem surgir erros na identificação principalmente porque as pessoas possuem (1) experiências e memória e diferentes níveis; (2) diferente percepção e subjetividade. Em contraste, os computadores podem executar software determinista em grandes bases de dados. Eles se destacam pelo uso intensivo de memória, lidam bem com problemas, como reconhecer cartazes de filmes ou caixas de cereal, etc. Isto sugere que um sistema visual composto por um ser humano e uma máquina pode realizar a tarefa, e fazê-lo de forma eficiente e colaborativa, combinando os pontos fortes dos dois envolvidos.

[1]Versão do código foi disponibilizado publicamente por Andrea Vedaldi (Vedaldi e Fulkerson (2010)) e utiliza pirâmides espaciais, múltiplos *kernels* de aprendizagem e a abordagem 1 contra todos para o classificador SVM.

Os algoritmos podem localizar e classificar os objetos no conjunto de dados de 200 classes em uma fração de segundos, utilizando detectores que são compartilhados entre as classes. O trabalho formaliza um método para avaliar a utilidade de diferentes tipos de intervenção humana ou da máquina. Os usuários humanos fornecem dois tipos diferentes de informação, clicando em partes do objeto de interesse. Os algoritmos são capazes de organizar 312 questões binárias, 29 questões de múltipla escolha e com múltiplas respostas e 15 questões para clicar em partes do pássaro. A máquina seleciona perguntas que possam capturar informações relevantes a fim de identificar o objeto de interesse, o mais rapidamente possível. A ideia da combinação de recursos é utilizada na construção de um guia para a identificação de pássaros. Os resultados experimentais demonstram a força do sistema híbrido que oferece rapidez e precisão na identificação para o conjunto de dados CUB 200 completo, apontando uma taxa de correta classificação de 56,5%, operando de forma isolada, sem a ajuda de um usuário em tempo real. O artigo compara os resultados com vários outros trabalhos da literatura, alguns deles descritos neste trabalho, enfatiza que esse desempenho é menor do que o esperado para uma aplicação deste tipo, mas tal desempenho foi alcançado depois de três anos de pesquisa em algoritmos de reconhecimento de granularidade fina.

O trabalho de Branson et al. (2014) propõe várias alterações desde os trabalhos de Branson et al. (2010) e Wah et al. (2011). O primeiro trabalho a utilizar o conjunto de dados CUB 200 também introduziu um sistema que combinava a interação humana e a visão computacional para que um pudesse complementar as limitações do outro (Branson et al., 2010). O sistema é disponibilizado aos usuários de forma interativa através de um jogo com 20 perguntas simples referentes a atributos visuais. O objetivo é identificar a classe verdadeira, minimizando o número de perguntas e utilizando o conteúdo visual. Neste caso, o sistema de visão computacional é auxiliar no processo de categorização. As imagens e as questões são apresentadas ao usuário e depois de cada interação as probabilidades, associadas às classes na procura da classe verdadeira, são recalculadas e redistribuídas. Os resultados obtidos são discutidos, considerando as seguintes vantagens e desvantagens: a) as respostas dos usuários são estocásticas; b) a visão computacional consegue reduzir o trabalho manual; c) as respostas do usuário podem conduzir a melhores resultados; d) a visão computacional consegue melhorar o desempenho global do sistema; e) perguntas diferentes são feitas quando se usa visão computacional e quando ela não é utilizada; e f) taxa de reconhecimento nem sempre significa sucesso. O trabalho de Branson et al. (2010) foi ampliado na proposta de Wah et al. (2011).

Um dos principais problemas para tratar granularidade fina é apontado por Wah et al. (2011). Tradicionalmente, os conjuntos de imagens disponíveis possuem poucas clas-

ses e muitos exemplos associados. O desenvolvimento e abrangência de novas pesquisas levaram a utilizar menor número de exemplos de treinamento na tarefa de aprendizagem e a criação de algoritmos de visão computacional mais robustos, bem como novas metodologias para a coleta de dados e rotulações de exemplos. O trabalho investiga a ambiguidade dos conceitos e a percepção humana. Por exemplo, a percepção da localização precisa de uma determinada parte (tal como o bico de um pássaro) e pode variar de pessoa para pessoa, assim como a percepção de brilho em um objeto. Isso faz com que algumas classes sejam, geralmente, reconhecidas apenas por especialistas humanos.

A abordagem de avaliação utilizada pelo trabalho é uma métrica de esforço baseada no tempo necessário para que um humano possa classificar um objeto. O desempenho é determinado pelo cálculo da quantidade média de tempo necessário para a classificação correta de uma imagem de teste. O computador apresenta ao usuário imagens da classe mais provável. Isso pressupõe que o usuário pode validar a classe correta em todos os casos, entretanto, isso nem sempre pode ser possível. A precisão da classificação é verificada nos seguintes casos: (1) método integrado, combinando algoritmos de classificação/localização de partes, cliques do usuário e atributos binários relativos às questões; (2) utilização de apenas perguntas binárias, não localizados por algoritmos de visão computacional, ou ganho de informação esperado para selecionar perguntas representativas; (3) não utilizando visão computacional; e (4) selecionando perguntas de forma aleatória. Os experimentos realizados no conjunto de dados CUB 200 mostra que o sistema é preciso e rápido. O tempo médio para responder a uma questão sobre partes pode variar de 3,01 até 0,26 segundos, em comparação com outros atributos que podem variar de 5,38 até 7,64 segundos. Neste contexto, as perguntas referentes às partes do pássaro são mais prováveis de aparecerem antes e podem interferir no desempenho global do sistema. Deste modo, o sistema deve apresentar ao usuário questões combinando atributos da classe e as informações armazenadas referentes às partes do pássaro (bico, cauda, cabeça, etc.).

Deng et al. (2013) seguiram a ideia de combinar as habilidades humanas e da máquina propondo um jogo *on-line* chamado *Bubbles*, que revela quais características são discriminativas. O exemplo é desenvolvido na identificação das espécies de pássaros e a taxa de acerto para o conjunto de dados CUB 200 é 32%. A cada rodada do jogo, um jogador vê imagens de exemplo para duas espécies de pássaros. Em seguida, uma nova imagem é apresentada e então é solicitado que o jogador classifique a ave em uma das duas espécies. O jogador ganha pontos na correta identificação e perde pontos caso contrário. Independentemente do resultado, o jogo avança para a próxima rodada com uma nova imagem e, possivelmente, um novo par de espécies de pássaros é apresentado. Nas próximas fases do jogo a imagem vai ficando com menos detalhes de forma e o

jogador só pode ver um esboço do pássaro. O jogador pode, no entanto, clicar para revelar pequenas áreas circulares da imagem, denominadas bolhas, para verificar detalhes da parte selecionada, porém isso gera uma penalidade na pontuação. Através de uma configuração adequada de recompensa, o jogo pode garantir que bolhas selecionadas por um jogador bem-sucedido poderá conter recursos discriminantes para identificar a espécie do pássaro. Segundo os autores, o jogo possui as seguintes vantagens: (1) *Domínio agnóstico* - A única suposição é que os seres humanos podem descobrir as características visuais discriminantes para um conjunto de exemplos. Na verdade, aprender a dizer que as categorias são ou não familiares sob pressão de tempo cria desafios e diversão. (2) *Garantia automática de qualidade* - Se os jogadores ganham altas pontuações, sabemos com certeza que as áreas escolhidas devem ser importantes. (3) *Baixo custo* - O jogo oferece entretenimento e as pessoas vão se voluntariar para jogar. Isso pode permitir a coleta de dados em larga escala, com baixo ou custo zero. Outro fato pertinente ao método proposto é o relacionamento com estudos de visão humana. O jogo tem o nome de uma técnica de psicologia conhecida por estudar características que os seres humanos usam para o reconhecimento facial, mas a abordagem é diferente na medida em que as bolhas são escolhidas pelo jogador. Outro relacionamento com estudos de visão humana é que por meio do jogo é possível acompanhar o olhar do jogador, revelando as posições que o jogador olha atentamente.

3.3 Métodos que consideram segmentação

O principal desafio da classificação em problemas de granularidade fina sem dúvida é a similaridade entre as classes. Para identificar quais características visuais são relevantes em um determinado problema, um especialista humano requer conhecimento de domínio específico. No entanto, um sistema automático vai encontrar alguns desafios adicionais, por exemplo, considerando a resolução do problema por meio de imagens obtidas em ambientes naturais, com fundos ricos de detalhes e desafiadores, onde o fundo da imagem pode tornar-se um fator de complicação para o algoritmo de reconhecimento. Enquanto em alguns problemas o fundo pode ser útil, isto é, as folhas podem expressar informações do contexto em imagens de flores, para outros problemas, como a identificação de pássaros, o fundo da imagem é variável, podendo apenas ser compartilhado por uma pequena quantidade de imagens. Desta forma, a segmentação consiste em separar os *pixels*da imagem que fazem parte do pássaro, daqueles que representam o cenário (Chai et al. 2011), (Das e Manmatha 2001).

Muitos trabalhos compartilham a hipótese de que a aplicação de segmentação antes

do procedimento de extração de característica implicará em melhores taxas de acerto ao sistema de identificação, pois seria possível descartar informações irrelevantes e isolar informações de interesse. Outra vantagem de um algoritmo de detecção e segmentação é que ele pode localizar o objeto, o que será benéfico, especialmente se o objeto não estiver no centro da imagem, ou for de tamanho ou formato diferente. Outro aspecto é a possibilidade da extração de contornos do objeto de interesse, que pode proporcionar características discriminares para a identificação. Alguns trabalhos descritos a seguir utilizam e discutem a importância da segmentação.

O trabalho de Chai et al. (2011) apresenta um algoritmo escalável, baseado na alternância de co-segmentação (Bicos). A co-tarefa de segmentação é representada como *pixels* e distribuições de cores para imagens individuais, e super-pixels com características aprendidas do conjunto de imagens, juntamente com o algoritmo GrabCut e classificador SVM em cada nível. A tarefa de classificação é realizada por meio da abordagem um contra todos e do classificador SVM. A taxa de classificação varia de 6,7%, sem qualquer segmentação a 16,2% quando o algoritmo Bicos é utilizado. Além disso, são apresentados resultados de 13,6% e 23,3%, utilizando o GrabCut (Rother, Kolmogorov, e Blake, 2004) e considerando a segmentação disponível juntamente com o conjunto CUB 200. O desempenho da segmentação e da classificação no conjunto de dados CUB 200 pode ser comparada com os resultados do trabalho de Chai et al. (2012) que apresenta várias melhorias e propõe o algoritmo TriCos (algoritmo de co-segmentação discriminativo de classes em três níveis para classificação de imagens). A taxa de classificação varia de 16,1%, com aplicações de diferentes versões do algoritmo BiCos ou aplicação do algoritmo proposto o TriCos que apresenta 25,5%. Além disso, são apresentados resultados 26,7% obtidos com a utilização do *ground truth* e 22,7% utilizando apenas o recorte da caixa delimitadora.

O artigo de Moghimi (2011) investiga o uso da cor para o reconhecimento de pássaros. São criados histogramas de cor de diferentes regiões de interesse em diferentes espaços de cores. Região de interesse ou ROI é definida como uma área de onde são extraídas as informações referentes às cores da imagem (neste caso, a partir de um histograma). A representação em um espaço de cor facilita a especificação de cores de acordo com um determinado padrão. Para a realização dos experimentos foram utilizados os espaços, RGB, HSV, YIQ, YCrCb, XYZ, LAB e LUV. Moghimi (2011) enfatiza que a segmentação automática não conseguiu bons resultados. Os resultados para a segmentação automática são semelhantes aos obtidos quando utilizada segmentação em partes da imagem, dependendo da quantidade de informação, tais como a segmentação completa, partes localizadas, ou apenas um ponto aleatório sobre o objeto. Outro fator que pode

influenciar no desempenho do classificador são os espaços de cores utilizados. Todos os resultados são inferiores a 18,9% obtidos através do espaço de cor YIQ.

O trabalho de Angelova e Zhu (2013) propõe um algoritmo eficiente para a detecção de objetos e segmentação que funciona para uma variedade de problemas de classificação do tipo de granularidade fina. A abordagem é baseada na identificação de regiões específicas do objeto de interesse, no momento da detecção. A ideia geral é criar detecções rudimentares, baseadas em recursos para a superclasse de objetos, por exemplo, pássaros. Estas detecções são bons indicadores da presença do objeto e podem ajudar a apontar para a possível localização do objeto. O método proposto apresenta melhores resultados quando comparado a abordagens anteriores que utilizam *ground truth* e caixa delimitadora. Outro fato importante é que a melhoria quando a caixa delimitadora é utilizada é maior do que quando não é utilizada. Essa melhora é verificada em todos os outros conjuntos de dados testados no trabalho, sendo atribuída ao fato de que a localização do objeto é facilitada.

O artigo de Chai et al. (2013) demostra que a precisão da classificação em problemas de granularidade fina pode ser melhorada quando a localização de partes e a segmentação são realizadas juntas, de modo que o resultado de ambos os processos pode ajudar um ao outro. A simbiose resultante alcança melhor desempenho quando comparado com a classificação obtida por uma simples concatenação de duas características. Apesar da sinergia entre a segmentação e a classificação já ter sido reconhecida em outros trabalhos, a abordagem em problemas de granularidade fina é capaz de atingir melhorias na taxa de acerto da identificação de espécies de pássaros no conjunto de dados Caltech CUB 200-2011. Outros trabalhos que seguem a abordagem de definição de partes para resolver problemas de granularidade fina são descritos na próxima seção.

3.4 Métodos que consideram partes

Algumas abordagens mais eficientes para o reconhecimento visual em problemas de granularidade fina são baseadas na detecção e extração de características de partes específicas de um objeto. Por exemplo, na classificação da raça de um cão podem-se extrair características do nariz e das orelhas. O reconhecimento facial é um caso muito claro de reconhecimento visual e de granularidade fina em que as subcategorias são casos individuais, e os melhores métodos de reconhecimento de faces utilizam a extração de características locais, tais como os cantos dos olhos (Berg e Belhumeur, 2013).

O trabalho de Berg e Belhumeur (2013) aponta resultados importantes para essa abordagem por meio de uma estrutura de aprendizagem para um grande conjunto de

características discriminativas de nível intermediário especializados para um determinado domínio de conjunto de partes identificado como uma parte versus uma característica (do inglês: *Part-based One-vs-One Features* - POOFs). Um conjunto de dados de imagens no domínio de pássaros é rotulado por classes e com localização de partes. Para qualquer par de classes ou para qualquer par de partes é possível extrair algumas características de baixo nível em uma grade de células que cobre as duas partes, e treinar um classificador linear para distinguir as duas classes entre si. Histogramas de gradiente ou cor são utilizados como características de baixo nível. Os pesos atribuídos por este classificador a diferentes células da grade indicam a região mais discriminativa em torno destas partes para o par de classes. A região de suporte é fixada com base nesses pesos. Depois é possível voltar a treinar o classificador para encontrar uma projeção discriminativa. O método proposto é totalmente automático para a construção de uma biblioteca (POOFS) de recursos de nível intermediário, discriminativamente treinado a partir de um conjunto de imagens com rótulos de classe e localização de partes. Uma das grandes vantagens é a demonstração de que o método consegue reduzir a necessidade de grandes conjuntos de treinamento e pode ser utilizado em outros domínios. Trata-se de uma significativa contribuição ao Estado da Arte, apontando uma taxa de classificação de 56,8% no conjunto de dados Caltech-UCSD CUB 200.

Várias abordagens têm sido propostas para descrever e localizar partes em objetos em problemas de granularidade fina. A normalização de poses foi discutida por Zhang et al. (2012). Essa abordagem procura ser invariante à postura, articulação e ângulo de visão de uma imagem obtida por uma câmera, visando obter a localização semântica de partes de objetos para a extração de características de aparência, no que diz respeito à localização de partes. Nestas abordagens, para a localização de partes foi utilizado um conjunto de filtros computacionalmente caros conhecidos como *Poselets*. O trabalho explora a questão de como esses conjuntos de características podem detectar comparações entre duas imagens. Os autores desenvolvem funções de similaridade que levam à ativação de *Poselets* que geram as características submetidas ao classificador. O método não exige representação em três dimensões e explora a ideia de que a normalização de poses é relevante para descritores de aparência. A implementação é baseada nas ferramentas disponibilizadas pela biblioteca VLFEAT (Vedaldi e Fulkerson 2010) que utiliza variações do algoritmo SIFT e classificador SVM com *kernel* linear para definição de sistema de referência. Os experimento, utilizam BOW-SIFT(*bag of words* em características extraídas pelo algoritmo SIFT) e PHOW (*pyramidal histogram of words*) como características. Os resultados obtidos pelo método são de 28,18%, superando o sistema de referência de 18,60% para o conjunto CUB 200-2011.

Em trabalho subsequente, Zhang e Farrell (2013) apresentam um novo método de descritores de partes que apresenta 51% de taxa de acerto para o mesmo conjunto de dados. O novo descritor de parte deformável (DPD) se apresenta como uma opção robusta e eficiente para descrever e normalizar posturas em um método de partes deformáveis (DPM). O artigo propõe dois descritores de partes: a) DPD forte que aproveita informações semânticas inerentes a partes de DPM que são fortemente controlados; e b) DPD-fraco que utiliza anotações semânticas para aprender correspondências cruzada de componentes de partes. O descritor DPD-forte atinge 50,05% de taxa de acerto enquanto o DPD-fraco atinge 50,98% sobre o mesmo conjunto de dados. Depois de prever as regiões de partes via DPM, são utilizados descritores do *kernel* para gerar vetores de características e posterior classificação. Especificamente, são utilizados quatro tipos de descritores de *kernel*: gradiente, LBP, de cor RGB, e cor RGB normalizado. São calculados descritores do *kernel* locais de tamanho 16 x 16 em relação a uma imagem, por meio de uma grade regular densa de tamanho 8. Em seguida, é aplicado, uma pirâmide espacial no topo. A quantização vetorial destes descritores utiliza um vocabulário de 1000 elementos, concatenando regiões do histograma em um único vetor para cada imagem. Estes vetores são fornecidos como entrada para um SVM linear. Uma importante contribuição deste trabalho é a indicação de qual parte é possível obter a informação mais discriminante. Por exemplo, localizar a cabeça do pássaro é especialmente importante, pois, quando as informações geradas por esta parte são retiradas do vetor de características a taxa de acerto do DPD fraco cai para 43,15%.

Diferente de trabalhos anteriores que utilizam partes, o trabalho de Gavves et al. (2013) propõe, em sua abordagem, não identificar as partes individuais, mas ao invés disso localizar detalhes distintivos de uma parte presente em uma imagem com uma parte presente em outra imagem, para que seja possível realizar alinhamento entre os objetos de interesse. Este alinhamento é aproximado e insensível às grandes variações de aparência para grande número de subcategorias. Além disso, o alinhamento aproximando não é somente de uma subcategoria específica, assim, a representação do objeto torna-se independente do número de classes ou imagens de treinamento. A primeira novidade deste trabalho baseia-se na observação de que todas as subcategorias pertencentes à mesma supercategoria compartilham características globais em relação à sua forma e posicionamento. A segunda novidade é baseada na observação de que a partir de alinhamentos irregulares, em vez da localização precisa das partes, alguns conflitos de aparência devem surgir mesmo entre pássaros muito semelhantes, devido às condições de aquisição da imagem, tais como pequenas translações, variações de ponto de vista e oclusões parciais. O método proposto indica que a utilização de vetores de Fisher consegue descrever melhor

o conteúdo visual de uma imagem do que HOG para problemas de granularidade fina. Os resultados são reportados a partir de dois conjuntos de dados muito conhecidos para tratar problemas de granularidade fina, a Caltech CUB 200 2011 e Stanford Dogs. Até o momento apresenta-se na literatura como o melhor desempenho obtido para o primeiro conjunto 62,7% com a aplicação de alinhamentos não supervisionados. Uma comparação entre alinhamentos de forma supervisionada e não supervisionada foi realizada com uma distribuição de classes parecida e o alinhamento de forma não supervisionada apresentou o melhor desempenho em termos de taxa de acerto e não utilizou o *ground truth* nem anotações de partes.

3.5 Métodos que consideram áudio

Independente da forma como é executada a tarefa de identificação de espécies de pássaros é altamente complexa e desafiadora. Recentemente, a bioacústica animal tem recebido atenção crescente devido a seus diversos benefícios potenciais para a ciência e a sociedade. Duas importantes iniciativas que pretendem resolver esse problema merecem destaque. Primeiro, em paralelo a *30th International Conference on Machine Learning* realizada em Atlanta, USA, em Junho de 2013 foi realizado o *Workshop on Machine Learning for Bioacoustics* - ICML4B (Glotin et al., 2013). Segundo, o laboratório LifeCLEF que propõe atividades para identificação de plantas e animais e organizou nos últimos três anos, com um número crescente de participantes, avaliar identificação de espécies de plantas, peixes e pássaros (Goëau et al., 2014).

Glotin et al. (2013) apontam que o aumento das expectativas de pesquisa bioacústica foram coincidentes com um aumento da coleta de dados acústicos. O gargalo neste momento não é o acesso aos dados brutos, mas a incapacidade de processar de forma eficiente, visualizar e interpretar grandes volumes de dados dentro de um avançado sistema de gerenciamento de dados. Entretanto, a falta de padronização, combinado, com os diferentes ambientes e contextos de coleta de dados em grande escala, cria uma adaptação de domínio e estrutura de aprendizado singular. O desafio para *Identificar espécies de pássaros por meio de gravações de áudio* considerou 35 espécies, representadas por gravações contínuas de áudio, obtidas de três locais diferentes. Os dados utilizados foram fornecidos pelo *Museu Nacional d'Histoire Naturelle*, uma das instituições de pesquisa de pássaros mais respeitadas no mundo. A métrica de avaliação adotada foi a curva ROC e o conjunto de características indicado o MFCC (*Mel-Frequency Cepstral Coefficients*). Alguns detalhes do conjunto de treinamento: áudio captado por um transdutor analógico de 16 bits, com frequência da amostragem de 44,1 kHz; 35 gravações, uma espécie por arquivo,

30 segundos por arquivo, totalizando 18 minutos de gravações. Cada uma das 35 espécies aparece, pelo menos uma vez, no conjunto de teste. Com relação ao conjunto de teste: os dados foram coletados por 3 microfones na mesma área, em 3 condições de ambientes (detalhados em documentos que acompanham os dados), transdutor analógico de 16 bits, frequência de 44,1 kHz. Também foram disponibilizado *ground truth* e metadados das 90 amostras. Os resultados obtidos com a participação de 77 equipes, variaram de 0,43663 até 0,83829 para a equipe vencedora.

Goëau et al. (2014) responsáveis pelo desafio LifeCLEF, justificam o uso de áudio ao invés de imagens pelo fato dos pássaros não serem facilmente fotografados, estarem na maioria das vezes escondidos no cenário, por exemplo, no alto de uma árvore, voarem rapidamente, assustadas com a presença humana, e as gravações de áudio serem mais fáceis de coletar. A identificação de espécies de pássaros no desafio LifeCLEF 2014 utilizou conjuntos de treinamento compostos por gravações de áudio, por exemplo, áudios de 501 espécies tipicamente brasileiras, obtidos por meio do website Xeno-Canto. O conjunto de dados tem entre 15 e 91 gravações por espécie, totalizando 14027 amostras de áudio. Para evitar qualquer viés na avaliação relacionada com os dispositivos de áudio utilizada, cada arquivo de áudio foi normalizado para frequência de 44,1 kHz e codificado com mais de 16 bits. As características extraídas como padrão foram o MFCC (essas configurações fazem parte dos melhores resultados obtidos no desafio ICML4B) e uma grande variedade de metadados: Informações taxonômicas de espécie, gênero, família, autor, data do registro, hora, qualidade da gravação de áudio, localidade, latitude, longitude, entre outros. Com um total de dez grupos de sete países, um total de vinte e nove rodadas, envolvendo métodos clássicos e originais. A Tabela 3.1 apresenta o resumo dos resultados obtidos por todas as execuções por meio da métrica média das precisões médias - *MAP-Mean Average Precision*.

O melhor desempenho foi obtido pela equipe *MNB TSA*. As características escolhidas são relacionadas a probabilidades e funções estatísticas de segmentos de áudio. Além disso, são combinados metadados do aúdio, como por exemplo: mês, ano, localidade e autor foram submetidos à classificação por meio de *randomized decision trees*. Em seguida, a abordagem baseada exclusivamente em áudio de equipe *QMUL*, indica que a aprendizagem de características não supervisionadas é um método simples e eficaz para aumentar o desempenho classificação quando o conjunto de dados é grande. Isso também pode ser verificado pelos resultados obtidos pela equipe *Inria ZENITH* que usou classificação baseada em instâncias, e características obtidas por MFCC. A equipe *Utrecht Univ.* que utilizou a abordagem profunda exclusivamente em relação a variações de MFCC não apresentou resultados competitivos e participou de apenas uma das rodadas.

Tabela 3.1: Resumo dos métodos, características e resultados para melhor atuação das equipes participantes.

Equipe	Características	MAP 1	MAP 2
BiRdSPec	Time Zero Crossings, Spectral Centroid, MFCC, Flux, Roll	0,119	0,144
Golem	HOG, LBP	0,105	0,129
HTL	MFCC, tempo médio, espectrogramas	0,289	0,272
Inria Zenith	MFCC	0,317	0,365
MNB TSA	Aceleração, probabilidades e funções estatísticas	0,453	0,511
QMUL	Características não supervisionadas em duas escalas de tempo	0,355	0,429
Utrecht Univ.	Variações de MFCC	0,123	0,140

Uma constatação importante é o uso de uma abordagem hierárquica, baseada na taxonomia das espécies, proposta pela equipe brasileira. Os resultados não foram expressivos, porém os organizadores destacam essa iniciativa e enfatizam que a métrica de avaliação utilizada não é indicada para esse método. Outra, refere-se à lentidão do processo. Segundo os organizadores, mesmo o método de melhor desempenho conseguiu utilizar apenas 96,8% dos dados de teste como previsto. Para que as equipes conseguissem respeitar os prazos, o conjunto de teste completo teve que ser abordado por soluções alternativas, mais rápidas, para que todas as gravações fossem identificadas.

A literatura aponta que a identificação de espécies de pássaros baseada em características de áudio vem apresentando resultados bastante interessantes. O problema da identificação automática de espécies de pássaros, por áudio, foi abordado por Lopes et al. (2011). Os experimentos para encontrar a espécie de um pássaro específico por meio da gravação de seu canto foram limitados somente a três espécies de pássaros: *Taraba major* (32 amostras), *Cercomacra tyrannina* (34 amostras) e *Thamnophilus doliatus* (35 amostras). Os cantos foram representados por três conjuntos de características extraídas por meio do Marsyas (o vetor de características final possui 64 dimensões referentes conteúdo rítmico, textura timbral e características relacionadas ao tom), IOIHC (o conjunto de características final e representado por um vetor de características de 40 dimensões de periodicidade rítmica) e Sound Ruler (36 características de um sinal de áudio pulsado). Os melhores resultados de classificação considerando validação cruzada com 10 partições, foram obtidos pelas características geradas por meio do IOIHC e Marsyas. O melhor caso ocorreu quando o conjunto IOIHC foi aplicado ao classificador Naive Bayes, apresentando 99,7% de taxa de acerto.

Em trabalho subsequente Lopes et al. (2011) utilizam dois conjuntos de dados. A

primeira, com gravações de áudio completas de 73 espécies. A segunda, derivada da primeira, obtida das gravações originais, divididas de acordo com pulsos presentes no áudio. Um pulso é definido como um intervalo de som curto com grande amplitude. Os conjuntos de características foram obtidos pelo Marsyas por meio do cálculo de médias e variâncias das características de timbre, obtidas nos intervalos, para *spectral centroid, rolloff, flux, time-domain, zero crossings* e ainda 12 MFCCs de cada caso; o conjunto completo inclui 64 características. Os experimentos realizados utilizaram diferentes classificadores para tentar identificar o mais adequado para o problema. Os classificadores utilizados foram: Naive Bayes; kNN ($k\bar{3}$); árvore de decisão (J.48); uma rede neural MLP treinada (*back-propagation e algorithm momentum*) e o classificador SVM com diferentes funções de *kernel*. A divisão dos conjuntos de dados adotada foi a validação cruzada de 5 partes.

Os resultados mostram que a utilização de pulsos conseguiu melhorar o desempenho da classificação. A explicação para esse fato é que a informação acústica presente nos pulsos contém menos ruído ambiental e incorpora as características mais importantes do canto do pássaro. Os experimentos realizados consideram diferente número de classes, sendo: 3, 5, 8, 12e 20 classes. Os melhores resultados usando pulsos foram obtidos com classificador MLP e SMO: para 3 classes, o *F-measure* obtida foi de 95,1% para as classes mais frequentes, usando SMO-Pearson, e 96,4% para classes selecionadas aleatoriamente, usando MLP; para 5 classes os valores *F-measure* correspondentes foram de 73,2% (com SMO-Pearson) e 83,1(com MLP); finalmente, por 8 classes os valores foram de 85,7% (com SMO-Pearson) e 89,7% (com SMO-Pearson). Esse classificador também apresentou os melhores resultados para os experimentos com 12 e 20 espécies (82,9% e 78,2%). Em trabalhos futuros, os autores pretendem aplicar as técnicas de classificação hierárquica e definir uma ontologia de espécies de pássaros para melhorar a capacidade de identificação do método.

O trabalho de Stowell e Plumbley (2014) considera a classificação automática de espécies de pássaros em larga escala e a aprendizagem de características. O artigo apresenta uma técnica para a aprendizagem de características a partir de grandes volumes de gravações de áudios de pássaro, inspiradas por técnicas que têm mostrado bom desempenho em outros domínios. São comparadas experimentalmente doze representações derivadas de *Mel-frequency cepstral coefficient* - MFCC. São utilizados quatro bancos de dados de larga escala com vocalizações diversificadas de pássaros, e um classificador *Random Forest*.

Os conjuntos de dados utilizados são descritos como de amostras com rótulos simples ou amostras de múltiplos rótulos. Por amostras de rótulos simples, entende-se que existe apenas uma espécie presente em uma gravação, portanto uma saída para o classificador utilizado. Por amostras de múltiplos rótulos, entende-se que qualquer espécie

pertencente ao conjunto de classes poderá estar presente na gravação. Desta forma as seguintes abordagens podem ser aplicadas: a) *classificação binária de múltiplos rótulos:* corresponde a divisão da tarefa de classificação de múltiplos rótulos em tarefas binárias de rótulo simples. Desta forma, um classificador separado para cada um dos rótulos e utilizado; b) *classificação total de múltiplos rótulos:* corresponde a adoção de um único classificador sendo treinado para fazer previsões de presença e ausência de todos os rótulos simultaneamente. Os conjuntos de dados foram: a) *nips4b*: pássaros da França, 87 espécies e 687 amostras com múltiplos rótulos. b) *xccoverbl*: pássaros do Reino Unido e Europa, 88 espécies e 264 amostras com rótulo simples. c) *bldawn*: pássaros do Reino Unido, 77 espécies e 60 amostras com múltiplos rótulos. d) *lifeclef2014*: pássaros do Brasil, 501 espécies e 9688 amostras com rótulo simples.

Os experimentos demonstram que as características obtidas por MFCC são limitadas, considerando conjuntos de larga escala. Por outro lado, demonstra que a aprendizagem de características de maneira não supervisionada fornece uma melhora em relação aos resultados obtidos anteriormente com as características MFCC, sem aumentar a complexidade computacional depois que o modelo foi treinado. O aumento foi maior para a classificação que considera um único rótulo para amostras em grande escala. Muitos experimentos foram realizados, mas outros estudos da interação entre as características do conjunto de dados e escolha da representação característica ainda necessitam de outras análises. A aprendizagem de características não supervisionadas requer grandes volumes de dados para ser eficaz. Para o maior conjunto de dados - *lifeclef2014* - a aprendizagem de características elevou o desempenho do classificador para 85,4% (métrica AUC). Sem a aprendizagem o resultado foi de 82,2% para *Mel spectra* e 69,3% para MFCCs. Essa diferença foi percebida nos outros conjuntos, com execução apenas para o conjunto *bldawn*. Os resultados obtidos foram comparados aos resultados do desafio LifeCLEF 2014. Neste caso, os resultados da abordagem proposta são melhores. Por exemplo, no desafio a média das precisões médias (MAP) máximo quando considerado somente áudio foi de 42,9%. Apenas uma equipe superou a abordagem proposta atingindo 51,1% quando utilizados além do áudio metadados adicionais. Devido aos resultados experimentais, os autores não recomendam as características MFCCs para o reconhecimento de espécies de pássaros, embora essa prática tenha sido amplamente difundida.

3.6 Outros trabalhos relacionados

O trabalho de Yao et al. (2012) apresenta uma estratégia de classificação de imagens de granularidade fina, ou seja, classificação de objetos da mesma categoria, por

exemplo, diferentes espécies de pássaros que possuem similaridade visual. O método proposto é inovador e não é dependente de anotações humanas ou da utilização de *codebook* para referência. Os experimentos são realizados com o conjunto de dados de pássaros Caltech-UCSD (CUB 200). A partir do conjunto de imagens de treinamento são gerados, aleatoriamente, um grande número de *templates* na imagem (verificados pelas regiões da imagem identificadas por retângulos vermelhos). A imagem é selecionada se compartilhar as cores com os mapas. Uma imagem será representada por agrupamento de mapas de resposta que foram obtidos a partir da combinação de cores da imagem com um grande conjunto de *templates* gerados aleatoriamente. Um algoritmo de *bagging* é usado para a realização de combinação de classificadores SVM. São obtidos resultados entre 39,76% a 44,73%, desempenhos superiores aos apresentados na literatura se comparados com outros métodos que utilizam anotações humanas (37,12%) ou *codebook* como referência (40,25%).

O trabalho de Bo et al. (2013) propõe aprender características discriminantes diretamente das imagens, abandonando o paradigma da aprendizagem superficial, movendo-se para o paradigma da aprendizagem profunda. A abordagem combina codificação esparsa *Multipath Sparse Coding* por meio múltiplas camadas, utilizando vários fragmentos da imagem de tamanho variável *Hierarchical Matching Pursuit Liefeng.* A codificação de cada fragmento é obtida por meio de diferentes caminhos com um número variável de camadas. O estudo apresenta comparações em três tipos de tarefas de reconhecimento visual: reconhecimento de objetos, reconhecimento de cenas e reconhecimento de objetos de granulação fina (conjunto CUB 200 2011). Os resultados obtidos com conjunto de 200 classes, utilizando o recorte da caixa delimitadora e a divisão de 15 imagens de cada uma das espécies de treinamento e o restante para o teste, apontam 30,3% de taxa de acerto. De forma geral, os resultados são extremamente encorajadores, indicando que os sistemas de reconhecimento visual podem ser significativamente melhorados com a utilização de aprendizagem a partir de imagens, sem etapas de pré-processamento.

O trabalho de Zhang et al. (2014) explora a localização de partes para isolar explicitamente as diferenças de aparência sutis associadas as partes do objeto de interesse e à aprendizagem de características por meio da abordagem convolucional. De forma breve, o método proposto aprende detectores e modelos de partes de todas as regiões da imagem, impõe restrições geométricas entre as partes e a caixa delimitadora do modelo aprendido. Neste caso, as características foram obtidas por meio do modelo convolucional ajustado e informação semântica de partes. O *ground truth* da caixa delimitadora para o pássaro é utilizado não apenas na fase de treinamento, mas também na fase de teste. A precisão de classificação sem ajuste fino obtém 68,1% de taxa de correta classificação. Com o ajuste fino o resultado obtido é de 76%. Os resultados experimentais em relação

ao conjunto de dados CUB 200 2011 indicam que o método proposto supera os demais métodos do Estado da Arte, indicando 76,37% de taxa de correta classificação.

3.7 Abordagens anteriores

Trabalhos que abordaram a classificação de espécies de pássaros por meio de imagens antes do surgimento do conjunto CUB 200, também foram analisados. Inicialmente, as abordagens consideram poucas imagens e técnicas de processamento de imagens para reconhecimento de pássaros.

Uma comparação de processamento de imagens foi apresentada por Nadimpalli et al. (2006). Algumas técnicas de processamento de imagem e uma série de operações morfológicas, *template matching* e variações de redes neurais artificiais foram implementadas para o reconhecimento de duas espécies de pássaros (os conjuntos de dados utilizados são de: 30, 60 e 75 imagens, respectivamente). As imagens são organizadas em três níveis considerando o nível de dificuldade de reconhecimento humano: (a) tipo 1: facilmente reconhecível; aparência similar; (b) tipo 2: um grau adicional de complexidade é adicionado com diferentes fundos e aves sem qualquer similaridade; (c) tipo 3: imagens ilegíveis devido às condições do ambiente e do equipamento utilizado para aquisição das imagens.

A implementação foi realizada em diferentes dois espaços de cores: HSV e RGB. Com a utilização de morfologia em níveis de cinza foram obtidos os seguintes desempenhos: tipo 1 = 96,5%; tipo 2 = 93,7%; tipo 3 = 80,9%; com alteração para imagens coloridas no espaço de cor RGB os resultados foram: tipo 1 = 86,3%; tipo 2 = 85,4%; tipo 3 = 80,9% e finalmente para o espaço de cor HSV: tipo 1 = 97,7%; tipo 2 = 97,9%; tipo 3 = 80,9%. Os resultados obtidos através de redes neurais artificiais foram os mais atrativos: para o tipo 1 = 100%; tipo 2 = 60%; tipo 3 = 50%. Utilizando *template matching* os resultados obtidos foram: tipo 1 = 90%; tipo 2 = 90%; tipo 3 = 40%.

O trabalho de Liu et al. (2007) indica que algumas pesquisas mostraram que o conhecimento de domínio específico pode ajudar na compreensão do conteúdo da imagem. A pesquisa em classificação baseada em forma é geralmente fundamentada por vários recursos visuais. O uso de ontologias para a representação explícita do domínio de conhecimento é sugerido, pois, nem sempre é considerado na recuperação de imagens ou na classificação. Os conceitos e as relações contidos em uma ontologia podem ser usados para descrever o conhecimento do domínio de raciocínio (Liu et al. 2007).

O artigo de Liu et al. (2007) propõe a construção de uma ontologia de forma automática, a partir de um conjunto de objetos, utilizando uma abordagem de agrupamento, visto que, a construção manual de uma ontologia é uma tarefa demorada. Os resultados

obtidos na geração de agrupamentos de 105 imagens de pássaros (quatro conjuntos: *duck, penguin, treecreeper, fairy-wren*, respectivamente: 25,28,28 e 24 imagens). As métricas *precision* e *recall* e F1 são informadas para cada conjunto e possuem variações de 45,5% até 87%. Na mesma direção, artigo proposto por Wang et al. (2010) propõe converter as imagens originais do formato RGB para o espaço HMMD (Hue, Max (nível de preto), Min (nível de branco), Diff (nível de cinza)) que é o espaço de cor padrão do formato MPEG-7. O descritor de estrutura de cor (CSD) do MPEG-7 é semelhante ao histograma de cores tradicional, mas é capaz de distinguir a estrutura espacial, o que não pode ser feito utilizando um histograma de cor. O trabalho propõe um algoritmo para encontrar formas dos pássaros: cauda, cabeça, bico, asa, entre outras e o operador Laplaciano para detectar as bordas da imagem. Os resultados experimentais mostram uma taxa de 84% de acerto quando as duas características são agrupadas. Os resultados para as características isoladas apontam 73,46% de taxa de acerto para cor e 66,31% para forma. Tais resultados são obtidos por meio de um sistema de consulta de 1726 imagens de pássaros de Taiwan.

Para finalizar, uma pesquisa sobre um mecanismo automático para o reconhecimento de espécies de pássaros com base em imagens coloridas é realizada por (Lin e Chen 2011). O conjunto de 120 imagens no formato JPG são submetidas a um pré-processamento para compensação de luz e remoção de fundo, em seguida são extraídas características (cor, forma e textura). A classificação é realizada a partir de árvores de decisão. Para cada uma das características é utilizado um classificador específico: (1) para características extraídas por um histograma de cores; (2) para características estatísticas extraídas a partir do histograma de cores: média, variância; (3) para características extraídas de formas a partir da densidade de um retângulo delimitador atribuído a imagem do pássaro; (4) para características extraídas de texturas por meio de Gabor e *Daubechies wavelet*; um classificador SVM é utilizado para finalizar a análise de resultados.

Os resultados da classificação e reconhecimento indicados por Lin e Chen (2011) apresentam uma taxa de acerto de 71,67% quando utilizadas características de cores, apenas um valor de medidas estatísticas de cores, forma e texturas e 73,33% quando utilizados os dois valores de medidas estatísticas de cores. O trabalho de Lin e Chen (2011) teve continuidade e reporta em Lin e Guo (2012) um sistema de monitoramento que poderá ser aplicado em casos de necessidade de monitoramento de epidemias.

De forma geral, os trabalhos descritos nesta seção apontam soluções que, apesar dos resultados mostrarem uma boa taxa de acerto, eram muito específicas e bastante restritas. Entretanto, alguns indicam abordagens que são adotadas em trabalhos recentes, como a identificação de formas, ou partes do objeto de interesse na imagem.

3.8 Análise dos trabalhos correlatos

Os principais desafios para a classificação em problemas de granularidade fina, por meio de identificação em imagens são: a quantidade de classes, a similaridade entre as classes (em geral diferenças sutis), a grande variabilidade intraclasse e as condições inerentes à maneira da aquisição da imagem. Neste caso, para identificar quais características são relevantes o especialista humano necessita de conhecimento de domínio específico. No entanto, um sistema automático vai encontrar alguns desafios adicionais. Infelizmente, em alguns casos o desempenho das técnicas disponíveis no Estado da Arte não é bem compreendido, necessitando de estudos aprofundados devido ao elevado grau de complexidade associado.

De maneira geral, os resultados recentes estão longe de atingir as exigências do mundo real em termos de ferramenta automática ou semi-automatizada aplicada a identificação de espécies. A maioria dos trabalhos ou ferramentas disponíveis conseguem identificar algumas dezenas ou centenas de espécies, apenas uma pequena parcela se comparada à totalidade das espécies catalogadas[2]. Além disso, a taxa de acerto, para abordagens baseadas em imagens, ou até mesmo para abordagens baseadas em áudio, na maioria dos casos é moderada, sendo necessárias novas alternativas que possam melhorá-las.

Comparando a identificação de pássaros por áudio com a identificação por imagens observa-se que os recursos visuais ainda não são bem explorados para a classificação de pássaros. Além dos problemas já mencionados, as propriedades visuais, por exemplo, cor, forma, partes, entre outras, são importantes para o reconhecimento de espécies e podem ser melhor exploradas, dada à complexidade do problema e à grande similaridade entre as espécies.

Esta tese aborda um problema de granularidade fina por meio da identificação automática de espécies de pássaros em imagens, e utiliza como ponto de partida os conjuntos de dados CUB 200 e CUB 200-2011. As Tabelas 3.2 e 3.3 organizam os principais trabalhos que envolvem esses conjuntos de dados de acordo com o ano de publicação e um breve resumo das principais características e classificadores utilizados para obter as taxas de acerto.

As Figuras 3.1 e 3.2 apresentam a evolução de desempenho medido pela taxa de acerto relacionada aos principais trabalhos. Neste caso, observa-se uma tendência de crescimento se considerarmos os trabalhos produzidos no último ano. Entretanto, muitos resultados são similares, mas abordam diferentes custos de implementação que podem ser avaliados pelas técnicas descritas nas Figuras 3.1 e 3.2.

[2]Atualmente, 9040 espécies estão catalogadas no website Xeno-Canto.

Tabela 3.2: Resumo de alguns trabalhos que utilizam o conjunto de dados CUB 200.

Referência	Características e Classificadores	Taxa de Acerto
Welinder et al. (2010)	Tamanho da imagem + KNN	0,6%
Welinder et al. (2010)	Histograma de cores + KNN	1,7%
Branson et al. (2010)	SIFT + *Spatial Pyramids* + SVM	19%
Chai et al. (2011)	SIFT + sem segmentação + SVM	6,7%
Chai et al. (2011)	SIFT + BiCos-MT + segmentação + SVM	16,2%
Chai et al. (2011)	SIFT + *Ground Truth* + segmentação + SVM	23,3%
Moghimi (2011)	Cor + segmentação + KNN	18,9%
Chai et al. (2012)	SIFT + TriCos e segmentação + SVM	25,5%
Chai et al. (2012)	SIFT + *Ground Truth* e segmentação + SVM	26,7%
Yao et al. (2012)	Codebook free + SVM	39,7%
Deng et al. (2013)	*Bubbles Game* + *Crowdsourcing* + SVM	32,8%
Angelova e Zhu (2013)	Detecção de objetos e segmentação + SVM	17,5%
Berg e Belhumeur (2013)	POOFs + SVM	56,8%
Bo et al. (2013)	Sparse Coding	30,3%

Tabela 3.3: Resumo de alguns trabalhos que utilizam o conjunto de dados CUB 200 2011.

Referência	Características e Classificadores	Taxa de Acerto
Wah et al. (2011)	Localização de partes + SVM	10,3%
Zhang et al. (2012)	*Poselet* + SVM	28,2%
Zhang e Farrell (2013)	Descritores de partes + SVM	51,0%
Chai et al. (2013)	Segmentação + Partes + SVM	59,4%
Gavves et al. (2013)	FGVC + Alinhamentos + SVM	62,7%
Branson et al. (2014)	Localização de partes + Atributos + SVM	56,5%
Zhang et al. (2014)	Partes + Convolução + SVM	76,37%

Cabe enfatizar o aspecto temporal relacionado aos trabalhos descritos neste capítulo. Desde o ano de 2010 até o presente momento, muitos trabalhos são executados em paralelo. Isso prova a recenticidade do tema e o grande interesse da comunidade científica em buscar soluções para problemas de granularidade fina. Desta maneira, acompanhar o Estado da Arte não é uma tarefa fácil. Entretanto, muitas contribuições podem ser apresentadas para uma área muito ativa e ainda com inúmeros problemas em aberto.

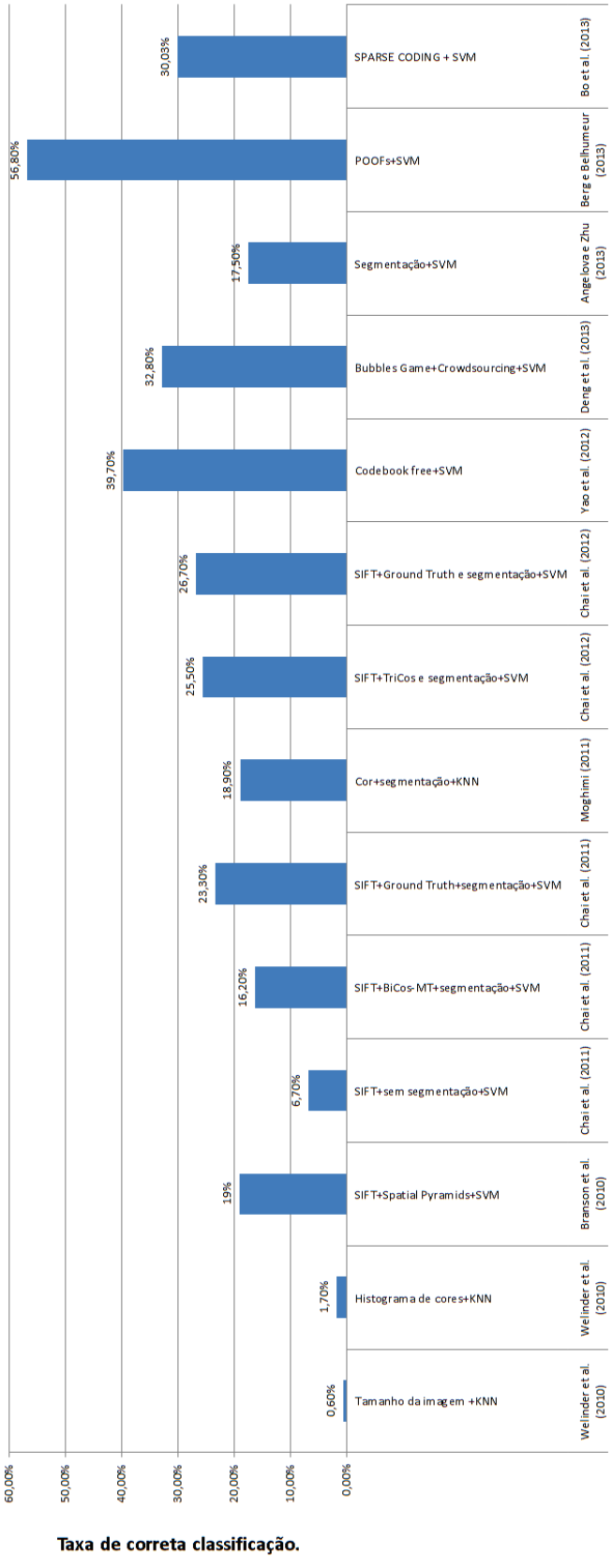


Figura 3.1: Evolução temporal (2010 - 2013) dos resultados para o conjunto CUB 200. Descrição detalhada dos tipos de características e classificadores utilizados.

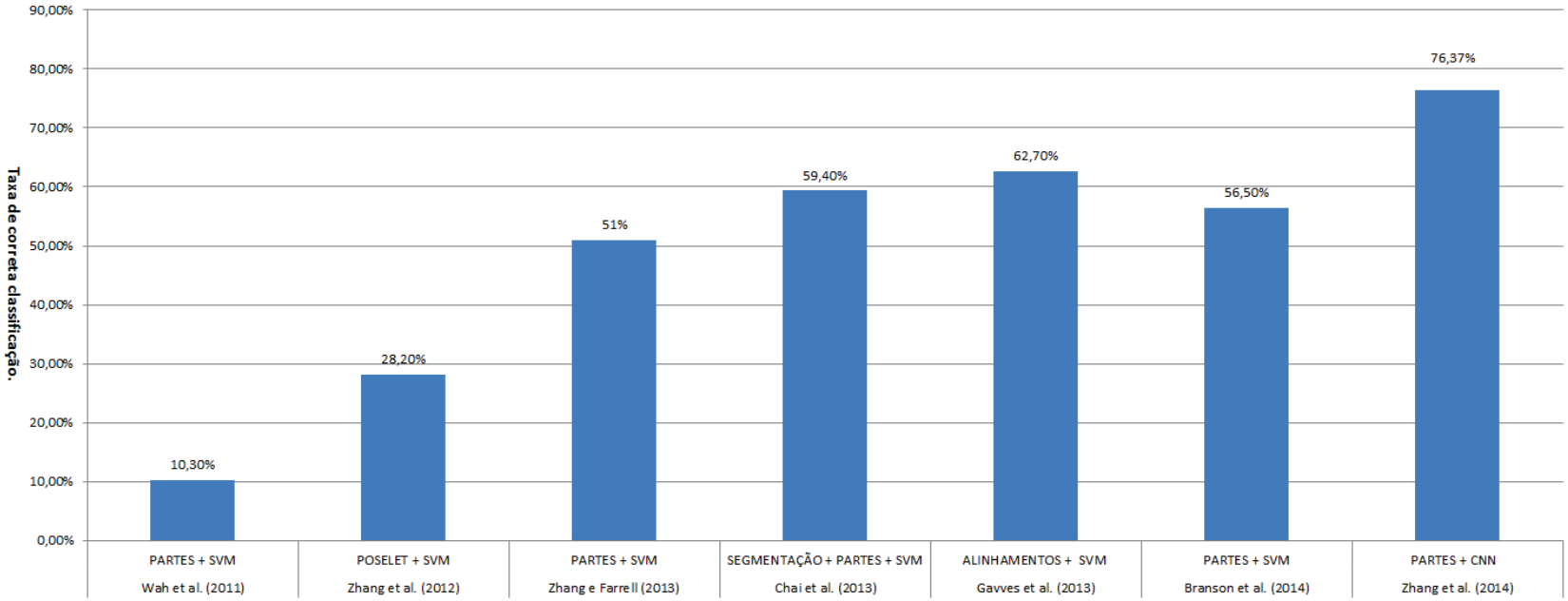


Figura 3.2: Evolução temporal (2011 - 2014) dos resultados para o conjunto CUB 200 2011. Descrição detalhada do tipo de coleta de características e classificadores utilizados.

3.9 Considerações finais

Este capítulo apresentou algumas das principais abordagens para resolver problemas de granularidade fina, com ênfase a propostas existentes para a identificação de espécies que utilizam características associadas à percepção visual, como cores, partes, entre outras. A partir do próximo capítulo serão apresentados os métodos implementados e desenvolvidos nesta tese, os quais empregam diferentes abordagens para a identificação de espécies de pássaros.

Capítulo 4

Identificação visual de espécies

Este capítulo descreve as estratégias distintas que compõem os métodos para a identificação visual de espécies propostos neste documento[1]: a) conjuntos e subconjuntos de dados; b) técnicas de pré-processamento; c) segmentação de imagens; d) extração e aprendizagem de características; e) classificadores; f) métricas de avaliação adotadas.

4.1 Conjuntos de Dados

Para facilitar a manipulação dos dados e investigar algumas condições específicas foram criados subconjuntos de dados. Cabe enfatizar que o conjunto de 200 classes representa o conjunto original para ambos os casos. A partir do conjunto de dados CUB 200 foram criados os subconjuntos descritos na Tabela 4.1. Para o conjunto de dados CUB 200 2011 os subconjuntos criados são detalhados na Tabela 4.2. A organização do subconjunto de 50 espécies foi estabelecida pelas espécies, às quais foi possível relacionar amostras de áudio. A seção 5.1 do capítulo 5 apresenta mais informações em relação à construção deste conjunto de áudio. O Apêndice B lista as espécies pertencentes a cada um dos subconjuntos utilizados. As Tabelas B.1 e B.2 apresentam as 200 espécies pertencentes aos conjuntos de dados CUB 200 e CUB 200 2011.

Tabela 4.1: Detalhamento do conjunto CUB 200 e subconjuntos derivados.

Conjunto	Total de amostras	Classes	Treinamento	Teste
2 espécies	68	2	28	40
5 espécies	151	5	62	89
17 espécies	519	17	212	307
200 espécies	**6.033**	**200**	**3.000**	**3.033**

[1]O Apêndice A apresenta os ambientes e ferramentas utilizados no desenvolvimento deste trabalho.

Tabela 4.2: Detalhamento do conjunto CUB 200 2011 e subconjuntos derivados.

Conjunto	Total de amostras	Classes	Treinamento	Teste
2 espécies	120	2	60	60
5 espécies	289	5	150	139
17 espécies	979	17	510	469
50 espécies	2.979	50	1.499	1.480
200 espécies	**11.788**	**200**	**5.994**	**5.794**

A Figura 4.1 apresenta imagens de espécies que formam os subconjuntos com 2, 5 e 17 espécies. A Figura 4.2 apresenta algumas imagens de espécies pertencentes ao subconjunto de 50 espécies.

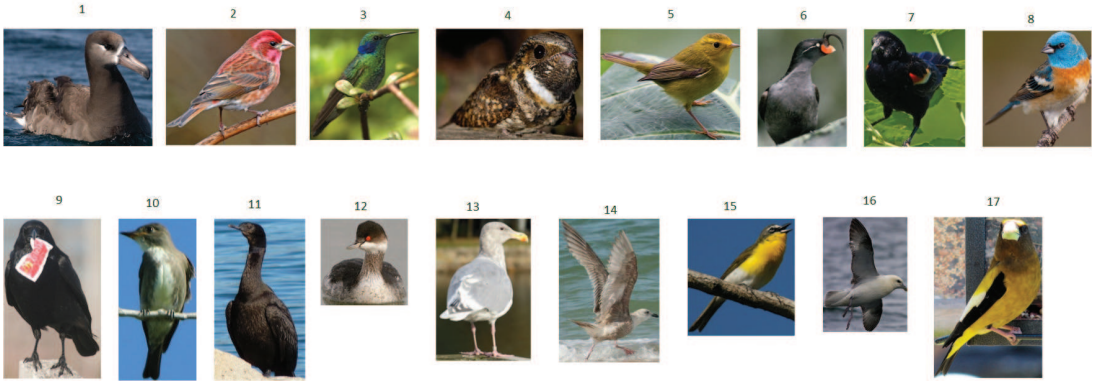


Figura 4.1: Visão geral das espécies pertencentes aos subconjuntos. A partir da numeração na parte superior da imagem é possível verificar a qual conjunto a espécie pertence.

4.2 Pré-processamento

As imagens pertencentes aos conjuntos de dados não possuem restrições para o processo de aquisição de imagens. Desta forma, alguns problemas podem ser encontrados. Por exemplo, na variação das cores, luminosidade, rotação, escalas e influência de elementos do cenário, visto que as imagens dos pássaros são frequentemente obtidas em seu habitat natural. Em decorrência, um grande problema para a etapa de extração de características é o fato de que algumas características podem ser raras ou exclusivas de uma imagem. Neste contexto, a etapa de pré-processamento é extremamente importante, já que visa a padronizar algumas variações específicas para facilitar a execução das etapas seguintes. No formato original as imagens pertencem ao espaço de cor RGB e ao formato

Figura 4.2: Imagens de 13 espécies, aleatoriamente escolhidas, pertencentes ao subconjunto de 50 espécies.

JPG[2]. Para investigar o comportamento de algumas técnicas de extração de características as imagens foram convertidas para os espaços de cor HSV e para a escala de cinza. A Figura 4.3 apresenta a conversão para diferentes espaços de cores.

(a) Imagem RGB. (b) Imagem HSV. (c) Imagem em escala de cinza.

Figura 4.3: Exemplos de representação de uma imagem em diferentes espaços de cores utilizados neste trabalho.

Visando descartar informações desnecessárias de uma imagem, uma caixa delimitadora é utilizada para isolar o objeto de interesse. No contexto deste trabalho existirá uma garantia de que o pássaro está presente na imagem e torna possível a extração de características mais discriminantes, utilizando o interior da caixa delimitadora. Os valores para a definição da caixa delimitadora são fornecidos juntamente com o conjunto de dados CUB 200 2011 (Wah et al., 2011). Tais informações foram utilizadas para recortar as imagens originais e gerar novas imagens[3]. Os conjuntos e subconjuntos descritos na seção 4.1 foram pré-processados. A Figura 4.4 apresenta a relação imagem original e imagem

[2]O formato JPEG (*Joint Photographic Experts Group*) pode utilizar as extensões: Jpg, Jpeg, Jpe, Jfif, Jif.

[3]O Apêndice C na seção C.0.3 apresenta o impacto do uso de imagens recortadas por meio dos valores de caixa delimitadora em relação ao uso das imagens originais.

(a) Imagem original. (b) Caixa delimitadora.

Figura 4.4: Exemplo da definição de caixa delimitadora. O contorno amarelo onde a imagem será recortada.

com a região de interesse delimitada.

4.3 Segmentação de imagens

Em nosso contexto, segmentação consiste em separar os *pixels* da imagem que fazem parte do pássaro, daqueles que representam o cenário (Chai et al. 2011), (Das e Manmatha 2001). O objetivo é realizar a extração de característica nas regiões da imagem em que possam existir indícios de que o objeto de interesse esteja presente. A abordagem proposta para segmentar pássaro em imagens coloridas com fundo complexo é baseada nas cores presentes na imagem. A Figura 4.5 apresenta o método proposto para segmentação baseada em cores.

A premissa adotada considera o fato de que a maioria das imagens mostram os pássaros centralizados na imagem e afastados de sua borda. A extração prévia de informações do fundo colorido da imagem faz-se por um levantamento das cores presentes nas bordas da imagem. Um percentual de *pixels* de cada borda (superior, inferior, direita e esquerda)[4] fornece uma estatística prévia do fundo colorido. A análise é expandida aos *pixels* fora da região de borda. Caso esses *pixels* apresentem valores de cores pertencentes às faixas extraídas da borda, eles serão considerados como fundo. Caso contrário, eles serão considerados pertencentes ao pássaro. Assim, é possível isolar as cores do pássaro presente na imagem e separá-lo do fundo da imagem. O método de segmentação proposto foi comparado com a segmentação ideal (em nosso contexto quando ocorre a efetiva separação do que é fundo e do que é pássaro). O *ground truth* utilizado é disponibilizado juntamente com os conjuntos de dados CUB 200. A seção 4.7.2 descreve as métricas de avaliação utilizadas para a avaliação da segmentação baseada em cores. Cabe enfatizar que o processo de segmentação foi utilizado apenas nos experimentos iniciais realizados

[4] O Apêndice C na seção C.0.2 apresenta como parâmetro de porcentagem de borda foi definido.

Figura 4.5: Visão geral do método de segmentação baseado em cores.

no conjunto de dados CUB 200. Nos demais experimentos foram utilizadas as imagens recortadas com as informações da caixa delimitadora.

4.4 Abordagem superficial

Esta seção descreve os principais elementos pertencentes à abordagem superficial.

4.4.1 Extração e avaliação de características

Neste trabalho são implementados três métodos de extração de características que produzem características globais e locais. As características globais são extraídas da imagem inteira, sendo representadas por vetores de tamanho fixo. As características de textura e cores podem ser consideradas globais. As características locais são as que utilizam regiões da imagem. O algoritmo SIFT é responsável por prover características locais referentes aos conjuntos de imagens. A literatura refere-se a essa forma de tratamento,

fortemente, direcionada à extração de características como abordagem superficial. Neste caso, o foco principal encontra-se na extração e seleção de características discriminantes para representação de um problema.

4.4.1.1 Extração de características de cor

Dentre os descritores de cor utilizamos o histograma de cores para extração de características. O uso de histogramas extraídos a partir dos canais de cor é simples e discriminativo. A saída do histograma gerado com parâmetros específicos resultará no vetor de características de cores. O histograma de uma imagem é produzido por meio da amostragem das cores em um número de faixas (*bins*) de cores. Em seguida, é realizada a contagem do número de *pixels* da imagem para cada faixa. O processo de formação de um histograma se baseia na construção de diversas pilhas também conhecidas como regiões de cores, uma para cada cor presente na imagem. Uma justificativa importante com relação à escolha deste descritor deve-se ao fato deste ser independente da resolução da imagem. Salienta-se que, em nosso contexto, não existem restrições para a formação dos conjuntos de imagens utilizados. Os histogramas utilizados consideram faixas de tamanho 10. Desta forma, o vetor de características de uma imagem RGB, por exemplo, possui 30 valores, sendo 10 para cada canal. O tamanho 10 foi definido após uma abordagem experimental entre 10 e 300. Observou-se que esse tamanho conseguiu, na maioria dos casos representar adequadamente a quantidade de cores presentes na imagem[5].

Cabe relembrar que a escolha da característica cor para identificação visual de espécies de pássaros deve-se, entre outros fatores já discutidos, ao fato de que uma cor pode se repetir em várias imagens do conjunto. Mesmo em imagens de contextos diferentes e com grande de variabilidade intraclasse, observa-se que as cores pertencentes ao pássaro são recorrentes. Além disso, na maioria dos casos são muito diferentes da cor do fundo da imagem. Tal constatação motivou a criação de uma abordagem de segmentação baseada exclusivamente nas cores presentes na imagem. A Figura 4.6 apresenta detalhadamente o método completo de identificação de espécies de pássaros que utiliza segmentação de imagens baseada em cor. Primeiro a imagem de entrada é segmentada pelo método descrito na Figura 4.5, em seguida os canais de cores presentes na imagem são separados. Um histograma de cores para cada canal é construído. A quantidade de faixas escolhida para cada canal fixa a quantidade de características de cores utilizadas. A Figura 4.7 apresenta um exemplo de um vetor de características, que será submetido a um classificador, gerado

[5] O Apêndice C na seção C.0.4 apresenta o impacto do uso de diferentes quantidades de faixas no histograma de cores.

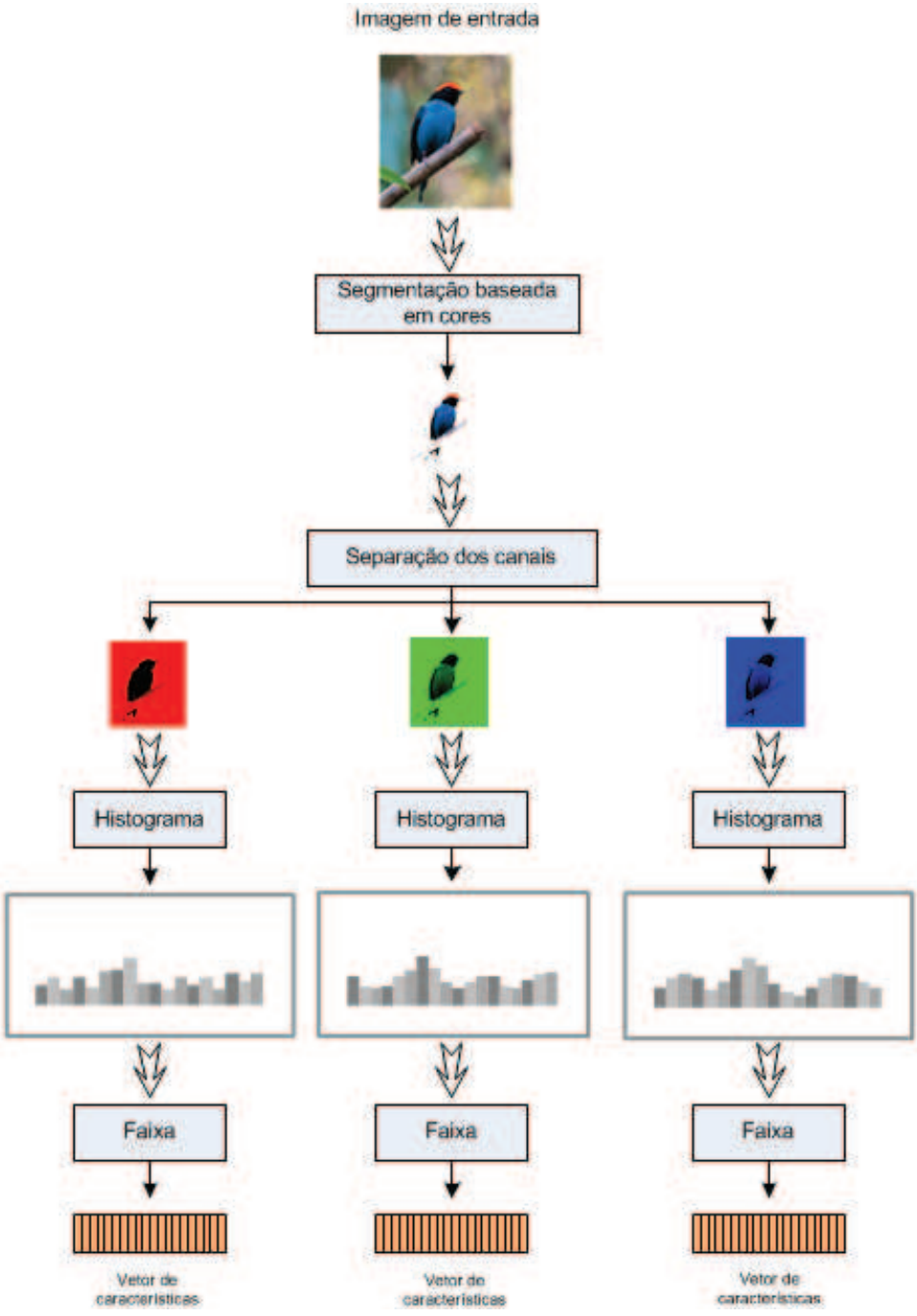


Figura 4.6: Detalhamento do método de identificação de espécies de pássaros que utiliza segmentação de imagens baseada em cor.

por meio de uma imagem no formato RGB. Os canais são agrupados e a amostra rotulada para compor o vetor de características.

Figura 4.7: Extração e concatenação de cores no nível de vetor de características.

4.4.1.2 Extração de características de textura

Apesar da dificuldade da definição formal de uma textura, a definição visual, embora não homogênea na totalidade, torna a abordagem estrutural de textura muito atrativa para a identificação de espécies de pássaros. A repetição de padrões (sendo que um conjunto de *pixels* é caracterizado por esses atributos), como por exemplo, a textura do bico, da perna, ou a disposição das penas no corpo do pássaro podem gerar características altamente discriminantes.

A Figura 4.8 apresenta uma visão geral do método empregado para a construção dos vetores de características baseados exclusivamente em texturas. Assim, os seguintes vetores são gerados por meio dos descritores $LBP_{P,R}$ descritos no capítulo 2 na seção 2.3.2.1:

- LBP_{R-G-B} - imagens no formato original são processadas em um canal de cada vez gerando diferentes vetores para cada canal.
- LBP_{GRAY} - imagens são convertidas do espaço RGB para escala de cinza e um novo vetor é gerado.
- LBP_{H} - imagens são convertidas do espaço RGB para o espaço HSV. Em seguida apenas o canal H é utilizado para gerar o novo vetor.

Dada uma imagem de entrada no formato RGB essa será convertida para os formatos HSV e GRAY. Depois da conversão da imagem para o espaço de cor HSV apenas o

canal H - *hue* (matiz ou tonalidade) é utilizado. Tal escolha deve-se ao fato de este canal possuir informação de cor em outro formato de representação de informações desconhecidas para o espaço RGB podem ser consideradas. A escolha deste espaço de cor impactou em bons resultados na abordagem de cores (discutidos no capítulo 6 na seção 6.1.1). O formato GRAY refere-se à conversão da imagem para a escala de cinza. Essa conversão foi realizada para verificar o comportamento do operador $LBP_{P,R}$ em sua concepção original. Em seguida os operadores $LBP^{u2}_{P,R}$, $LBP^{ri}_{P,R}$, $LBP^{riu2}_{P,R}$ são executados com os valores 8 e 12 para P e 2 e 1.5 para R[6].

Fator importante para um problema de classificação é a seleção das características mais discriminativas. Em aplicações com textura, dependendo do tipo do operador e os parâmetros P e R utilizados, é possível trabalhar centenas de características. Nos experimentos aqui descritos foram considerados os vetores de características originais, sem a aplicação de métodos de seleção ou redução de características. A Tabela 4.3 apresenta a quantidade de características geradas de acordo com as configurações de descritores e escolha de operadores em cada um dos vetores.

Tabela 4.3: Quantidade de características geradas pelos operadores utilizados.

	P=8, R=1	P=12, R=1.5
$LBP^{u2}_{P,R}$	59	36
$LBP^{ri}_{P,R}$	36	352
$LBP^{riu2}_{P,R}$	10	14

4.4.1.3 Extração de características de forma

As características locais, em geral, conseguem ser diferente de uma região vizinha e podem ser descritas como padrões na imagem. Um descritor representa uma característica local sendo obtido por meio de informações retiradas de uma região em torno de uma característica local. Em decorrência, o conjunto de descritores representa a forma (estrutura) do objeto contido em uma imagem. Neste trabalho, as principais etapas de extração de características de forma são:

- detecção de pontos-chave;
- atribuição de descritores de pontos para um conjunto pré-determinado por meio de um algoritmo de quantização vetorial, ou seja, a definição de um vocabulário;
- construção de um vetor de pontos-chave, ou seja, a implementação de um histograma;

[6]O Apêndice C na seção C.0.5 indica os trabalhos que serviram de referência para os valores de P e R

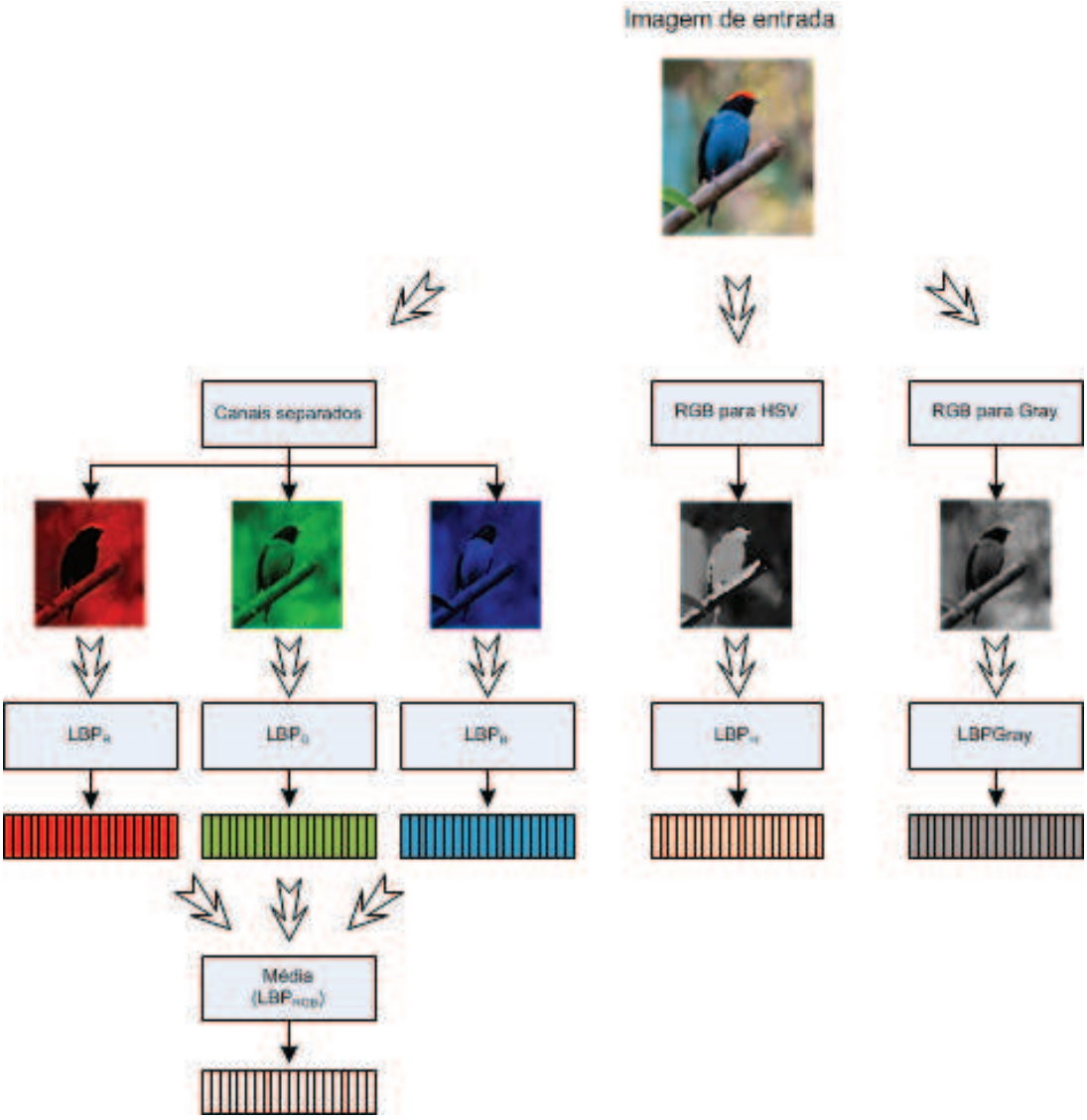


Figura 4.8: Detalhamento do método de extração de textura. As abordagens iniciais (na variações do conjunto de dados CUB 200) calculavam a média dos canais RGB para reportar os resultados. Essa abordagem foi substituída pela escolha do canal com a melhor representação. A seção 6.2.2 do capítulo 6 discute essa escolha.

Essas etapas visam a maximizar a precisão da classificação. Assim, o descritor extraído no primeiro passo não deve sofrer interferência a variações de iluminação, escala, rotação, etc. A detecção de pontos-chave é realizada por meio do algoritmo SIFT, descrito na seção 2.3.3.1 e a variação Dense SIFT[7] disponibilizada por Vedaldi e Fulkerson (2010). Além das vantagens já descritas para esse algoritmo, Csurka et al. (2004) enfatizam que os pontos-chaves, são derivados de Gaussianas lineares simples. Neste caso, espera-se mais estabilidade perante as perturbações típicas de imagem, como o ruído de derivados de Gaussianas superiores ou invariantes diferenciais. Os valores gerados por meio dos histogramas de orientação representam gradientes de uma região na imagem (128 valores) o que torna possível uma representação rica e potencialmente mais discriminativa.

Como as imagens possuem número variável de pontos, podem gerar vetores de alta dimensionalidade. O próximo passo é como resumir o conteúdo para um tamanho fixo, sendo que apenas um conjunto desses descritores irá compor o vocabulário. O tamanho do vocabulário utilizado no segundo passo deve ser grande o suficiente para distinguir mudanças relevantes em partes da imagem, mas não tão amplo para distinguir variações irrelevantes ou ruídos. De acordo com Sivic e Zisserman (2003) o vocabulário deve ser construído a partir de partes de uma imagem, levando em conta a precisão de correspondência de cada parte e o poder de expressão avaliados sobre o restante do conjunto. Para agrupar algumas centenas de milhares de descritores visuais em um vocabulário o algoritmo Elkan é utilizado. O algoritmo proposto por Elkan (2003) realiza a quantização vetorial por meio do algoritmo K-*means* que permite uma aceleração ao processo com uma grande quantidade de pontos-chave, mas calcula exatamente o mesmo resultado que o algoritmo padrão. Finalmente, os pontos-chave devem ser armazenados em uma estrutura indexada. Cada ponto-chave possui uma referência na imagem de origem Sivic e Zisserman (2003).

Csurka et al. (2004) exemplificam os vetores de características quantizados de pontos-chave em analogia com palavras-chave na categorização de textos. No entanto, neste caso os pontos-chave não têm necessariamente repetição de pontos, por exemplo, olhos de uma pessoa ou rodas de um veículo. Em vez disso, o objetivo é usar um vocabulário de pontos-chave que permita auxiliar a identificação no conjunto de dados de treinamento. A Tabela 4.4 apresenta as variações utilizadas de diferentes tamanhos de vocabulário e diferentes regiões de coleta na imagem.

[7] A biblioteca VLFeat utiliza uma função para calcular mais rapidamente o conjunto de descritores SIFT. Neste caso, o descritor PHOW (*Pyramid Histogram Of visual Words*) permite o cálculo em diferentes resoluções para cada imagem de entrada. Algumas configurações adotadas: imagem no formato escala de cinza; espaçamento uniforme (*3 pixels*); escalas (4, 6, 8, e 10 *pixels*); tamanho da janela (1.5) e limiar de constraste (0.005).

Tabela 4.4: Características geradas por meio da detecção de pontos-chave, vocabulários e histogramas.

Tamanho do Vocabulário	Histograma	Quantidade de características
600	Imagem inteira	600
1.200	Imagem inteira	1.200
600	4 regiões + 16 regiões	12.000
1.200	4 regiões + 16 regiões	24.000

A Figura 4.9 detalha o processo de divisão de regiões da imagem para gerar os histogramas (vetor de pontos-chave) e construir um vetor de 24.000 características. A alta dimensionalidade do vetor deve-se ao fato do tamanho do vocabulário utilizado. Dada a dificuldade de encontrar um compromisso entre o tamanho do vocabulário ideal e a representação real do problema de identificação de espécies para um conjunto formado sem restrições, optou-se por vocabulário abrangente de 1.200 itens.

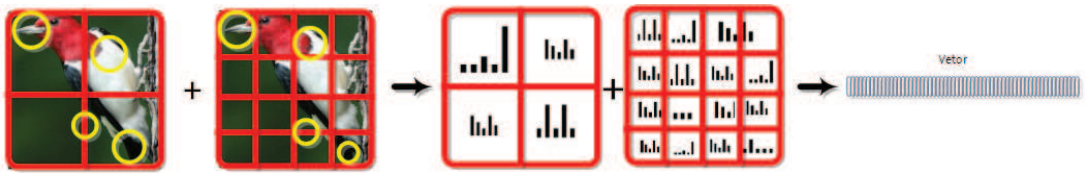


Figura 4.9: Detalhamento da construção do vetor de alta dimensionalidade. Inicialmente a imagem é dividida em 4 regiões em seguida 16 regiões. As saídas destes histogramas são concatenadas. Como o tamanho do vocabulário escolhido é 1.200, então 4x1.200 + 16x1.200 resultará em um vetor de 24.000 características.

4.4.1.4 Classificação

Para os métodos de abordagem superficial, por meio dos vetores de características, a classificação supervisionada poderá ser executada. Neste trabalho, as SVMs foram implementadas por meio da biblioteca LIBSVM disponibilizada por Chang e Lin (2013). Para as características de cores e texturas o *kernel* utilizado foi o *Radial Basis Function* e a estratégia *um-contra-um*, construindo um classificador para cada par de classes. A tarefa de classificação mapeada para o LIBSVM (SVC - *support vector classification*), para duas ou múltiplas classes, envolve a separação dos dados em conjuntos de treinamento e teste. Desta forma, cada instância no conjunto de treinamento contém os rótulos de classe e vários atributos. O objetivo da SVM é produzir um modelo, com base nos dados de treino, capaz de predizer os valores dos dados de teste. Os conjuntos de treinamento e

teste foram detalhados na seção 4.1, a divisão dos conjuntos segue a proposta em trabalhos de referência (muito próxima de 50% para treinamento e 50% para teste (Welinder et al., 2010) e Wah et al. (2011)). Para os vetores de alta dimensionalidade (de 600 a 24.000 características), gerados pela extração de pontos-chave, foi utilizada a biblioteca LIBLINEAR disponibilizada por Fan et al. (2008) e a estratégia *um-contra-todos*, construindo classificadores binários que distinguem cada uma das classes das demais.

4.5 Abordagem profunda

Em contraste com essa prática predominante de extração de característica, recentemente muitos trabalhos adotam a abordagem de aprendizado de características. Movendo-se para abordagens profundas, o maior diferencial em relação aos métodos com foco em extração de características é o fato de, neste caso, não existir a etapa de extração de característica. Ao invés disso, o próprio modelo aprende os descritores relevantes ao problema. Por exemplo, na abordagem superficial, o descritor LBP busca os padrões locais na imagem, gerando um código para a vizinhança de cada *pixel* e é o histograma dos diferentes códigos que é transformado em vetor de características para ser aplicado ao classificador SVM. No caso da abordagem que incorpora o aprendizado de características, as imagens são tratadas diretamente pelas camadas convolucionais que obtêm as características de *pixels* e as envia para tratamento nas próximas camadas.

4.5.1 Aprendizagem de características

A aprendizagem de característica no nível de *pixel* pode ser feita utilizando uma rede neural convolucional (*Convolutional Neural Network* - CNN), (Krizhevsky et al., 2012). Essa seção detalha a definição da arquitetura e os parâmetros da rede neural convolucional utilizada neste trabalho. São utilizados camadas de convoluções seguidas de agrupamento, camadas conectadas localmente, terminando com uma camada totalmente conectada. A Figura 4.11 apresenta uma visão geral da arquitetura de aprendizagem profunda utilizado neste trabalho. A seguir cada camada é detalhada:

- *Camadas de convolução*: os parâmetros desta camada dependem do tipo da imagem de entrada. As camadas convolucionais geram mapas de características que são aplicados às imagens. A Figura 4.11 apresenta a utilização de duas camadas convolucionais. Na primeira, as imagens de entrada têm o tamanho de 64x64 com 3 canais (RGB). Os filtros aplicados são de 64x64 a distância entre as aplicações dos filtros é *stride* 1. O tamanho da janela é definido em 5x5. Fazem parte destas ca-

madas neurônios com função de ativação Relu (Krizhevsky et al., 2012). A segunda possui a mesma configuração, a única alteração ocorre nos valores de entrada, já que, são provenientes de camadas anteriores.

- *Camadas de agrupamento*: estas camadas, inseridas logo após as camada convolucionais, têm o objetivo de reduzir a dimensionalidade e capturar pequenas invariâncias de translação. Na definição destas camadas é necessário informar o tamanho da janela que será utilizado. Neste caso o valor utilizado é de 3x3.

- *Camadas conectadas localmente*: estas camadas apenas conectam os neurônios dentro de uma pequena janela para a próxima camada. São semelhantes às camadas convolucionais, mas não compartilham os valores dos pesos. A Figura 4.11 apresenta duas camada conectadas localmente, uma seguida da outra. Na primeira, os filtros aplicados são de 64x64 e o *stride* é 1. O tamanho da janela é 3x3. Fazem parte desta camada neurônios com função de ativação Relu (Krizhevsky et al., 2012). Na segunda, os filtros aplicados são de 32x32. Os demais parâmetros são equivalentes.

- *Camada totalmente conectada*: esta camada depende do número de classes consideradas. Fazem parte desta camada neurônios com função de ativação *identidade*[8].

Para aumentar o número de exemplos na etapa de treinamento do modelo utilizamos a extração de subimagens da imagem original. A metodologia adotada segue as etapas propostas por Hafemann et al. (2014) e realiza translações da imagem original, obtendo um número de subimagens[9]. A possibilidade de considerar um número maior de imagens (mais exemplos distintos) previne a ocorrência de sobreajuste na etapa de treinamento. Entretanto, essa condição ocorre em tempo de execução e ao final do processo o número de imagens de conjunto de dados é igual ao inicial.

A metodologia proposta por Hafemann et al. (2014), inicialmente, realiza a extração aleatória de subimagens a partir de imagens do conjunto de treinamento. O modelo é treinado em relação a partes da imagem, considerando a dimensão de fragmentos iguais para todas as imagens. Para tanto, é necessária uma estratégia para dividir as imagens de teste em partes, executá-las por meio do modelo e combinar os resultados. A solução adotada foi selecionar uma grade de subimagens em relação à imagem, ou seja, utilizar o conjunto de todas as subimagens não sobrepostas.

Para a fase teste, as subimagens são combinadas para a imagem inteira. Durante a execução do modelo, cada subimagem possui uma probabilidade de pertencer a uma

[8]Função *logistic* ou *identidade* é representada por: $f(x) = 1/(1 + e^{-x})$.
[9]Subimagem é a tradução para o termo em inglês: *Patch*.

determinada classe. Para combinar as subimagens de uma imagem de teste é utilizada a regra Sum[10]. Desta forma, a combinação para uma imagem de teste é obtida da classe que maximiza a soma das probabilidades em todas as subimagens de uma imagem.

Em síntese, a metodologia pode ser descrita da seguinte forma: a) padronização do tamanho das imagens para 64x64, com 3 canais de cores no formato RGB; b) extração de subimagens das imagens originais (neste caso o tamanho da subimagem é 56 e o tamanho da borda desconsiderada é 4; c) treinamento do modelo com imagens e subimagens; d) utilização do conjunto de testes.

Algumas arquiteturas foram a para a identificação de espécies. Essas arquiteturas são apresentadas na Figura 4.10 e foram inspiradas nos trabalhos de (Krizhevsky e Hinton, 2009), (Krizhevsky et al., 2012), (Ciresan et al., 2012). A arquitetura apresentada na Figura 4.11 foi a que mais se adaptou ao problema de identificação de espécies de pássaros[11].

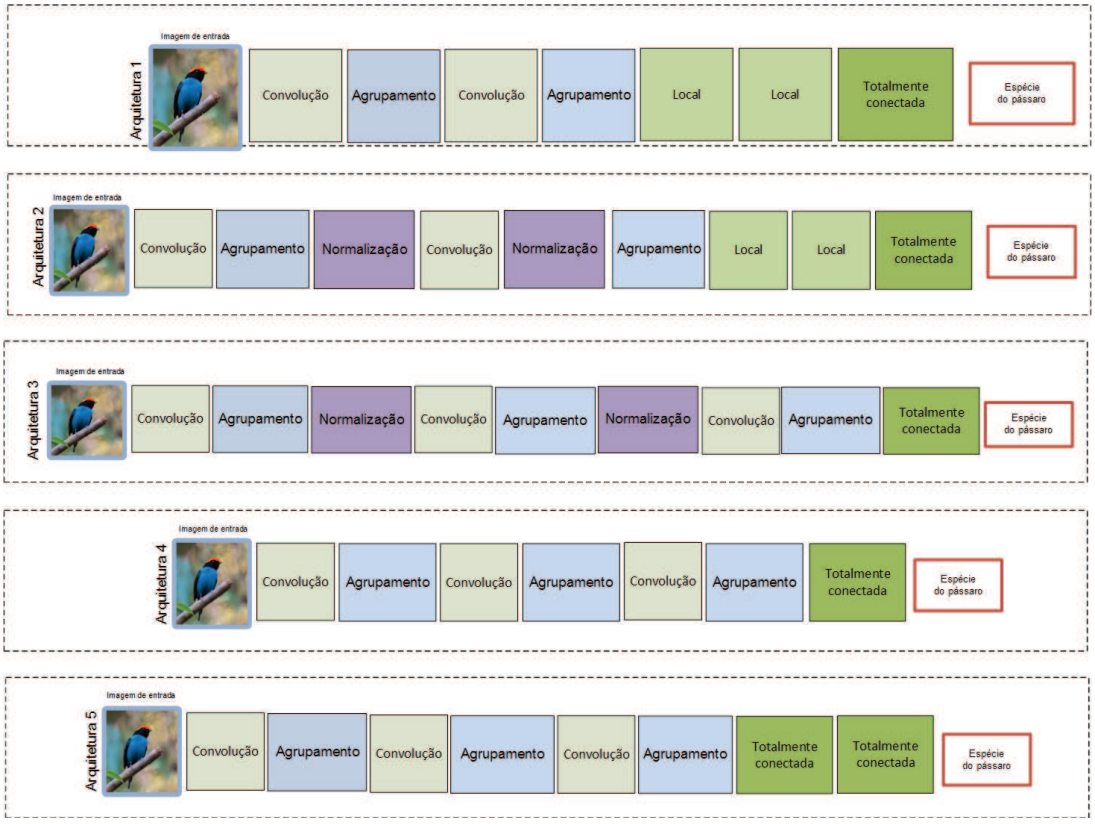


Figura 4.10: Diferentes arquiteturas avaliadas para rede neural convolucional em relação à identificação de espécies de pássaros.

[10]Mesma regra descrita na seção 4.6

[11]O Apêndice C na seção C.0.6 apresenta resultados de 10 execuções e a variância obtida para as cinco arquiteturas indicadas na Figura 4.10. Os experimentos descritos nas Tabelas C.6 utilizam imagens de 32x32 *pixels* e os indicados na Tabela C.7 utilizam imagens de 64x64 *pixels*.

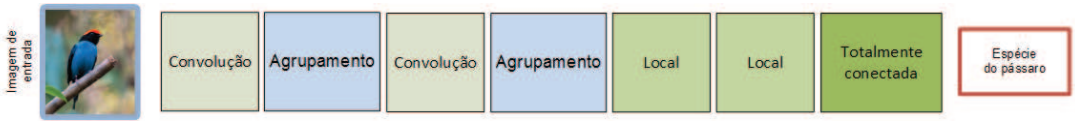


Figura 4.11: A arquitetura adotada para rede neural convolucional.

4.5.2 Parametrização

A abordagem adotada para a aprendizagem profunda utiliza uma rede neural com múltiplas camadas, conforme descrição na seção 2.5.2. Para o funcionamento de uma rede neural convolucional é necessário definir uma arquitetura e alguns parâmetros de aprendizagem. A seção 4.5.1 apresentou a arquitetura da rede que permanecerá fixa. A seção atual descreve os parâmetros de aprendizagem que podem ser alterados enquanto testes experimentais são realizados para encontrar o melhor desempenho.

A Figura 4.12 apresenta um resumo do processo de classificação para uma CNN. Neste caso, o conjunto de dados é dividido em três conjuntos: treinamento, validação (parte do conjunto de treinamento) e teste. As imagens são submetidas à metodologia de geração de subimagens. A arquitetura é inicializada por uma camada convolucional, seguida de camada de agrupamento, e camadas conectadas localmente. A camada totalmente conectada recebe como entrada os pesos da camada convolucional inicializados com $initW=0.0001$. A saída dessa camada é igual à quantidade de classes que o problema leva em conta. O parâmetro *[probs]* indica o uso de uma função *softmax*. A função *softmax* foi detalhada na seção 2.5.1.

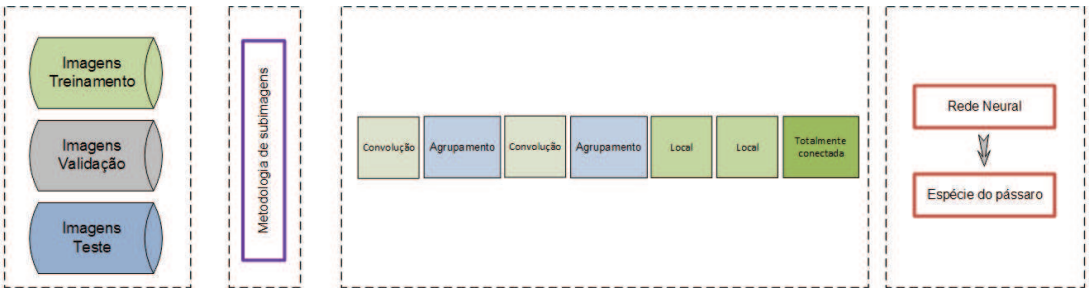


Figura 4.12: Visão geral do processo de identificação de espécies por meio de CNN.

A rede é treinada por meio do método gradiente descendente estocástico, um algoritmo executa iterativamente pequenos passos em direção à diminuição do erro, sendo definido por uma função de perda e alguns parâmetros adicionais. Para isso considera-se

que os dados de treinamento sejam avaliados por uma função de perda. O algoritmo *backpropagation* é usado da mesma forma como em uma rede totalmente conectada. A rede neural convolucional deve definir uma função objetivo para otimização, neste caso a função utilizada é a regressão logística. Os parâmetros *[logprob]* e *[cost.logreg]* consideram dois valores: os rótulos reais e as probabilidades estimadas. Os parâmetros de custos utilizam um coeficiente escalar para função objetivo. Isso fornece uma maneira fácil de ajustar a taxa de aprendizagem global da rede. Os principais parâmetros de aprendizagem utilizados são, Krizhevsky (2014): *[epsW = 0,001]* = taxa de aprendizagem para pesos; *[epsB = 0,002]* = taxa de aprendizagem para *bias*; *[momW = 0,9]*= *momentum* para pesos; *[momB = 0,9]*= *momentum* para *bias*; *[wc = 0]*= regularização do tipo L2; *[dropout]*= 0.5 = valor de abandono.

4.6 Combinação de classificadores

A combinação de classificadores tem como objetivo central buscar o melhor conjunto de classificadores, bem como o melhor método de combinação desses conjuntos. A viabilidade destas estratégias se apresenta na melhora dos resultados na combinação em relação ao resultado obtido pelos classificadores individuais. As possíveis maneiras de combinar saídas de L classificadores em conjuntos dependem das informações obtidas por meio dos classificadores individuais. Neste trabalho, consideramos: i) $L = 7$ (todos os classificadores individuais); ii) $L = 3$ (os três melhores desempenhos individuais); iii) $L = 2$ (os dois melhores desempenhos individuais). Além disso, duas formas de combinação: i) no nível de medida, que considera a confiança atribuída a cada classe. Neste nível foram implementadas as estratégias: máximo, soma, produto, soma ponderada e produto ponderado; ii) em nível abstrato, quando cada classificador produz um rótulo. Por meio destes rótulos são implementados o Oráculo, voto majoritário relacionado à medida estatística moda, voto majoritário ponderado pela taxa de acerto do classificador individual e o voto majoritário ponderado pelo tipo de caraterística utilizada (cor, textura, etc).

Uma abordagem para a combinação de classificadores é a fusão de rótulos de classes preditas por um classificador. Esse tipo de fusão assume que o classificador conhece todo o espaço de características e a classificação pode ser resultado de uma opinião coletiva. Assim, sendo $D_1, D_2, ...D_L$ um conjunto de L classificadores, é possível obter diferentes tipos de saídas para esse conjunto (Kuncheva, 2007):

- *Nível abstrato*: $D_i, i = 1, 2, ..., L$ produz um rótulo $S_i \in \Omega = w_1, w_2, ..., w_c$ conjunto de classes. Portanto, para qualquer $X \in R^n$, a coleção produz um vetor $S =$

$[S_1, S_2, ...S_L^t] \in \Omega^t$.

- *Nível de medida*: cada $D_i\, i = 1, 2, ..., L$, produz um vetor $[d_i, 1d_i, 2...d_i, c]^t$ de medidas entre $[0, 1]$ que representam o suporte para a hipótese que x seja da classe $w_j, j = 1, 2, ..., c$.

- *Nível oráculo*: é usado com z_1[12] para verificar se D_i produz a saída correta ou incorreta para X. Neste caso, o conjunto produz um vetor $[y_1, y_2, ...y_L^t]$ de valores binários $y_1 \in 0, 1, i = 1, 2, ...L$ que indicam uma classificação correta ou incorreta. O oráculo contabiliza classificação correta nos casos em que pelo menos um D_i indique classificação correta.

Para implementar regras de decisão, no nível de medidas, assumimos um vetor *d-dimensional* de características, x e c classes possíveis, rotuladas ω_1 a ω_c organizado como um conjunto de rótulos. De forma que um conjunto de k classificadores $D_1...D_k$ onde cada classificador D_i produz na saída $[P_{Di}(\omega_1|x), P_{Di}(\omega_2|x), ..., P_{Di}(\omega_j|x)]$, onde $P_{Di}(\omega_j|x)$ representa o suporte para que hipótese de que o vetor x seja da classe w_j. As seguintes regras de decisão podem ser adotadas:

- *Max*: a classe com o nível de confiança mais elevado é a vencedora. Conforme a Equação 4.1:

$$\begin{aligned} \widehat{\omega} = arg\, max\, P_{Di}(\omega|x) \\ i \in [1, k] \\ \omega \in \Omega \end{aligned} \tag{4.1}$$

- *Sum*: baseada no somatório dos níveis de confiança fornecidos pelos classificadores. Os níveis de confiança são somados para cada classe e a classe cuja soma resultante for a mais elevada é declarada vencedora. Conforme a Equação 4.2.

$$\begin{aligned} \widehat{\omega} = arg\, max \sum_{i=1}^{k} P_{Di}(\omega|x) \\ \omega \in \Omega \end{aligned} \tag{4.2}$$

- *Prod*: baseada na multiplicação dos níveis de confiança fornecidos pelos classificadores. Os níveis de confiança são multiplicados para cada classe, e a classe cujo

[12]Parte do conjunto Z. Sendo que $Z = \{z_1, ..., z_N\}, z_j \in R^n$ e conhece os rótulos para todos os valores z_j.

produto resultante for o mais elevado, é declarada vencedora. Conforme a Equação 4.3.

$$\widehat{\omega} = arg\,max \prod_{i=1}^{k} P_{Di}(\omega|x) \qquad \omega \in \Omega \tag{4.3}$$

- *wSum*: é possível adicionar pesos às saídas dos classificadores. Neste caso, o nível de confiança de cada classificador é multiplicado por k pesos w_i, e os pesos são específicos para cada classificador. Conforme a Equação 4.4.

$$\widehat{\omega} = arg\,max \sum_{i=1}^{k} w_i P_{Di}(\omega|x) \qquad \omega \in \Omega \tag{4.4}$$

- *wProd*: segue o mesmo procedimento da soma ponderada. Conforme a Equação4.5.

$$\widehat{\omega} = arg\,max \prod_{i=1}^{k} [P_{Di}(\omega|x)]^{w_i} \qquad \omega \in \Omega \tag{4.5}$$

As regras de combinação adotadas para as saídas no nível abstrato correspondem ao voto da maioria, ou voto majoritário. As estratégias simples e computacionalmente eficientes, consistindo apenas de contagem do número de classificadores que retornaram a uma determinada classe. Nesse caso, será vencedora a classe que recebe maior número de votos dos classificadores. O trabalho com rótulos no nível abstrato, de certa forma restringe o número de fusões que podem ser realizadas e, consequentemente, as conclusões que podem obtidas. O voto majoritário se apresenta como uma alternativa viável para a verificação de regras nesses casos.

A fusão por voto majoritário pode ser formalizada da seguinte forma, sendo: $[d_{i,1}, d_{i,2}, ...d_{i,c}^{t}]$ um vetor que $d_{i,j} \in 0, 1$ indica a saída do classificador $D_i, i = 1, 2, ..., L$ com relação a uma amostra x pertencer ou não a uma classe $\omega_j, j = 1, 2, ..., c$ o voto majoritário pode escolher a classe ω_k conforme a Equação 4.6:

$$g_k(x) = \sum_{i=1}^{L} d_{i,k} = \max_{j=1}^{c}\{g_j(x) = \sum_{i=1}^{L} d_{i,j}\} \tag{4.6}$$

A decisão final poderá levar em conta uma classe extra ω_{c+1} para nenhuma das

alternativas, quando a medida não ultrapassa um limiar $\alpha L, 0 < \alpha \leq 1$ para a classe ω_k. Sendo a decisão final expressa na Equação 4.7:

$$\begin{cases} \omega_k, & \text{se} \sum_{i=1}^{L} \mathrm{d}_{i,k} \geq \alpha L \\ \omega_{c+1} & \text{caso contrário.} \end{cases} \tag{4.7}$$

Quando $x = \frac{1}{2} + \in, 0 <\in< \frac{1}{L}$ temos a maioria simples, ou seja 50%+1 dos votos. Quando $\alpha = 1$ existe a unanimidade.

A implementação da regra que considera o voto majoritário ponderado é indicada quando os classificadores no conjunto não apresentam desempenho similar, caso em que o classificador mais competente poderá indicar a decisão final para o voto do conjunto. Quando o voto majoritário ponderado é utilizado a Equação 4.6 precisa ser alterada para 4.8. Onde w_i é coeficiente de peso para o classificador D_i. Por convenção, $\sum_{i=1}^{L} w_i = 1$ e $w_i \alpha p_i$ (p_i é a probabilidade de acerto do classificador D_i).

$$g_k(x) = \sum_{i=1}^{L} d_{i,k} = \max_{j=1}^{c} \{g_j(x) = \sum_{i=1}^{L} w_i d_{i,j}\} \tag{4.8}$$

O voto ponderado nem sempre é melhor do que o voto do melhor classificador individual, , mas normalmente é mais exato que o voto majoritário sem ponderação. As regras de decisão propostas que adotam o voto majoritário são:

- $Voto maj_{moda}$ - neste caso, os rótulos atribuídos pelos classificadores são analisados para verificar a ocorrência da medida estatística moda[13]. Caso a medida seja identificada ela é comparada a saída real do classificador. Caso contrário, à saída do melhor classificador individual é considerada.

- $Voto maj_{acc}$ - nesta regra a contagem do voto majoritário é realizada de forma ponderada. A ponderação é realizada pela taxa de acerto dos classificadores individuais e depende de quantos classificadores fazem parte da combinação.

- $Voto maj_{feature}$ - nesta regra o voto majoritário ponderado poderá obter a opinião de acordo com o tipo de característica utilizada. A ponderação depende do tipo de características utilizadas pelos classificadores que fazem parte da combinação.

[13]Moda é a medida de tendência representada pelo valor observado com mais frequência em um conjunto de dados.

4.7 Métricas de avaliação

Para avaliar os resultados obtidos é necessário definir algumas medidas de avaliação para as estratégias de classificação e segmentação.

4.7.1 Avaliação das estratégias de classificação

O resultado de uma classificação de um problema de duas classes, onde as classes são positivas ou negativas poderá obter uma predição correta ou incorreta apresentadas na Tabela 4.5. Para um problema com múltiplas classes observe a Tabela 4.6. Por conseguinte, algumas medidas podem ser definidas e aplicadas para avaliar diferentes contextos. Dado um classificador e suas instâncias, há quatro resultados possíveis em uma matriz de confusão que podem ser utilizados para avaliar o desempenho do classificador.

Tabela 4.5: Matriz de confusão clássica.

	Predição Correta	**Predição Incorreta**
Positivo	Verdadeiro positivo (VP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadeiro negativo (VN)

A abordagem de problemas com mais de duas classes pode ser mais complexa e difícil de administrar, tendo em vista que as instâncias podem ser corretamente ou incorretamente classificadas em relação a qualquer classe. Com c classes, a matriz de confusão se transforma em uma matriz contendo os *"y"* resultados do classificador. Apesar disso este método é viável na maioria dos casos. A Tabela 4.6 apresenta a matriz de confusão para problemas de múltiplas classes. Neste contexto, *"c"* representa a classe e *"n"* a amostra pertencente à classe.

Tabela 4.6: Matriz de confusão para problemas com múltiplas classes.

	Predição c_1	**Predição** c_2	**Predição** $c_{...}$	**Predição** c_c
Rótulo c_1	n_{11}	n_{12}	$n_{...}$	n_{1c}
Rótulo c_2	n_{21}	n_{22}	$n_{...}$	n_{2c}
Rótulo $c_{...}$	$n_{...}$	$n_{...}$	$n_{...}$	$n_{...}$
Rótulo c_c	n_{1c}	n_{2c}	$n_{...}$	n_{cc}

Na matriz de confusão é possível analisar a distribuição entre as classes e os relacionamentos entre as linhas. Assim, qualquer medida de desempenho que utilize valores de ambas as colunas será necessariamente sensível à desproporção entre as classes. A partir da matriz de confusão, diferentes métodos quantitativos podem ser derivados e algumas métricas podem ser calculadas. Neste trabalho as métricas adotadas são:

- *Taxa de acerto / Acurácia*: é o número de classificações corretas (a soma dos resultados Verdadeiros Positivos e Verdadeiros Negativos) dividido pelo número total de classificações (a soma de todos os itens). Conforme representa a equação 4.9.

$$Taxa\ de\ acerto = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.9}$$

- *Taxa de erro*: é o número de classificações incorretas (a soma dos Falsos Positivos e Falsos Negativos) dividido pelo número total de classificações. A taxa de erro assume um papel importante, uma vez que o objetivo constante é construir classificadores com baixa taxa de erro quando são apresentados novos exemplos, conforme a equação 4.10.

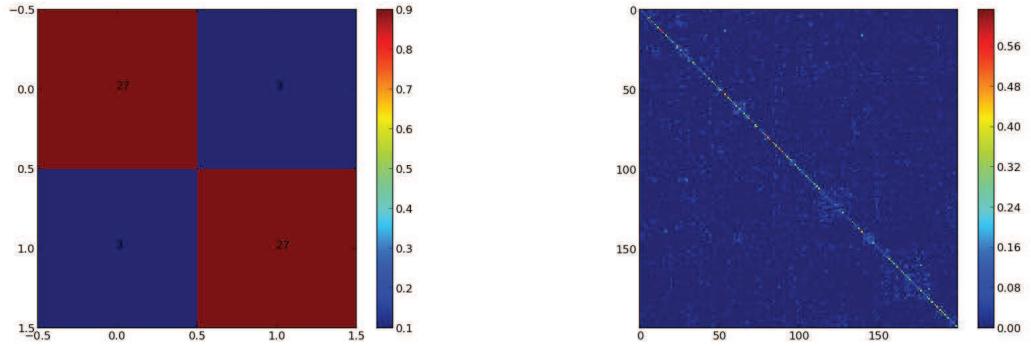
$$Taxa\ de\ erro = \frac{FP + FN}{VP + VN + FP + FN} \tag{4.10}$$

A matriz de confusão poderá representar de maneira gráfica os resultados da classificação. Entretanto, quando o número de classes é elevado, essa representação é inviável. Visando facilitar a análise da matriz de confusão com grande quantidade de classes, este trabalho utilizará uma forma de representação baseada em cores. A Figura 4.13 apresenta um exemplo desta representação baseada nas opções disponíveis pela ferramenta *Scikit-learn*[14]. Neste exemplo, a imagem à esquerda representa uma matriz de confusão para um problema que considera duas classes. A imagem à direita representa uma matriz de confusão para um problema que considera 200 classes (na forma convencional de representação é uma matriz de 200x200). A escala de cores é iniciada por variações de azul e finalizada em variações de vermelho. Isso significa que, quanto mais a cor vermelha estiver presente, melhor o desempenho do classificador. No caso do problema com duas classes é possível observar que para a diagonal principal da matriz foi atribuída variação da cor vermelha. Por outro lado, para os demais valores, variações de azul.

4.7.2 Avaliação das estratégias de segmentação

A avaliação de uma técnica de segmentação é complexa. A partir de uma análise visual a segmentação de imagens de pássaros pode indicar a eficiência do método aplicado, porém, dificilmente fornecerá justificativa sólida da validade da técnica. Deste modo, uma abordagem numérica e automática se faz necessária para comparar a segmentação obtida e a segmentação ideal (em nosso contexto quando ocorre a efetiva separação do que é fundo

[14]http://scikit-learn.org/stable/auto_examples/plot_confusion_matrix.html



(a) Matriz de confusão resultante de 2 classes.

(b) Matriz de confusão resultante de 200 classes.

Figura 4.13: Exemplos da representação da matriz de confusão por escala de cores.

e do que é pássaro). Neste trabalho a segmentação ideal utilizada é disponibilizada, juntamente com os conjuntos de dados CUB 200 e CUB 200 2011. Esse *ground-truth* será sobreposto ao resultado apresentado pela técnica proposta para mensurar a qualidade da segmentação em termos de precisão e revocação). As métricas para avaliação da segmentação proposta são apresentadas nas Equações 4.11 e 4.12. A Tabela 4.7 apresenta uma matriz de confusão utilizada para a avaliação segmentação. Os detalhes para a construção da matriz são descritos a seguir:

- Um *pixel* verdadeiro positivo (VP) é um *pixel* pertencente ao pássaro na segmentação ideal, o qual também é considerado um *pixel* do pássaro na técnica de segmentação proposta;
- Um *pixel* verdadeiro negativo (VN) é um *pixel* pertencente ao fundo segmentação ideal, o qual também é considerado um *pixel* de fundo na técnica de segmentação proposta;
- Um *pixel* falso negativo (FN) é um *pixel* pertencente ao pássaro na segmentação ideal, o qual na segmentação proposta foi considerado como fundo;
- Um *pixel* falso positivo (FP) é um *pixel* pertencente ao fundo na segmentação ideal, o qual na segmentação proposta foi considerado como pássaro;

$$Taxa\ de\ Erro = \frac{FP + FN}{VP + FP + VN + FN} \tag{4.11}$$

$$Taxa\ de\ Segmentação = \frac{VP + VN}{VP + FP + VN + FN} \tag{4.12}$$

Tabela 4.7: Matriz de confusão para avaliação da segmentação a partir da relação *ground-truth*, e resultado da segmentação proposta.

	Predição Correta	**Predição Incorreta**
Segmentação	VP (pássaro|pássaro)	FP (fundo|pássaro)
Não Segmentação	FN (pássaro|fundo)	VN (fundo|fundo)

4.8 Considerações finais

Neste capítulo foram descritos os principais métodos implementados e desenvolvidos neste trabalho. Destacamos as funções e os parâmetros adotados em cada método, a arquitetura selecionada, demonstrando em todos os casos estruturas e configurações. Foram apresentados inicialmente os conjuntos de dados, atividades de pré-processamento, segmentação, extratores de características e configurações, passando em seguida a detalhes do processo de classificação e métricas de avaliação. No próximo capítulo será descrito o método proposto para fusão de informação visual e acústica. Os resultados experimentais serão apresentados e discutidos no capítulo 6.

Capítulo 5

Fusão de informação visual e acústica

O propósito deste capítulo é apresentar uma descrição detalhada do novo método para a identificação de espécies de pássaros que emprega a combinação de recursos visuais e sonoros. A abordagem primordial deste trabalho é a identificação de espécies de pássaros por meio de características visuais representadas na forma de imagens. Os capítulos anteriores discutiram alguns problemas inerentes à coleta destas imagens e a influência disso no processo de identificação.

A abordagem do problema por meio de áudio é justificada por Goëau et al. (2014) pelo fato das gravações de áudio serem mais fáceis de coletar do que as imagens. O fato de os pássaros estarem na maioria das vezes escondidos no cenário, por exemplo, no alto de uma árvore, voarem rapidamente, assustados com a presença humana, limita a coleta de imagens. Entretanto, alguns problemas são compartilhados, tanto para a aquisição de imagens quanto para a aquisição de gravações de áudio. Por exemplo, o caso do cenário pode ser comparado aos ruídos presentes no ambiente gerando, coletas sem restrições e similaridade visual comparada à similaridade acústica.

Considerando o cenário real, onde o número global de espécies de pássaros supera 9.000 espécies catalogadas, a escalabilidade das abordagens atuais é limitada. Desta forma, ainda é necessário o desenvolvimento de novos métodos de extração de características para fornecer informações adicionais e auxiliar o processo de classificação. Muitas pesquisas utilizam com sucesso recursos visuais ou acústicos de forma isolada para identificar automaticamente espécies de pássaros. Aproveitando o fato de que o pássaro pode ser identificado, tanto pela sua imagem quanto pelo seu canto, a fusão audiovisual atua na solução do problema com toda a experiência adquirida pelos métodos individuais e se apresenta como uma alternativa viável para a implementação.

A Figura 5.1 apresenta uma visão geral do método proposto para a fusão de informação visual e acústica. Além disso, destaca-se a principal inovação para o sistema de

identificação de espécies proposto: ele poderá ser aplicado em casos em que as imagens e sinais de áudio não possuem relação um-para-um. As próximas Seções apresentam detalhes em relação ao método proposto. O capítulo 6 e a seção 6.2.6 apresentam a aplicação do novo método ao conjunto de dados CUB 200-2011. Os resultados dos testes mostram que o algoritmo proposto consegue melhorar a taxa de identificação por meio de imagens.

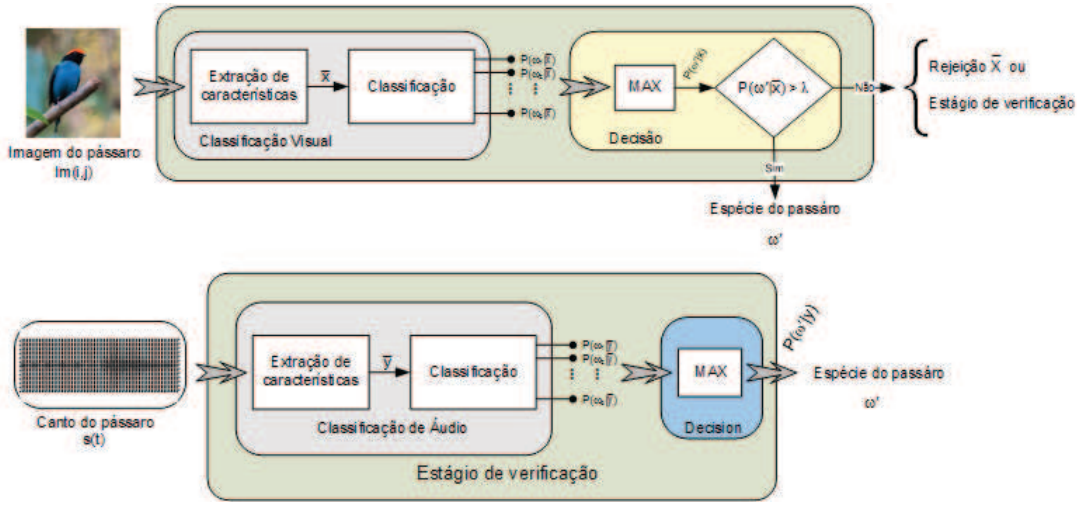


Figura 5.1: Visão geral do método para a classificação visual com opção de rejeição e verificação acústica.

5.1 Conjunto de dados

A etapa inicial para a implementação do novo método foi a preparação do conjunto de dados. Com base nas informações e as imagens disponibilizadas no conjunto CUB 200 2011, um subconjunto de áudios relacionados foi organizado. O nome taxonômico da espécie foi a chave para organizar espécies e ter acesso às amostras disponíveis no website do Xeno-Canto[1]. De acordo com Goëau et al. (2014) o acesso a toda a informação disponível sobre as espécies, na web, literatura, livros ou revistas, é realizado pelo nome taxonômico das espécies de plantas ou animais. Por exemplo, o nome taxonômico da classe *119 Field Sparrow* é *Spizella Pusilla*.

A ideia inicial era constituir o conjunto de áudio com as 200 espécies correspondentes às espécies disponíveis no conjunto CUB200 2011. Entretanto, algumas espécies, por serem mais raras ou de canto não harmonioso aos ouvidos humanos, possuem poucas amostras disponíveis. A única restrição de escolha foi considerar apenas o canto, e

[1]Mantido por aproximadamente 1.800 contribuintes ativos, disponibiliza mais de 190 mil gravações de mais de 9.000 espécies.

desconsiderar outros áudios, como da comunicação, do acasalamento, etc. Verificamos no decorrer do processo que essa limitação não restringiu significativamente a quantidade de classes.

Outras formas de aquisição de amostras foram iniciadas, porém, nenhuma delas tornou-se viável para o andamento do trabalho. Para a resolução do problema foi necessária a redução do número de classes, utilizando apenas as 50 espécies (a quantidade de amostras pertencentes a uma classe pode variar de 15 a 22 amostras por espécie). O subconjunto construído recebeu o nome de **CUB 50 Songs**[2]. A Figura 5.2 resume o processo de construção do conjunto de áudios.

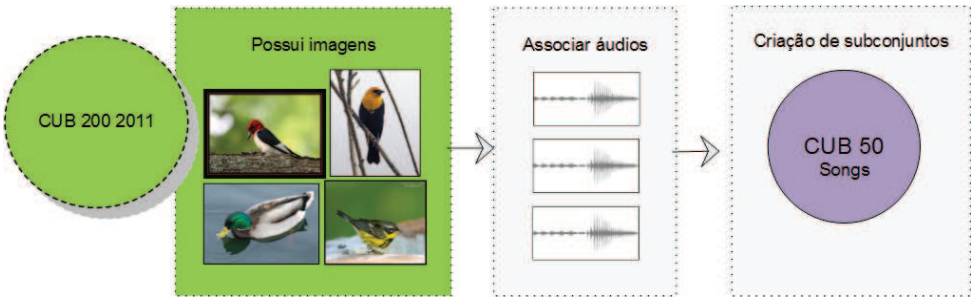


Figura 5.2: Visão geral do processo de criação do subconjunto de áudios CUB 50 Songs.

Os arquivos originais de áudio no formato *WAV*[3] possuem duração entre de 30 a 60 segundos. Estes arquivos foram segmentados manualmente para remover momentos de silêncio e ruídos indesejáveis das amostras. Desta forma, o conjunto original CUB 50 Songs derivou um novo conjunto com as amostras segmentadas. A Figura 5.3 apresenta a forma de onda do sinal de áudio do canto de um pássaro. Em seguida, a mesma gravação após a segmentação manual, processo o qual concatenou a informação relevante pertencente a uma amostra.

5.2 Extração de características

Por considerar duas formas de extração de características distintas, as subseções 5.2.1 e 5.2.2 detalham cada uma delas.

[2]Disponível em: http://www.ppgia.pucpr.br/ andreia

[3]WAV - abreviatura para *WAVEform audio format*, formato de arquivo de áudio para armazenamento.

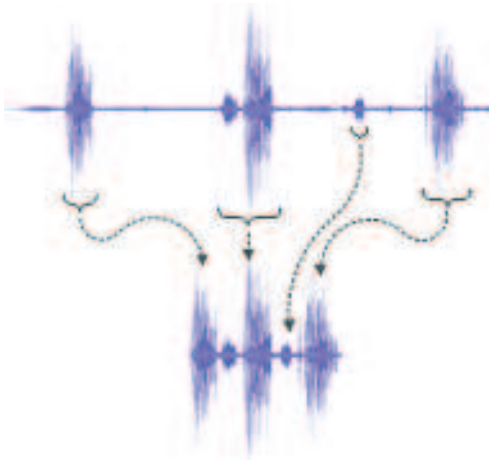


Figura 5.3: Exemplo de um sinal acústico do canto um pássaro. Exemplo superior: áudio original; Exemplo inferior: sinal pré-processado e os intervalos de silêncio removidos.

5.2.1 Extração de características visuais

A Figura 4.9 da seção 4.4.1.3 apresentada no capítulo 4 descreveu a formação do primeiro vetor de características visuais utilizado para a fusão audiovisual. Resumidamente, a detecção de pontos-chave é realizada pelo algoritmo SIFT; a definição do vocabulário é realizada por meio da atribuição dos descritores dos pontos para um conjunto pré-determinado (de 1.200 itens de conteúdo visual) a aplicação de um algoritmo de quantização vetorial, e a construção dos histogramas. O vetor gerado por esse método é de alta dimensionalidade contendo 24.000 características. A escolha desse vetor deve-se ao fato de representar uma grande quantidade de informação visual associada a uma amostra. Espera-se que dentre esse número elevado de características estejam presentes mais características discriminantes que possam diferenciar as espécies. Quanto à escolha das características visuais de forma, justifica-se pelos resultados individuais obtidos, eles são melhores quando comparados aos obtidos por meio de cores ou texturas quando consideradas apenas abordagens superficiais. Cabe salientar que no momento da construção do método apenas as estratégias superficiais estavam implementadas.

A Figura 4.11 da seção 4.5.1 apresentada no capítulo 4 descreveu a formação do segundo vetor de características visuais utilizado para a fusão audiovisual. A escolha desse vetor deve-se ao fato de que os melhores resultados em relação a taxa de correta classificação foram obtidas por esse vetor. Espera-se que esse resultado possa impactar

de forma positiva nos resultados finais da fusão audiovisual.

5.2.2 Extração de características acústicas

Para a extração de características acústicas as amostras de áudio são convertidas para um espectrograma. A representação espectral é obtida por meio da Transformada de Fourier (STFT) no espectro de frequências do sinal. A STFT de curta duração, em torno de 10 ms de duração por quadro, teve o eixo de frequência transformado para a escala de Mel, conforme (Costa et al., 2012). Essa escala relaciona as frequências físicas às frequências percebidas pelo sistema auditivo humano, sendo muito utilizada em aplicações de aprendizagem de máquina que envolve áudio, incluindo atividades de bioacústica e tem como objetivo analisar e modelar cantos de pássaros.

As características MFCCs foram extraídas por meio do *framework Music Analysis, Retrieval and SYnthesis for Audio Signals* - MARSYAS, proposto por Tzanetakis e Cook (2002). A Figura 5.4 apresenta a definição de quadros para um sinal acústico. Para cada amostra do conjunto de áudio (contendo 50 espécies de pássaros a CUB 50 Songs) foram extraídas 13 MFCCs[4] representadas ao longo do tempo por meio de média e desvio padrão. São extraídos quatro quadros gerando um vetor concatenado de 52 características para cada amostra do conjunto. A principal vantagem da utilização de MFCCs é o fato dos valores das características não serem correlacionados.

Figura 5.4: Exemplo de sinal acústico do canto um pássaro e a definição dos quadros para gerar o vetor de características. Quatro quadros de

[4]Algumas configurações relevantes: a) Tamanho da Janela: 512; b) Tamanho do salto: 512; c) Tamanho da memória (em análise de janela): 20; d) Tempo entre um quadro e outro: 10 milisegundos; e) total de quatro quadros.

5.3 Classificação

As características visuais e acústicas são submetidas ao classificador SVM com múltiplas classes. O vetor visual de alta dimensionalidade utilizou *kernel* linear. O vetor sonoro *Kernel* gaussiano e os parâmetros de custo e gama foram ajustados pela aplicação de validação cruzada 5 partes no conjunto de treinamento. A estratégia um-para-um é usada na classificação de múltiplas classes. A fusão da informação visuais e acústicas é implementada no nível de classificação, pois nem sempre é possível a mesma disponibilidade de amostras, ou seja, uma relação de imagem e áudio um-para-um. A seção 5.5 detalha o funcionamento dos classificadores individuais em seguida ao funcionamento do método proposto.

5.4 Mecanismo de rejeição

Para aprimorar o processo de classificação no reconhecimento de espécies de pássaros será incorporado o conceito de rejeição. Por meio da utilização de estratégias de rejeição é possível produzir sistemas classificadores mais confiáveis, além de melhorar os resultados medidos em relação à taxa de erro apresentada pelo classificador de base (Marini e Koerich, 2008). Em particular a esse trabalho, o mecanismo de rejeição possui a função adicional de possibilitar a fusão audiovisual nos casos em que a relação de imagem e áudio não é um-para-um.

A rejeição ocorre quando um padrão ambíguo, propenso a ser incorretamente classificado, é deixado de lado para uma posterior classificação. A decisão de aceitar ou rejeitar um exemplo é controlada por um limiar λ. Quando um exemplo possui medida de confiança $\geq \lambda$ ele é aceito, enquanto amostras que possuem nível de confiança $\leq \lambda$ são rejeitadas. Neste trabalho, dependendo do nível de confiança e da disponibilidade de dados de áudio, uma instância pode ser rejeitada ou enviada para uma fase de verificação que emprega as características acústicas.

Encontrar um valor ótimo para λ é o objetivo de um mecanismo de rejeição. Quando este valor é encontrado, o mecanismo de rejeição consegue rejeitar todos os exemplos incorretamente classificados pelo classificador e também aceitar todos os exemplos corretamente classificados. O valor de λ é obtido de sua variação entre 0 e 1. De modo que, 0 significa aceitar todas as amostras e 1 rejeitar todas as amostras.

5.5 Método para fusão de informação visual e acústica

Nesta seção é apresentada a metodologia proposta para a realização da fusão de informação visual e acústica. Esta metodologia considera o problema da classificação de espécies de pássaros como: dada uma imagem de um pássaro ou canto de um pássaro da mesma espécie é necessário atribuir uma classe entre um número amplo e fixo de possibilidades. De forma que, $\Omega = (\omega_1, \omega_2, \omega_c)$ seja o conjunto de todas as classes e $\omega' = arg\,max\, P(\omega_i|x)$ sendo que $i \in \Omega$, seja a classe atribuída pelo classificador.

O método para fusão de informação visual e acústica pode ser descrito em quatro etapas:

1. **Classificação visual** - Dada uma imagem, as características visuais são extraídas e organizadas em um vetor de características. Sendo um vetor $\bar{X}$ o classificador atribui uma probabilidade P para cada amostra de que ela possa pertencer a uma classe c, sendo $P(\omega_1|\bar{X})$, $P(\omega_2|\bar{X})\,...\,P(\omega_c|\bar{X})$ as possíveis probabilidades e classes. Em seguida, o operador Max escolhe a classe que oferece a maior probabilidade. Desta forma, a classe com maior probabilidade, ou seja ω' indica a espécie à qual o pássaro pertence. A Figura 5.5 exemplifica esta etapa. Esse procedimento é o mesmo utilizado na seção 6.2.4 para apresentar os resultados obtidos por meio de características de forma.

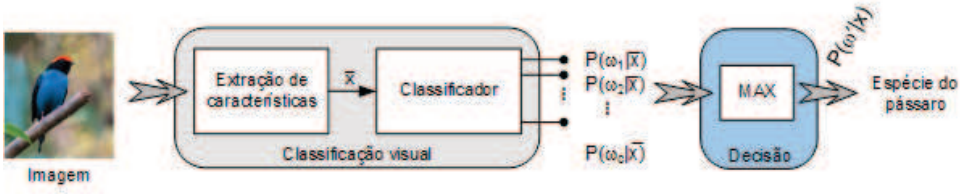


Figura 5.5: Visão geral do método para classificação visual.

2. **Classificação visual com a opção de rejeição** - A classificação visual é aprimorada com o emprego de um mecanismo de rejeição na saída do classificador. O conceito de rejeição admite que o classificador possa recusar uma amostra caso ele não esteja certo o suficiente sobre a classe a ser indicada. Neste caso, foram utilizadas as probabilidades atribuídas à classificação da espécie como parâmetro para implementar a rejeição de uma amostra. O valor da probabilidade da amostra deverá ser maior que um limiar λ para que ela seja rejeitada ou enviada ao estágio de verificação. A Figura 5.6 exemplifica o emprego desta etapa.

3. **Classificação acústica** - Dada uma amostra de áudio, as características acústicas

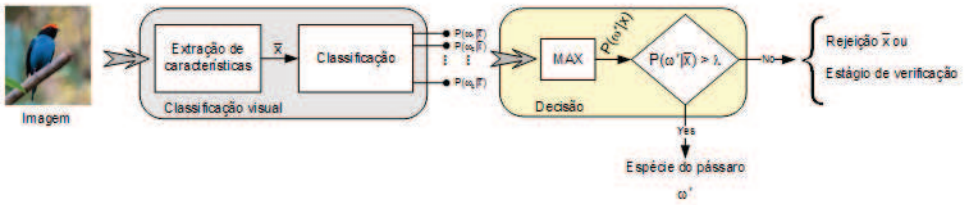


Figura 5.6: Visão geral do método para classificação visual com opção de rejeição.

são extraídas e organizadas em um vetor de características submetido ao classificador SVM. De posse de uma amostra de áudio $\bar{y}$ o classificador atribui uma probabilidade para cada amostra de que ela possa pertencer a uma classe ω_c de Ω. Em seguida o operador Max escolhe a classe que fornece a maior probabilidade $P(\omega'|y)$. A Figura 5.7 exemplifica esta etapa.

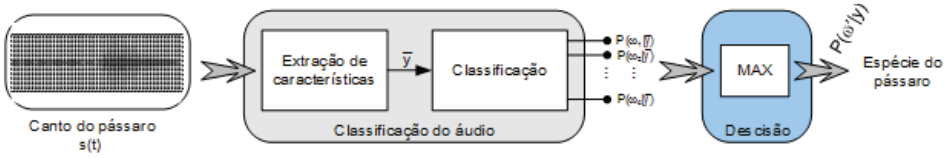


Figura 5.7: Visão geral do método para classificação acústica.

4. **Adição do estágio de verificação** - Dada uma amostra de imagem, dependendo do valor da probabilidade associada à classe, bem como da disponibilidade de amostras de áudio, uma amostra de imagem poderá ser rejeitada ou enviada para uma etapa de verificação, que emprega as características de áudio como auxiliar ao processo de classificação. Nesta etapa é onde ocorre a efetiva fusão de informação visual e acústica.

A fusão da informação visual e acústica é implementada no nível de classificação, sendo que dois esquemas de fusão são propostos:

1. Amostras rejeitadas na classificação visual são reclassificadas, considerando informações acústicas, se estas estiverem disponíveis. A decisão é tomada baseada unicamente no estágio de verificação;

2. Além das amostras rejeitadas na classificação visual poderem ser reclassificadas de acordo com a disponibilidade de informação acústica. A saída do classificador visual

e acústico pode ser combinada no nível de probabilidades, usando de regras, tais como, soma, produto e máximo.

A Figura 5.1 resume a abordagem proposta para a classificação das espécies de pássaros utilizando fusão de informação visual e acústica. A principal inovação tratada por esse método é a forma de como combinar informação visual e acústica em situações onde não há dados de áudio disponíveis em número suficiente para uma relação um-para-um com as amostras de imagens. De tal modo, para uma determinada instância, podemos ter apenas representação visual (imagem da espécie) ou a representação visual e acústica (canto da espécie). Em particular, para o conjunto de dados utilizado apenas um quarto das instâncias possui relacionamento um-para-um entre imagem e áudio. Portanto, a fusão é abordada no nível de pós-processamento, pois não é possível concatenar diretamente os vetores.

5.6 Considerações finais

Independentemente dos recursos utilizados, a tarefa de identificação de espécies de pássaros é altamente complexa e desafiadora. Cabe enfatizar que, até o presente momento, não foram identificados trabalhos similares na literatura que possam servir de comparação ao método de fusão de informação na identificação de pássaros. O capítulo 6 apresenta os resultados obtidos. O resultado do vetor de características visuais é superado por meio da fusão com informações acústicas. A partir dessa constatação, considera-se que a melhoria da abordagem proposta, otimizando tanto as características visuais quanto as acústicas, pode produzir resultados superiores aos obtidos por meio do classificador individual. Certamente, trabalhos futuros serão desenvolvidos visando ao aprimoramento do método de fusão audiovisual.

Capítulo 6

Resultados experimentais

Este capítulo apresenta os resultados dos experimentos realizados. A seção 6.1 descreve os resultados obtidos por meio dos conjuntos CUB 200, utilizada nos experimentos iniciais desta pesquisa. Em seguida, a seção 6.2 descreve os resultados obtidos com o conjunto CUB 200 2011. Além disso, a seção 6.2, agrega abordagens adicionais, sendo: resultados da combinação de classificadores na seção 6.3; análise de erros na seção 6.4. Por fim, a seção 6.5 discute e apresenta uma análise geral dos resultados obtidos.

A escolha de iniciar os experimentos com o conjunto CUB 200 deve-se ao fato de que alguns trabalhos na literatura reportarem resultados para esse conjunto de dados. Esses trabalhos foram utilizados como sistema de referência para os resultados preliminares dessa pesquisa. No entanto, em curto período de tempo, surgiram trabalhos que reportavam resultados para a CUB 200 2011 e foi possível estabelecer um novo sistema de referência. Além disso, o aumento da quantidade de imagens por classes na segunda versão do conjuntos de dados, seria fundamental para a realização dos novos experimentos.

Os resultados são expressos de forma progressiva inicialmente para 2 classes, 5 classes, 17 classes, 50 classes e o conjunto original com 200 classes. A forma de apresentação considera 3subconjuntos com menor número de espécies. Nesses casos, a problematização relacionada a granularidade fina, não fica tão evidente, pelo fato de que ainda é possível perceber algumas diferenças discriminantes entre as classes (espécies). Entretanto, para os conjuntos de 50 e 200 classes, a condição de granularidade fina está completamente presente, a qual exige o tratamento de diferenças sutis entre as classes. Isso implica diretamente nos resultados descritos e também foi objeto de estudo deste trabalho.

O Apêndice C apresenta outras formas de organização de subconjuntos de dados e discute impacto dos resultados obtidos por meio de subconjuntos de dados aleatórios (organização utilizada para apresentar os resultados no decorrer deste capítulo), conjuntos similares (forçando a presença de granularidade fina por meio de espécies pertencentes a

mesma família) e conjuntos não similares (sem qualquer similaridade visual), para 2, 5 e 17 classes.

6.1 Resultados para CUB 200

Esta seção tem como principal objetivo apresentar os resultados dos experimentos iniciais, conduzidos para avaliar isoladamente as características de cores e texturas. Todos os resultados descritos nesta seção são obtidos por meio do conjunto de dados CUB 200 que possui 6.033 imagens de 200 espécies. Para cada espécie, estão disponíveis de 20 a 40 amostras. Entretanto, em casos em que as classes possuem poucos exemplos são consideradas pelo menos 15 amostras de cada classe para a etapa de treinamento. Essa parametrização foi estabelecida pelo trabalho de Welinder et al. (2010) sendo seguida pelos demais trabalhos que reportam resultados para esse conjunto de dados.

6.1.1 Característica visual: Cor

Esta seção apresenta os resultados obtidos exclusivamente por meio de características de cor. O objetivo principal é avaliar o comportamento da característica cor em duas condições: extração de características da imagem inteira, ou seja, sem a aplicação de segmentação e, em seguida, com a aplicação de uma abordagem de segmentação baseada em cores. Adicionalmente, é realizada uma verificação da eficácia do método de segmentação proposto em relação às imagens do conjunto de dados.

Os vetores de características foram obtidos por meio de histogramas que geram uma representação compacta da imagem (cada *pixel* é associado a uma cor) em um número pré-definido de faixas (neste experimento 10 para cada canal que em seguida são concatenadas para formar o vetor, conforme foi descrito no capítulo 4 na subseção 4.4.1.1). Uma cor pode ser descrita de diferentes formas, dependendo do espaço de cor utilizado. Neste experimento são utilizados os espaços de cores HSV e RGB para imagens segmentadas e não segmentadas.

O primeiro experimento considera a imagem completa (a cor presente na imagem é extraída independente de pertencer ao pássaro ou ao fundo). O segundo experimento aplica a segmentação baseada em cores antes da extração de características e desta forma ocorre uma separação entra as cores do pássaro e as cores pertencentes ao fundo, possibilitando que as cores do pássaro participem do procedimento de extração de características. Em ambos os casos o vetor é composto pela concatenação dos três canais RGB ou HSV.

A Tabela 6.1 apresenta os resultados da abordagem baseada exclusivamente em

cores. O classificador utilizado é SVM com *kernel Radial Basis Function* e otimização para os parâmetros de custo e *gamma* por meio de *grid search* em relação ao conjunto de treinamento com aplicação de validação cruzada de cinco partições.

Tabela 6.1: Resumo da classificação de espécie de pássaros utilizando características de cores para 2,5,17 e 200 classes.

	Taxa de correta classificação (%)							
	Sem segmentação				Com segmentação			
Espaço de cor	2	5	17	200	2	5	17	200
HSV	83,82	47,02	25,05	8,17	92,64	48,34	25,63	8,60
RGB	73,53	39,07	16,96	4,16	77,94	40,39	18,48	6,86

Um dos objetivos desse experimento é avaliar o impacto da segmentação na classificação. Para mensurar a eficácia do método de segmentação proposto, as imagens segmentadas pelo algoritmo proposto são comparadas à segmentação aproximada disponibilizada juntamente com o conjunto de dados CUB 200. Como as imagens do conjunto possuem dimensões diferentes, um percentual de largura para a borda de verificação de cores foi definido na horizontal e na vertical para representar as cores do fundo. Testes foram realizados com percentuais de 2% a 10% de largura, independente das dimensões da imagem para um identificador do valor mais adequado para o problema. A Tabela 6.2 apresenta os resultados da avaliação do método de segmentação proposto, considerando 2% de largura de borda.

Tabela 6.2: Taxa de correta segmentação para o método de segmentação proposto em relação ao conjunto de teste de 200 espécies com um total de 3.033 amostras.

Espaço de cor	Taxa de segmentação (%)
RGB	71,00
HSV	75,00

A primeira conclusão obtida, por meio da análise dos resultados, aponta que as características de cor não são discriminativas o suficiente para lidar com um grande número de classes para o problema de classificação de espécies de pássaros. Quanto à segmentação, percebe-se um impacto favorável sobre as taxas de classificação, proporcionando um aumento de 8,82% na taxa de classificação de duas classes, mas esse impacto não é tão significativo quando o número de classes aumenta, atingindo 0,43 % para o conjunto de 200 classes. Além disso, quando se comparam os resultados alcançados pelos espaços de cor RGB e HSV, os melhores resultados foram obtidos utilizando o espaço de cor HSV. A diferença na taxa de classificação varia de 10,29% para duas classes até 1,74% para 200 classes.

A Tabela 6.3 mostra os resultados da segunda abordagem de classificação, onde as características extraídas de cada canal da imagem são utilizadas individualmente como vetor de características, ou seja os resultados obtidos com os vetores P_H, P_S, P_V. Apenas os resultados para o espaço HSV são apresentados nesta tabela, uma vez que esse espaço sempre obteve melhores taxas de correta classificação. Esses resultados são interessantes porque, nesse caso, usa-se menos informação para a classificação de espécies de pássaros do que a abordagem anterior, que utiliza um vetor de características que concatena três canais da imagem. No entanto, os resultados apresentados na Tabela 6.3 não são superiores aos resultados apresentados na Tabela 6.1.

Tabela 6.3: Resumo da classificação de espécie de pássaros, utilizando os canais H, S e V separadamente.

	Taxa de correta classificação (%)			
Canais do espaço de cor HSV	2	5	17	200
P_H	82,35	45,03	20,42	6,60
P_S	77,94	42,38	15,22	4,34
P_V	79,41	42,38	10,60	2,87

A classificação por meio de SVMs pode gerar estimativas de probabilidades a posteriori relacionadas ao grau de confiança de uma determinada amostra pertencer a uma classe. Essas probabilidades podem ser combinadas considerando as regras: Max (valor máximo), Sum (soma), $Prod$ (produto), $WSum$ (soma ponderada) e $WProd$ (produto ponderado).

A Tabela 6.4 mostra os resultados obtidos através da combinação da saída dos classificadores para os canais H, S e V, por meio das regras já citadas. Os resultados mostrados na Tabela 6.4 quando comparados aos apresentados na Tabela 6.1 apontam que o segundo método de classificação só supera o primeiro método de classificação quando são consideradas 5 classes. Para todos os outros casos, a primeira abordagem de classificação alcança os melhores resultados. Isto indica, a segunda conclusão para este experimento: a separação do vetor de características e a combinação dos classificadores pode não ser adequada para lidar com o problema de classificação de espécies quando descritores exclusivos de cores são abordados.

Nos experimentos que podem avaliar o impacto do método de segmentação por cor em imagens, percebe-se que existe um impacto da segmentação nos resultados de classificação. Ambos os espaços de cor HSV e RGB, conseguem segmentar corretamente, mais de 70% dos *pixels*. Contudo, e o impacto em relação a classificação de espécies de pássaros não pode ser considerado relevante, devido à baixa variação de 8,82% para o subconjunto de 2 classes e 0,43% para o conjunto de 200 classes.

Tabela 6.4: Resumo da classificação de espécie de pássaros utilizando fusão dos canais HSV no nível de regras de fusão.

	Taxa de correta classificação (%)			
	Número de classes			
Regra de fusão	2	5	17	200
Max	85,29	47,68	19,65	6,76
Sum	86,76	49,01	22,16	7,16
$Prod$	88,24	49,67	22,54	7,25
$WSum$	89,71	51,66	23,89	7,59
$WProd$	91,18	51,66	23,70	8,03

Essa variação, sugere que a segmentação não desempenha um papel importante no problema de identificação de espécies, em particular quando o número de classes é alto. Embora essa conclusão seja válida somente para os recursos de cores empregados neste experimento, não é possível estender a outros tipos de descritores de características. Além disso, pode-se concluir que a abordagem de segmentação proposta não apresenta uma boa escalabilidade, devido o resultado obtido para o conjunto de 200 classes.

Finalmente, a análise geral dos resultados indica baixo impacto do processo de segmentação e na identificação de espécies de pássaros. Por outro lado, o processo de segmentação implica em alto custo adicional à identificação. Por esses motivos, os experimentos seguintes substituem a etapa de segmentação pela utilização de imagens recortadas de acordo com informações de caixa delimitadora em relação ao objeto de interesse disponibilizada juntamente com o conjunto de dados original. Apesar de neste caso não haver separação do objeto de interesse e o fundo da imagem, tem-se a garantia de que o objeto de interesse esta presente na imagem e também que menos informação da imagem é descartada.

6.1.2 Característica visual: Textura

Esta seção apresenta os resultados obtidos por meio dos descritores de textura baseados em LBP. Foram realizados três experimentos considerando os operadores: $LBP^{u2}_{P,R}$, $LBP^{ri}_{P,R}$ e $LBP^{riu2}_{P,R}$.

O primeiro experimento foi realizado em imagens coloridas codificadas no formato RGB. A análise dos canais de cores é realizada individualmente para cada canal a fim de que não ocorra perda de informação da relação entre canais de um mesmo *pixel*. O resultado de LBP_{RGB} é apresentado por meio da média do processamento dos três canais. Tal espaço de cor em conjunto com o operador $LBP^{u2}_{P=8,R=1}$ apresenta o melhor desempenho utilizando duas classes, porém, nos demais casos fica evidenciado que em um

problema com um número maior de classes, a taxa de acerto é decrementada, variando de 82,3% a 7,4%. Considerando 200 classes, a taxa de acerto neste espaço de cor e para o operador indicado é de 7,5%. Observa-se que a melhor taxa de correta classificação foi obtida com uma vizinhança de 12 *pixels* que gerou um vetor de 36 características. Na sequência, o operador $LBP^{u2}_{P,R}$ é aplicado em imagens convertidas para o espaço de cor HSV, canal H representado por LBP_H e para escala de cinza representada por LBP_{GRAY}. A Tabela 6.5 apresenta os resultados obtidos por meio do operador $LBP^{u2}_{P,R}$ em diferentes parâmetros de P e R.

Tabela 6.5: Resumo da classificação de espécies de pássaros utilizando e o operador $LBP^{u2}_{P,R}$ para 2,5,17 e 200 classes.

Taxa de classificação (%)								
	$LBP^{u2}_{P=8,R=1}$				$LBP^{u2}_{P=12,R=1.5}$			
Espaço de cores	2	5	17	200	2	5	17	200
LBP_{GRAY}	77,9	48,3	30,4	8,0	79,4	38,8	28,7	9,0
LBP_{RGB}	82,3	42,4	25,4	7,44	85,3	45,0	27,7	7,5
LBP_H	76,5	42,4	24,7	9,0	79,4	54,3	24,5	8,8

A primeira conclusão obtida para este experimento refere-se à similaridade entre os resultados obtidos em imagens em escala de cinza ou para o canal H. O conjunto de dados com imagens em escala de cinza representadas pelo vetor LBP_{GRAY} quando submetidas ao classificador SVM atingiram 9% de correta classificação.

O segundo experimento, considerando padrões locais binários, utilizou o operador $LBP^{ri}_{P,R}$ invariante à rotação e escala, uma condição importante para o processo de extração de características. Entretanto, neste experimento observa-se que a condição não impactou na melhora dos resultados descritos na Tabela de 6.6. Um forte indício que pode explicar esses resultados deve-se ao fato de todas as imagens apresentarem certo padrão (em geral a imagem é coletada do melhor ponto de vista do observador humano). A rotação não adiciona significativa informação adicional e os resultados se apresentam menores do que o processo LBP normal.

Tabela 6.6: Resumo da classificação de espécie de pássaros utilizando o operador $LBP^{ri}_{P,R}$

Taxa de classificação (%)								
	$LBP^{ri}_{P=8,R=1}$				$LBP^{ri}_{P=12,R=1.5}$			
Espaço de cores	2	5	17	200	2	5	17	200
LBP_{GRAY}	76,5	39,7	21,8	7,4	72,1	46,4	26,6	1,4
LBP_{RGB}	74,5	42,8	22,9	7,2	71,1	38,9	21,5	1,9
LBP_H	75,0	45,7	24,9	8,5	76,5	39,7	20,4	3,3

O terceiro experimento realizado com o operador $LBP^{riu2}_{P,R}$ apresenta os resultados

descritos na Tabela 6.7. A taxa de classificação apresentada pelo descritor aponta tal descritor como menos influenciado pelos parâmetros adotados para P e R. Em todos os experimentos realizados os melhores desempenhos foram obtidos pelo operador $LBP^{u2}_{P,R}$, seguidos dos operadores $LBP^{ri}_{P,R}$ e $LBP^{riu2}_{P,R}$ respectivamente.

Tabela 6.7: Resumo da classificação de espécie de pássaros utilizando o operador $LBP^{riu2}_{P,R}$

Taxa de classificação (%)								
	$LBP^{riu2}_{P=8,R=1}$				$LBP^{riu2}_{P=12,R=1.5}$			
Espaço de cores	2	5	17	200	2	5	17	200
LBP_{GRAY}	61,8	39,1	20,8	5,8	64,7	37,1	23,7	5,5
LBP_{RGB}	71,6	43,7	20,4	5,5	64,2	37,1	19,2	5,7
LBP_{H}	69,1	34,4	20,6	6.7	72,1	33,8	17,3	6,5

A Figura 6.1 apresenta uma comparação das taxas de classificação obtidas por meio do classificador SVM em relação aos operadores, parâmetros e espaços de cores utilizados. Analisando a Figura percebe-se que os parâmetros $LBP_{P=8,R=1}$ apresentaram melhor desempenho quando comparados a $LBP_{P=12,R=1.5}$. O primeiro conjunto quando aplicado ao $LBP^{ri}_{P=8,R=1}$ apresentou melhores taxas de acerto para todos os espaços de cores. Também em outros dois casos, quando consideradas 200 classes, LBP_H obteve por meio de $LBP^{u2}_{P=8,R=1}$ 9% e por meio de $LBP^{riu}_{P=8,R=1}$ 6,7%. Com relação ao segundo conjunto observa-se os melhores desempenhos associados aos operadores $LBP^{u2}_{P=12,R=1.5}$ e $LBP^{u2}_{P=12,R=1.5}$.

De forma geral, nos resultados obtidos com as características de textura, observa-se que o operador LBP possui um bom poder de discriminação quando aplicado ao problema de classificação de espécies de pássaros. Para fins de comparação, os experimentos aqui descritos apresentam desempenho médio ligeiramente superior quando comparados aos resultados apresentados pela característica de cor na seção 6.1.1. Os resultados obtidos experimentalmente em uma abordagem incremental demostraram que é possível atingir resultados similares por meio de texturas obtidas de imagens em escala de cinza ou coloridas. Ainda, no caso do conjunto de características baseado em escala de cinza, observou-se o bom desempenho do operador LBP em sua concepção original. Embora os resultados obtidos sejam interessantes, apontam que somente os descritores de textura LBP não oferecem informações suficientes para classificação das espécies de pássaros.

A comparação deste estudo com outros trabalhos disponíveis na literatura possui certo grau de dificuldade, devido ao uso de diferentes protocolos experimentais. Até onde esta pesquisa alcançou não foram encontrados trabalhos que utilizem o operador LBP e seus descritores para identificação de espécies. Dessa forma, entende-se que um novo sistema de referência foi estabelecido. Uma comparação indireta pode ser estabelecida

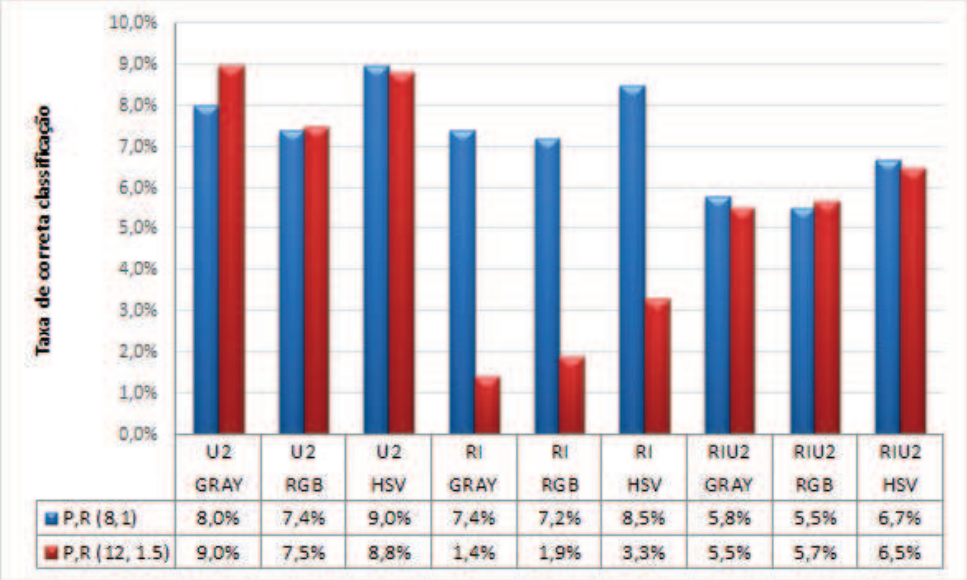


Figura 6.1: O eixo Y do gráfico apresenta a taxa de classificação obtida por meio do classificador SVM em relação aos operadores, parâmetros e espaços de cores descritos no eixo X.

com outros trabalhos que utilizam o conjunto de dados CUB 200 e o classificador SVM. Desse ponto de vista, as taxas de acerto obtidas pelos trabalhos relacionados, comparadas as obtidas neste estudo, indicam que as características extraídas por meio de textura de imagens são uma alternativa interessante para a solução do problema de identificação de espécies de pássaros. A principal vantagem é não necessitar da etapa de segmentação, proporcionando ao processo menor tempo de processamento e complexidade.

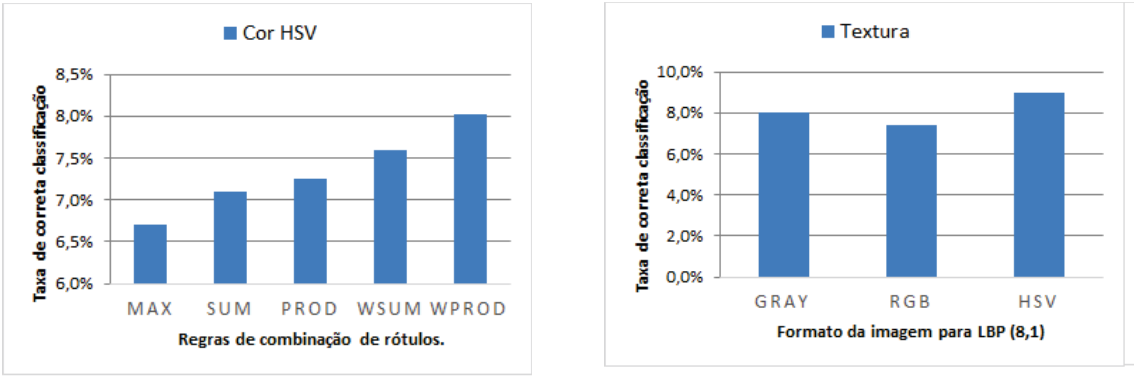
6.1.3 Considerações para cor e textura

Os experimentos realizados conseguem avaliar o desempenho de abordagens baseadas exclusivamente em cor e texturas. As cores presentes na imagem e o descritor de textura LBP são avaliados em condições representativas de aplicações reais de classificação de imagens sem restrições.

Por meio dos resultados experimentais pode-se destacar algumas importantes conclusões, relacionadas a cores e texturas. Tais conclusões orientam os experimentos descritos na próximas seção. Embora os resultados obtidos sejam interessantes, apontam para ambos os casos (cores e texturas) que os descritores isolados não oferecem informações suficientes para classificação das espécies de pássaros. As características de cor não são discriminativas o suficiente para lidar com um problema que possui um grande número de classes, particularmente com a evidência da granularidade fina. Por outro lado, observa-

se que as características de textura extraídas por meio do operador LBP apresentam-se mais discriminantes para o caso de um problema com o número de classes elevado, mas não discriminante o suficiente para resultados significativos (similares ao Estado da Arte) para lidar com problemas de granularidade fina.

A Figura 6.2 apresenta uma comparação direta entre as características de cores e texturas, considerando o conjunto de imagens completo com 200 classes. As melhores taxas de correta classificação para descritores de cor são obtidas por meio do uso dos procedimentos: conversão para o espaço de cor HSV; aplicação de regras de fusão de rótulos da saída ao classificador SVM. No entanto, para descritores de textura, a imagem é convertida para diferentes espaços de cores, mas os resultados do classificador individual e o operador $LBP^{u2}_{P=8,R=1}$ já indicam melhor desempenho, sem a aplicação da fusão de rótulos na saída do classificador SVM.



(a) Melhores resultados para cor.

(b) Melhores resultados para textura.

Figura 6.2: O eixo Y do gráfico apresenta a taxa de classificação obtida por meio do classificador SVM. O eixo X apresenta os melhores resultados para cor e textura para 200 classes.

6.2 Resultados para CUB 200 2011

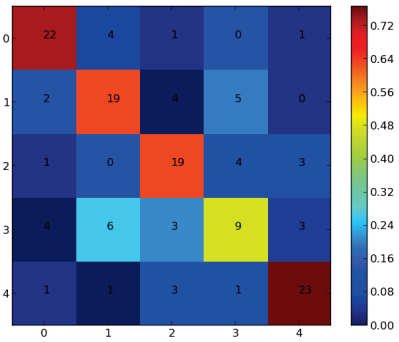
Nesta seção são apresentados os resultados dos experimentos realizados para identificação de espécies de pássaros em relação ao conjunto CUB 200 2011. Cabe enfatizar que a partir desta seção em todos os experimentos são consideradas imagens recortadas por uma caixa delimitadora e não imagens segmentadas. Alguns dos experimentos descritos na seção 6.1, considerando cor e textura são estendidos ao novo conjunto.

6.2.1 Característica visual: Cor

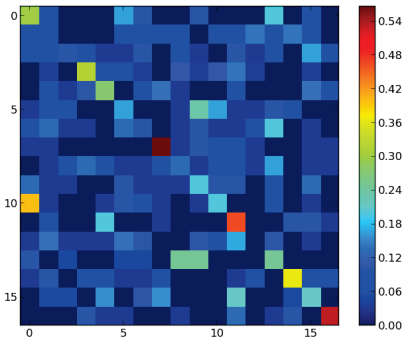
A característica cor foi discutida na seção 6.1.1. A Tabela 6.8 apresenta os resultados obtidos exclusivamente por meio da característica cor em imagens recortadas pela caixa delimitadora. O vetor utilizado foi obtido por meio da saída concatenada de um histograma de cores de 10 faixas, em cada canal do espaço de cor utilizado.

Tabela 6.8: Resumo da classificação de espécie de pássaros utilizando características de cores para o conjunto CUB 200 2011 e o classificador SVM.

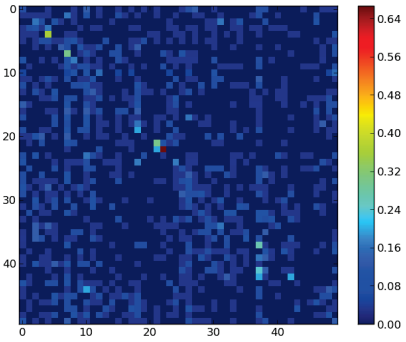
Taxa de correta classificação (%)		
Número de classes	Cor_{RGB}	Cor_{HSV}
2 classes	66,66	81,66
5 classes	27,33	66,18
17 classes	13,43	25,79
50 classes	3,85	9,34
200 classes	1,82	3,88



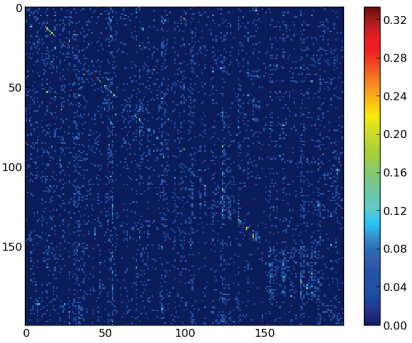
(a) Matriz para 5 espécies (66,28%).



(b) Matriz para 17 espécies (25,79%).



(c) Matriz para 50 espécies (9,34%).



(d) Matriz para 200 espécies (3,88%).

Figura 6.3: Matrizes de confusão resultantes da aplicação da característica Cor_{HSV}.

As matrizes de confusão relativas ao espaço de cor HSV, correspondentes a identificação exclusivamente baseada em cores são apresentadas na Figura 6.3. É possível

Tabela 6.9: Resumo da classificação de espécie de pássaros utilizando o operador $LBP^{u2}_{P,R}$.

	Taxa de correta classificação (%)		
Número de classes	LBP_{RGB}	LBP_{GRAY}	LBP_H
2 classes	95,00	83,33	75,00
5 classes	53,23	52,51	53,23
17 classes	26,44	47,54	29,21
50 classes	6,62	5,74	7,22
200 classes	5,07	4,84	3,69

constatar nestas matrizes a dificuldade do classificador em lidar com o problema utilizando exclusivamente cores. As melhores taxas de acerto encontra-se entre 81,66% para 2 classes e 3,88% para 200 classes. Tais valores são considerados baixos para grupos com espécies quando a condição de granularidade fina não é tão evidente. Isso indica uma forte evidência que os descritores de cor podem não obter resultados satisfatórios em larga escala. Novamente os resultados obtidos por meio dos descritores puros de cor não fornecem informações suficientes para a identificação das espécies de pássaros. Esta constatação auxilia a responder a uma das hipóteses iniciais deste trabalho que será tratada no capítulo 7.

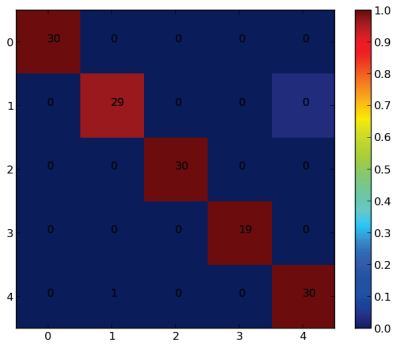
6.2.2 Característica visual: Textura

A característica de textura refere-se a um padrão visual que possui homogeneidade. De acordo com os resultados descritos na seção 6.1.3 tais atributos conseguem ser mais discriminantes do que o uso de cores para identificação de espécies. Nesta seção, apenas as melhores práticas percebidas na aplicação de descritores de textura no conjunto CUB 200 foram repetidas para CUB 200 2011. Um exemplo é a alteração da forma de obter resultados para o espaço RGB. A média dos canais utilizada para avaliar os resultados no conjunto CUB 200 não se mostrou eficiente pois, de certa forma, reconstrói a correlação entre os canais. Os resultados descritos nesta seção são obtidos por meio da separação dos canais RGB, mas apenas o canal que apresentou melhor taxa de correta classificação é considerado. Outro exemplo é a aplicação do descritor $LBP^{u2}_{P,R}$ o qual se apresentou mais adequado à identificação de espécies.

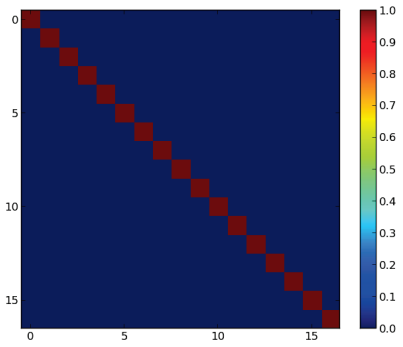
A Tabela 6.9 apresenta os resultados para o operador $LBP^{u2}_{P,R}$ em três casos considerados. Os resultados indicam que o descritor LBP também apresenta superioridade em relação a característica de cor quando aplicado ao problema de identificação de espécies. Entretanto, a análise do conjunto de resultados permite concluir que o descritor de textura não fornece informação suficiente para a identificação de espécies de pássaros.

O conjunto de 200 classes apresenta 5,07% de taxa de correta classificação obtido por meio das imagens no formato RGB. A superioridade de desempenho do vetor extraído por meio do canal H do espaço HSV não continuou evidente como nos experimentos realizados com o conjunto CUB 200. O vetor LBP_H destacou-se em apenas dois casos: 29,21% para conjuntos de 17, e 7,22% para o conjunto de 50 classes. Essa constatação pode indicar que variações de iluminação não geram impacto significativo na classificação de espécies. Cabe enfatizar que o uso de espaços de cores diferente do RGB podem ajudar a entender o comportamento do sistema em condições em que não existe variação de iluminação. Por exemplo, o espaço de cor HSV separa as informações de cromaticidade e intensidade, e neste caso, apenas a primeira foi considerada.

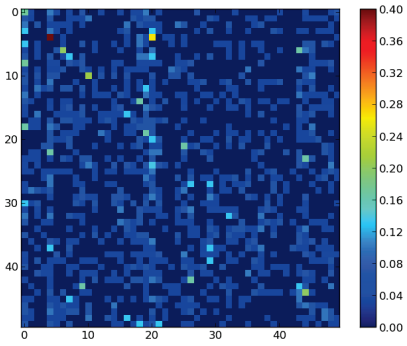
A Figura 6.4 apresenta as matrizes de confusão para o descritor LBP aplicado em imagens coloridas no formato RGB. Essas matrizes são resultantes da aplicação do classificador SVM conforme descrito no capítulo 4 na seção 4.4.1.4.



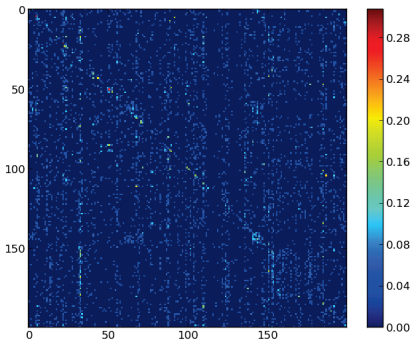
(a) Matriz para 5 espécies (53,24%).



(b) Matriz para 17 espécies (26,44%).



(c) Matriz para 50 espécies (6,62%).



(d) Matriz para 200 espécies (5,07%).

Figura 6.4: Matrizes de confusão resultantes da aplicação da característica de textura e operador LBP_{RGB}.

6.2.3 Considerações para cores e texturas

A extração de descritores de textura em imagens coloridas ou escala de cinza tem sido uma abordagem popular para a análise da textura. Uma das principais maneiras de analisar texturas em imagens coloridas é dividir o sinal de cor em componentes de luminância e crominância para processá-los separadamente (Zhu et al., 2011), (Zhu et al., 2010). De forma geral, a extração de descritores de textura, quando a informação de cor está presente, mostrou resultados melhores, quando a condição de granularidade fina é efetiva. Porém, ocorrem alguns casos, dependendo do tamanho do conjunto de imagens considerado, e especificamente para o subconjunto de 50 espécies, em que os descritores de textura coletados em escala de cinza são muito similares aos demais coletados em diferentes espaços de cores. A Figura 6.5 apresenta comparação de resultados para a relação cor e escala de cinza.

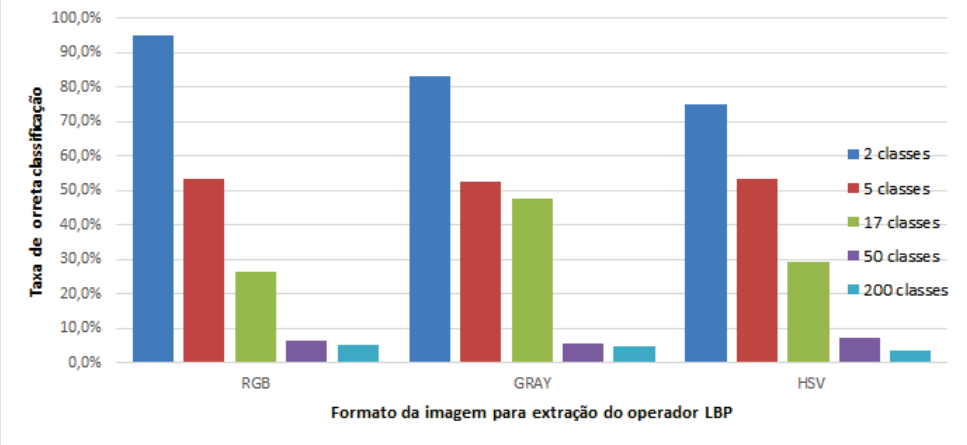


Figura 6.5: Visão geral dos resultados obtidos por meio de textura em imagens coloridas ou em escala de cinza.

A Figura 6.6 compara, indiretamente, os resultados obtidos por meio de características de cor e textura para os conjuntos CUB 200 e CUB 200-2011. Os resultados indicam que a metodologia utilizando o LBP obteve os melhores resultados em ambos os conjuntos de dados. A principal explicação para isso deve-se a constatação de que os descritores de textura conseguem ser mais discriminantes do que os descritores de cores para problemas de granularidade fina.

É interessante mencionar que os resultados obtidos experimentalmente reconstroem o cenário da percepção humana de padrões. De acordo com (Mäenpää et al., 2004), a investigação sobre o sistema visual humano tem proporcionado muitas evidências de que a percepção humana de padrões não está relacionada diretamente à cor presente em uma

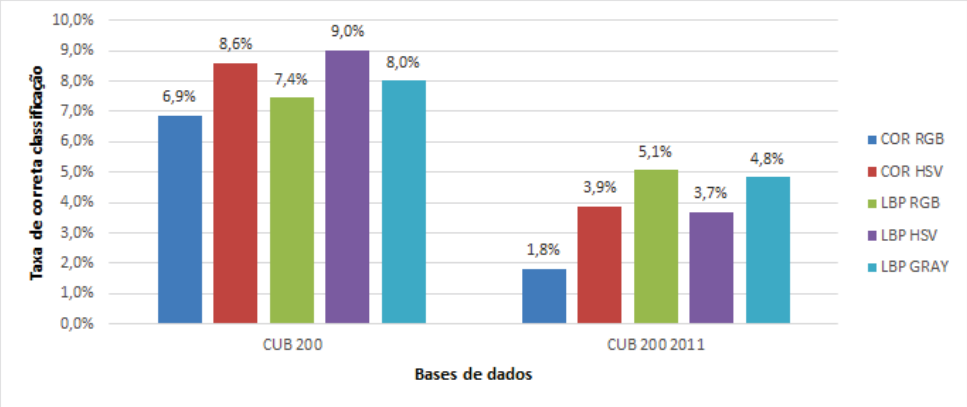


Figura 6.6: Visão geral da taxa de correta classificação para 200 classes.

imagem, mas que a percepção está mais relacionada à presença ou ausência de uma cor predominante, ou no grau de pureza de uma cor. Isso indica, que de forma geral, as pessoas usam informações de cor para julgar semelhanças. Cabe lembrar que, a constituição básica da representação de textura estrutural situa-se na orientação e regularidade de ocorrências de um determinado padrão.

6.2.4 Característica visual: Forma

O algoritmo SIFT adota uma estratégia que filtra sucessivamente uma imagem a fim de obter por um processo rápido a extração de pontos-chaves e garantir a invariância das características locais em escala. Podemos considerar esse algoritmo como descritor de forma, pois caso os pontos-chaves, e seus respectivos descritores sejam plotados, podemos observar a representação gráfica da forma do objeto pertencente a imagem. Os pontos-chaves podem ser organizados para produzir um vetor de características de tamanho fixo. O método de vetor de pontos-chave (Bok) é aplicado por meio de um vetor de quantização de descritores, para parte de uma imagem ou a imagem inteira, produzindo um vocabulário. Um vetor de pontos- chave pode gerar um histograma do número de ocorrências de determinados padrões visuais presentes em uma imagem. Os resultados obtidos por meio do algoritmo SIFT+BoK e suas variações são apresentados na Tabela 6.10.

Algumas variações do tamanho do vocabulário e regiões coletadas da imagem foram adotadas para quatro experimentos realizados. O primeiro experimento utiliza um vocabulário de 600 pontos chave, coletados na imagem inteira. Essa combinação apresentou bom desempenho apenas para o subconjunto de 2 classes.

Tabela 6.10: Resumo da classificação de espécie de pássaros por meio de variações SIFT+BoK.

	Taxa de correta classificação (%)			
Base	600 imagem inteira	600 [2+4]	1200 imagem inteira	1200 [2+4]
2 classes	88,33	85	85	78,33
5 classes	52,51	61,84	65,46	55,39
17 classes	28,14	43,28	37,10	43,07
50 classes	7,90	20,27	15,74	20,67
200 classes	6,38	18,24	13,23	16,07

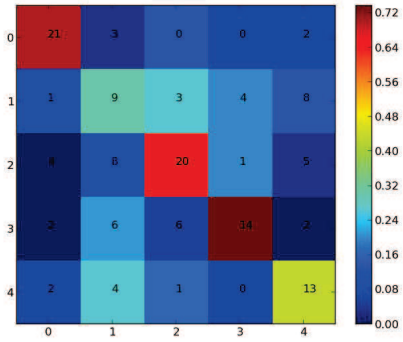
O segundo experimento aumentou o vocabulário para 1200 itens obtidos da imagem inteira. Os resultados foram melhores, porém, apenas nos subconjuntos de 5 e 17 classes. O terceiro experimento considera um vocabulário de 600 itens e concatenação de duas coletas para uma imagem (a imagem é dividida em 2x2 regiões e em seguida 4x4 regiões). A de taxa de correta classificação para tais parâmetros se destacou em relação às demais variações, obtendo 18,24% quando consideradas 200 classes. Acredita-se que uma vantagem dessa abordagem deve-se ao fato dos vetores de características gerados serem invariantes à translação, escala, rotação, mudança de iluminação, ruído na imagem e pequenas mudanças de perspectiva. Além disso, que o tamanho do vocabulário, com 600 itens, conseguiu representar bem algumas características visuais das imagens obtidas em ambientes sem restrições.

O quarto experimento adotou o mesmo procedimento de coleta, mas dobrou o número de itens do vocabulário. Nestas condições, os resultados são muito similares aos obtidos no experimento três. Entretanto, o custo computacional e dificuldade de manipulação de vetores de alta dimensionalidade torna essa abordagem mais trabalhosa quando comparada à anterior.

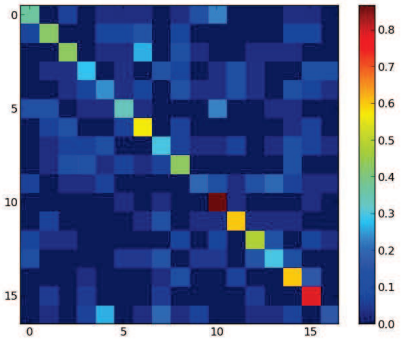
A Figura 6.7 apresenta as matrizes de confusão para o algoritmo SIFT+BoK. As características locais geradas pelos pontos-chave e abordagem BoK atingem melhores taxas de correta identificação utilizando o classificador SVM quando comparados às abordagens de cores e texturas, nos mesmos conjuntos de dados com todos os fatores de variabilidade já discutidos.

Em muitos trabalhos da literatura o algoritmo SIFT é utilizado para a extração de características (pontos-chave na imagem). Entretanto, a maioria desses trabalhos utiliza o conjunto CUB 200. Por esse motivo, apenas duas comparações podem ser realizadas.

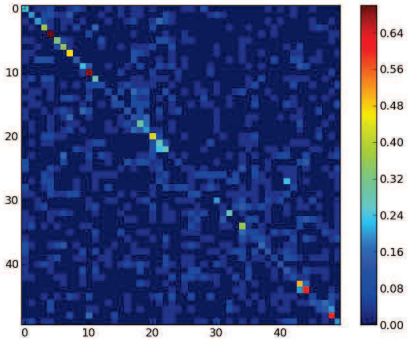
A primeira comparação pode ser realizada com os resultados obtidos por Wah et al. (2011). Neste caso, são utilizadas no máximo 52 amostras por classe, imagens inteiras no formato RGB, o algoritmo SIFT, construção de histograma de cores por meio de vetor de quantização e um SVM linear. O resultado obtido para a taxa de correta classificação



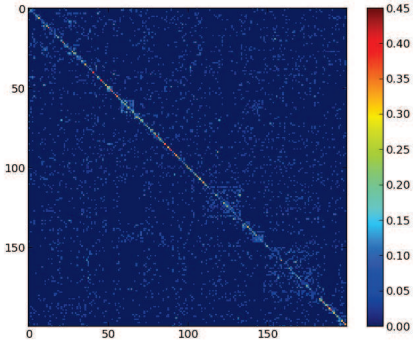
(a) Matriz para 5 espécies (61,84%).



(b) Matriz para 17 espécies (43,28%).



(c) Matriz para 50 espécies (20,27%).



(d) Matriz para 200 espécies (18,24%).

Figura 6.7: Matrizes de confusão resultantes da aplicação do algoritmo SIFT+BoK.

foi de 10,3%. Quando considerado o *ground truth* da localização de partes do pássaro presente na imagem foi possível alcançar 17,3% de taxa de correta classificação para o conjunto completo de 200 classes.

A segunda comparação é feita com o trabalho proposto por Zhang et al. (2012). O resultado de 14,14% é indicado como padrão para a combinação de ferramentas disponíveis por meio da biblioteca VLFeature (Vedaldi e Fulkerson, 2010) (SIFT + BoK + Histograma + SVM). Zhang et al. (2012) propõem um método que adiciona a normalização de poses por meio de um *framework* conhecido como *Poselet*. Em seguida, são realizados agrupamentos referentes às poses, considerando funções de similaridade e medidas de distância. Um vocabulário de 1024 itens é utilizado. O resultado obtido para o conjunto de 200 classes é de 24,21%.

De forma geral, os resultados obtidos por meio da aplicação SIFT+Bok são satisfatórios pela simplicidade da abordagem. A taxa de correta classificação de 18,69% é superior à primeira comparação que utiliza informação adicional de partes. Além disso, é superior ao valor de referência indicado pela biblioteca utilizada com algumas alterações de parâmetros e sem a adição de métodos computacionalmente caros.

6.2.5 Rede neural convolucional

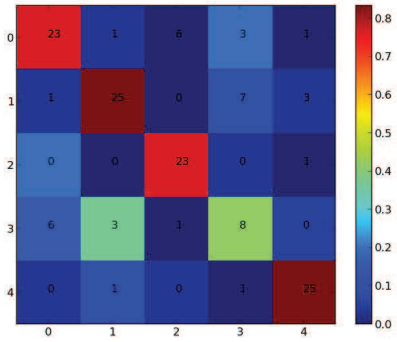
As redes neurais convolucionais (CNNs) têm alcançado resultados impressionantes em tarefas desafiadoras no reconhecimento de imagem e detecção de objetos (Krizhevsky e Hinton, 2009), (Krizhevsky et al., 2012), (Ciresan et al., 2012), aumentando significativamente o interesse da comunidade nesses métodos. As abordagens propostas para a aprendizagem de características são adequadas para trabalhar com imagens de alta resolução em larga escala. A ideia principal deste experimento é verificar o comportamento da abordagem quando não existe restrição para as imagens pertencentes aos conjuntos de dados e também quando o conjunto não é de larga escala.

Para esse experimento, foi necessário padronizar o tamanho das imagens, para fins de adaptação à biblioteca utilizada. As imagens foram redimensionadas em 64x64, mantendo os 3 canais de cores e formato RGB. Desta forma, uma imagem gera 12.228 entradas (64x64*3) para a rede e possui como saída o número de classes consideradas. Dada uma imagem de entrada um tamanho fixo de borda pode ser desconsiderado para gerar as novas subimagens. Neste caso, uma borda de 4 *pixels* foi removida, implicando que o tamanho das subimagens será 9408 entradas. As CNNs foram treinadas em uma GPU NVIDEA 630, utilizando a biblioteca Cuda Convnet[1]. O tempo de execução variou de alguns minutos para 2 classes (convergindo com 30 ciclos) a um hora para o conjunto de 200 classes (convergindo com 500 ciclos).

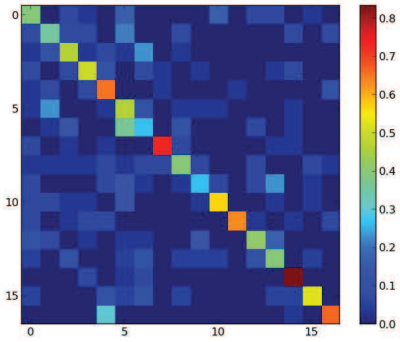
Os experimentos inicias realizados com imagens de 32x32 quando comparados aos experimentos realizados com imagens de 64x64 indicaram resultados inferiores de taxa de correta classificação. Devido a limitações de hardware não foi possível investigar adequadamente quanto o tamanho da imagem de entrada pode influenciar nos resultados. Ou se essa situação deve-se ao fato do conjunto não ser de larga escala.

A Figura 6.8 apresenta as matrizes de confusão obtidas com a aplicação de CNN ao conjunto CUB 200 2011. O subconjunto de 5 classes obtém 74,82% de taxa de correta classificação. O subconjunto de 17 classes obteve 50,96%. Com o aumento da quantidade de classes os resultados são melhores que os obtidos por meio de abordagens superficiais, sendo, 30,88% para 50 classes e 23,5% para 200 classes. De acordo com os resultados obtidos, verifica-se que essa abordagem pode funcionar de maneira satisfatória, com geração de resultados confiáveis de maneira automática, para identificação de espécies por meio de imagens. A análise dos experimentos demonstra que o modelo proposto por meio de aprendizagem de características apresenta melhor resultado do que outros métodos da literatura, sem necessitar de uma etapa prévia de extração de características. Contudo,

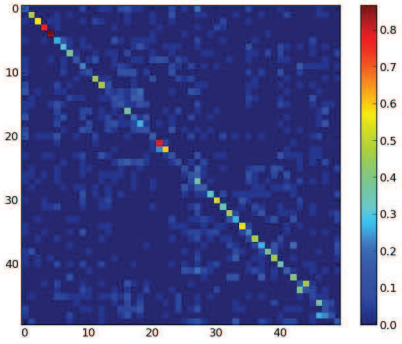
[1]http://code.google.com/p/cuda-convnet/



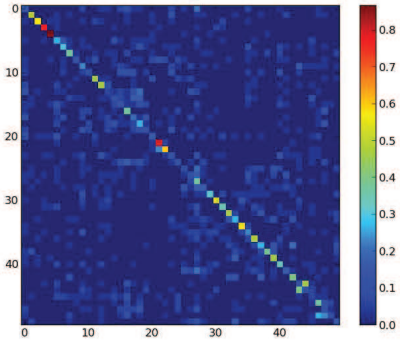
(a) Matriz para 5 espécies (74,82%).



(b) Matriz para 17 espécies (50,96%).



(c) Matriz para 50 espécies (30,88%).



(d) Matriz para 200 espécies (23,50%).

Figura 6.8: Matrizes de confusão resultante da aplicação de CNN.

ainda existem lacunas para essa abordagem. Sendo que, algumas poderiam ser contornadas por um procedimento de ajuste fino, e outras precisam ser melhor investigadas para que o desempenho de um sistema automático seja adequado para condições reais.

A Figura 6.9 apresenta um exemplo de filtros aprendidos na primeira camada convolucional de uma rede para classificação de imagens de espécies de pássaros do conjunto de 200 classes com imagens de tamanho 64x64 em relação a arquitetura adotada.

A comparação com trabalhos similares, mesmo que de forma indireta, pode ser realizada com o trabalho de Bo et al. (2013) que aplica o paradigma de aprendizagem profunda por meio de *sparse coding*. Alguns experimentos são realizados no contexto de granularidade fina, por meio de conjuntos de dados de imagens de cães, gatos e pássaros. O conjunto de dados CUB 200, com 15 imagens de cada classe para o conjunto de treinamento e o restante para o conjunto de testes consegue obter 30,3% de taxa de correta classificação. Outra comparação indireta pode ser estabelecida com o trabalho de Zhang et al. (2014) que utiliza o conjunto de dados CUB 200 2011 e aplica CNNs para implementação do paradigma de aprendizagem profunda. O principal diferencial do trabalho de Zhang et al. (2014) é detectar e identificar modelos de partes do pássaro no processo

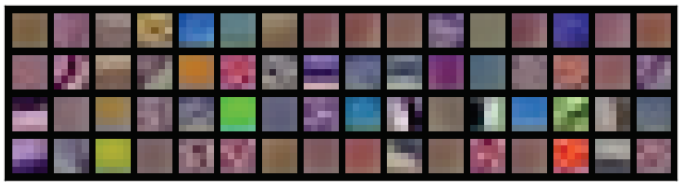


Figura 6.9: Exemplos de filtros aprendidos na primeira camada convolucional 5x5 de uma imagem de 64x64 pertencente ao conjunto de 200 classes.

de aprendizagem de características. Além disso, impõe restrições geométricas aprendidas entre a identificação de partes e o espaço definido pela caixa delimitadora que possui o modelo da parte detectada. Essa abordagem supera as demais presentes no Estado da Arte e consegue obter 76,37% de taxa de correta classificação.

6.2.6 Resultados da fusão audiovisual

O capítulo 5 apresentou a descrição e o novo método para combinação de informações visuais e acústicas. A seção atual apresenta os resultados obtidos com a aplicação do método, alicerçados nas melhores evidências disponíveis na literatura. Não há, no entanto, referência sobre outros métodos de combinação de informações visuais e acústicas aplicados à identificação de espécies. Devido a indisponibilidade de amostras de áudio, não foi possível estender os testes ao conjunto de 200 classes.

O processo de avaliação da fusão de informações visuais e acústicas é composto, basicamente, por duas etapas: 1) análise isolada dos resultados obtidos a partir das características visuais, após a aplicação do classificador (considerando ou não mecanismo de rejeição); 2) análise dos resultados obtidos por meio da fusão de informações visuais e acústicas, após a aplicação da classificação e mecanismos de rejeição.

A Tabela 6.11 apresenta a taxa da correta classificação para características visuais e acústicas por meio da verificação da hipótese $TOP\,N$. Neste caso, $TOP\,N$ indica se a espécie correta esta presente entre as N melhores saídas do classificador. A taxa de correta classificação é calculada por meio da soma das amostras corretamente classificadas divididas pelo número total das amostras participantes do teste. Por exemplo, a identificação correta da espécie do pássaro está entre as Top 6 hipóteses, para mais de 57% dos casos. Além disso, a Tabela 6.11 mostra que as características acústicas são mais discriminantes do que as visuais. A taxa de correta classificação para o classificador que utiliza informação acústica é cerca de 20% maior do que a obtida com o classificador que

utiliza informação visual, quando nenhum mecanismo de rejeição é aplicado. Para este experimento foram utilizadas 1480 amostras de imagens e 422 amostras de áudios para o conjunto de testes de 50 espécies.

Tabela 6.11: Taxa da correta classificação para características visuais e acústicas isoladas.

	Taxa de acerto (%)	
N melhores hipóteses	Visual	Áudio
TOP 1	27,03	45,97
TOP 2	36,76	57,58
TOP 4	48,92	72,04
TOP 6	57,77	79,62
TOP 8	64,05	84,36
TOP 10	68,72	86,97

6.2.6.1 Resultados da fusão audiovisual - Experimento 1

O Experimento 1 implementa a fusão audiovisual por meio dos vetores de características descritos no capítulo 5 nas seções: 5.2.1 e 5.2.2. A Tabela 6.12 apresenta resultados para o conjunto de testes de 50 espécies, por meio da combinação proposta para a fusão das saídas fornecidas por ambos os classificadores, visual e acústico. Os resultados são descritos em relação à aplicação de diferentes taxas de rejeição, sendo: 10%, 30% e 50%. De forma geral, a melhoria trazida com a utilização de características acústicas tende a aumentar com a utilização de um mecanismo de rejeição mais rigoroso, ou seja, uma taxa de rejeição mais alta. Cabe enfatizar que as características acústicas têm como objetivo auxiliar a classificação visual de espécies de pássaros. Os resultados obtidos com a re-classificação das amostras rejeitadas pelo classificador visual, usando as características acústicas, situam-se entre 1,2% e 2,9%, maiores que quando comparada ao classificador que utiliza apenas a informação visual.

Tabela 6.12: Taxas de acerto obtidas por meio da fusão de informação audiovisual SIFT e Áudio. Resultados para o conjunto de testes de 50 espécies e aplicação de 10%, 30% e 50% de taxa de rejeição.

Estratégia	Taxa de Acerto (%)		
	Taxa de Rejeição		
	10%	30%	50%
Visual	28,89	32,70	40,02
Visual + Acústico	30,10	35,65	42,20
Visual + Acústico (SUM)	29,71	35,22	41,90
Visual + Acústico (PROD)	26,96	35,25	42,04
Visual + Acústico (MAX)	26,96	35,25	42,04

Após a aplicação do novo método verificou-se que a utilização do mecanismo de rejeição é um forte aliado para a identificação de espécies. A associação deste permite a fusão de informação e a variação (aumento) nas taxas de correta classificação. Além disso, sua implementação feita de forma extremamente rápida e dinâmica, permitindo o acompanhamento por meio do compromisso erro e rejeição. A Figura 6.10 apresenta um gráfico deste compromisso para as estratégias abordadas: a) rejeição aplicada ao classificador visual + verificação da amostra de áudio; b) rejeição das amostras do classificador visual que possuem nível de confiança abaixo de λ; c) rejeição aplicada ao classificador visual + a combinação da saída do classificador acústico por meio da regra Sum; d) rejeição aplicada ao classificador visual + a combinação da saída do classificador acústico por meio da regra Prod.

A análise geral dos resultados e do gráfico que estabelece o compromisso entre as taxas de erro e de rejeição indica que a utilização do método de fusão de informação mostrou-se mais eficiente que o método tradicional, considerando apenas a imagem ou o áudio para a identificação da espécie de pássaros. As regras Sum, Prod e Max quando utilizadas em conjunto com o mecanismo de rejeição e a verificação acústica conseguem melhorar o desempenho em relação ao classificador visual. Entretanto, quando comparadas ao desempenho do classificador visual com verificação acústica e adoção de mecanismo de rejeição não conseguem indicar resultados competitivos, sendo inferior em todos os casos.

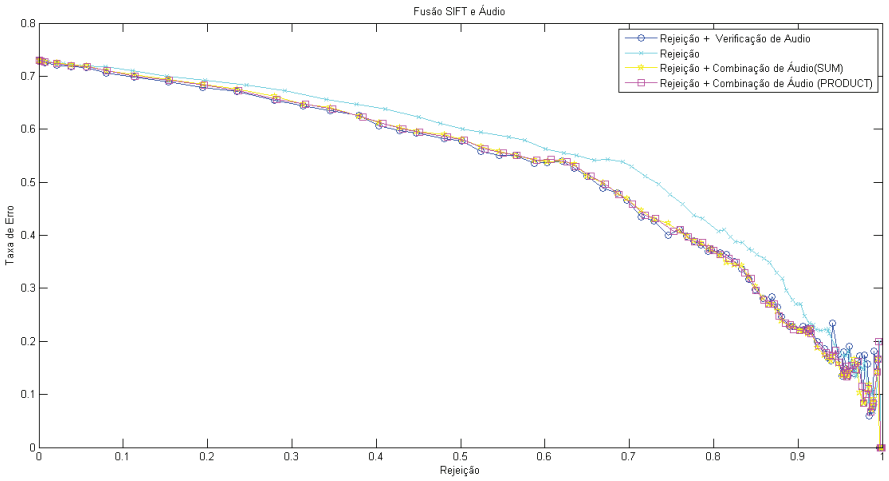


Figura 6.10: Representação do compromisso Erro e Rejeição para a fusão audiovisual: SIFT e Áudio.

6.2.6.2 Resultados da fusão audiovisual - Experimento 2

O Experimento 2 implementa a fusão audiovisual por meio do segundo vetor de características. Neste caso, o classificador visual utilizado é apresentado na Figura 4.11 da seção 4.5.1 descrita no capítulo 4. As características acústicas são descritas no capítulo 5 na seção 5.2.2. A Tabela 6.13 apresenta resultados para a fusão das saídas fornecidas por ambos os classificadores, visual (CNN) e acústico (Áudio). Os resultados são descritos em relação à aplicação de diferentes taxas de rejeição, sendo: 10%, 30% e 50%.

Tabela 6.13: Taxas de acerto obtidas por meio da fusão de informação audiovisual CNN e Áudio. Resultados para o conjunto de testes de 50 espécies e aplicação de 10%, 30% e 50% de taxa de rejeição.

Estratégia	Taxa de Acerto (%)		
	Taxa de Rejeição		
	10%	30%	50%
Visual	33,13	38,47	47,12
Visual + Acústico	34,46	41,65	50,56
Visual + Acústico (SUM)	34,22	41,65	50,56
Visual + Acústico (PROD)	34,36	41,66	50,16
Visual + Acústico (MAX)	34,36	41,66	50,16

A Figura 6.11 apresenta um gráfico deste compromisso para as estratégias abordadas. Os resultados obtidos por meio do Experimento 2 apresenta variação em escala positiva quando comparado ao Experimento 1. Quando considerados 10% de rejeição a

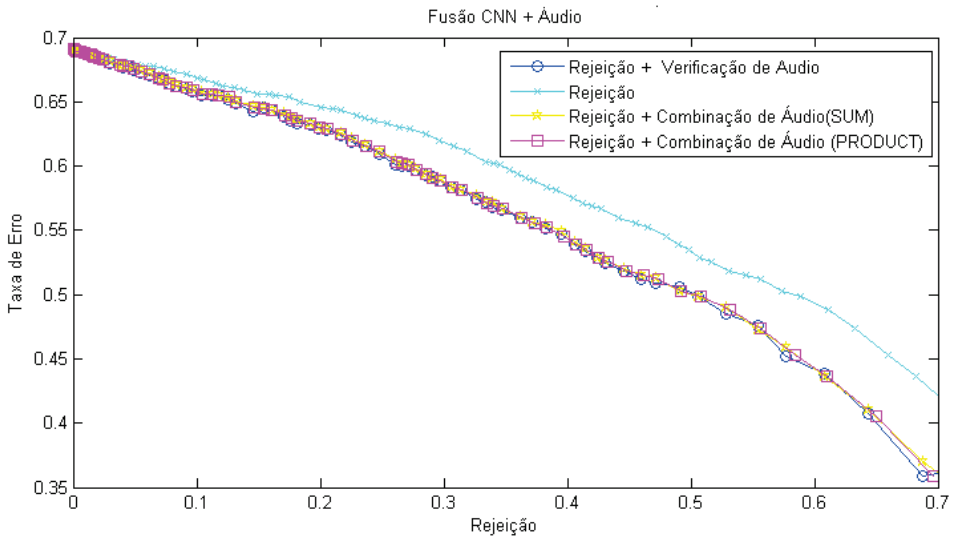


Figura 6.11: Representação do compromisso Erro e Rejeição para a fusão audiovisual: CNN e Áudio.

variação de 4,36%, com 30% de rejeição a variação é de 6%. Finalmente, com 50% de

rejeição a variação é de 8,36%. Os resultados obtidos fortalecem a hipótese de que fusão de informações é viável para melhorar o desempenho de um sistema de identificação de espécies de pássaros.

6.3 Combinação de classificadores

A aplicação de combinação de classificadores tem como objetivo buscar melhores resultados referentes à taxa de correta classificação. As principais vantagens obtidas com a utilização de regras de combinação é permitir que as deficiências de um determinado classificador pudessem ser suprimidas pelo bom desempenho de outros classificadores. As combinações dos classificadores consideraram três abordagens: i) todos os classificadores; ii) os três melhores desempenhos individuais; iii) os dois melhores desempenhos individuais; em todos os casos mantendo o número original de amostras em cada classe.

Para primeiro o experimento foram utilizados os sete classificadores individuais e seus respectivos rótulos de saída. Esses classificadores são descritos a seguir:

1. Cor_{RGB}: histograma de cores (30 faixas), classificador SVM, *kernel* RFB otimizado, características de cores obtidas por meio do espaço de cor RGB.

2. Cor_{HSV}: histograma de cores (30 faixas), classificador SVM, *kernel* RFB otimizado, características de cores obtidas por meio do espaço de cor HSV.

3. LBP_{RGB}: classificador SVM, *kernel* RFB otimizado, características de textura obtidas por meio de imagens no formato RGB. O resultado do melhor canal representa o conjunto.

4. LBP_{GRAY}: classificador SVM, *kernel* RFB otimizado, características de textura obtidas por meio de imagens em escala de cinza.

5. LBP_{H}: classificador SVM, *kernel* RFB otimizado, características de textura obtidas por meio de imagens no formato HSV. O resultado do canal H representa o conjunto.

6. SIFT+BoK: classificador SVM, *kernel* linear, características estruturais. O vetor de características foi obtido por meio do algoritmo SIFT, a construção de vocabulário, pontos-chave visuais e geração de histograma.

7. CNN: o classificador é uma rede neural convolucional. A abordagem de aprendizado de características é adotada.

A Tabela 6.14 apresenta os melhores resultados para os classificadores individuais (*Single Best* - SB) de acordo com número de classes consideradas. A Tabela 6.15 apresenta os resultados calculados para o Oráculo. Podemos considerar os valores obtidos por meio do Oráculo como o limite máximo de performance para o conjunto de classificadores. Além disso, a Tabela 6.15 apresenta à taxa de correta classificação versus a quantidade de classes, considerando o desempenho do classificador individual e os resultados para as regras que utilizam o voto majoritário: $Voto\,maj_{50\%+1}$ e $Voto\,maj_{moda}$; e as que utilizam o voto majoritário ponderado: $Voto\,maj_{acc}$ e $Voto\,maj_{feature}$.

Tabela 6.14: Melhores resultados para os classificadores individuais, independente da abordagem superficial ou profunda.

Classificadores Individuais	Taxa de acerto (%)
2 classes - LBP_{RGB}	95,00
5 classes - CNN	74,82
17 classes - CNN	50,96
50 classes - CNN	30,88
200 classes - CNN	23,50

Tabela 6.15: Definição do Oráculo e resultados para as diferentes combinações para o conjunto completo de 7 classificadores.

Taxa de correta classificação (%)						
Classes	Individual	*Oráculo*	$Voto\,maj_{50\%+1}$	$Voto\,maj_{moda}$	$Voto\,maj_{acc}$	$Voto\,maj_{feature}$
2 classes	95,00	100	98,33	98,33	95,00	95,00
5 classes	74,82	100	62,59	80,58	20,86	20,86
17 classes	50,96	100	19,62	58,85	9,38	9,38
50 classes	30,88	58,11	1,08	28,92	7,56	7,56
200 classes	23,50	45,96	0,41	23,5	1,65	2,14

As melhorias dos resultados utilizando as regras de combinação podem ser observadas quando o resultado do SB é comparado com o resultado obtido com a aplicação de uma regra. Para o subconjunto de 2 classes, a votação majoritária associada às regras, $Voto\,maj_{50\%+1}$ e $Voto\,maj_{moda}$, obtiveram resultados superiores aos obtidos ao classificador SB e muito próximo do resultado obtido pelo Oráculo. A regra $Voto\,maj_{moda}$ apresentou resultados superiores aos obtidos por meio do SB para os conjuntos de 5 e 17 classes.

Nos demais casos, as regras utilizadas não superam os resultados obtidos pelo SB. Uma explicação para isso deve-se à heterogeneidade dos classificadores combinados. Outra, à variação do número de classes e amostras. A combinação de todos os classificadores, considera os resultados obtidos pelas características de cores e texturas, mesmo que esses sejam muito inferiores aos obtidos por meio de SIFT+Bok ou CNN. Por exemplo, para

200 classes o desempenho do classificador que utiliza exclusivamente a característica cor, no formato RGB é 1,82%. Entretanto, a CNN obtém para o mesmo conjunto 23,50% de taxa de correta classificação.

Considerando o conjunto de teste completo com 5.794 amostras e 200 classes, algumas observações importantes acerca dos resultados são descritas: 3.131 amostras não foram classificadas por nenhum dos classificadores utilizados; dentre o conjunto completo de classificadores, pelo menos um deles classificou corretamente 2.663 amostras. Estendendo a busca para três ou mais classificadores, somente 24 amostras foram corretamente classificadas. E em nenhum caso, todos os classificadores (sete) conseguiram classificar a mesma amostra corretamente, ou seja, evidencia-se que é impossível obter a unanimidade entre os classificadores considerados.

A Tabela 6.16 apresenta os resultados calculados para o Oráculo quando considerados os 3 classificadores com melhor desempenho individual, sendo: LBP_{RGB}, SIFT+BoK e CNN. Quando apenas 3 classificadores são combinados, os resultados são visivelmente melhores do que os classificadores executados individualmente. A regra $Voto\,maj_{moda}$ obteve destaque em todos os conjuntos de dados. Em comparação com o resultado do SB, foi observada uma melhoria considerável para a taxa de correta classificação, embora os resultados se apresentem distantes dos obtidos pelo Oráculo. As regras que utilizam o voto majoritário ponderado obtiveram resultados superiores aos obtidos pelo SB, para os conjuntos de 2, 5 e 17 classes. Para os conjuntos maiores, indicaram resultados inferiores aos obtidos pelo SB.

Tabela 6.16: Definição do Oráculo e resultados para as diferentes combinações para o conjunto de 3 classificadores (com melhor desempenho individual em relação ao subconjunto ou conjunto de classes).

Taxa de correta classificação (%)					
Classes	Individual	Oráculo	$Voto\,maj_{moda}$	$Voto\,maj_{acc}$	$Voto\,maj_{feature}$
2 classes	95,00	100	98,33	100	100
5 classes	74,82	98,56	74,82	88,47	88,11
17 classes	50,96	81,88	59,91	58,41	58,41
50 classes	30,88	48,31	31,55	9,25	8,64
200 classes	23,5	39,13	24,49	7,76	8,07

A Tabela 6.17 apresenta os resultados calculados para o Oráculo quando considerados os 2 classificadores com melhor desempenho individual, sendo: SIFT+BoK e CNN. Neste caso, não foram observados melhorias nos resultados por meio da aplicação de nenhuma das regras consideradas. No caso da regra $Voto\,maj_{moda}$ os resultados obtidos foram equivalentes aos obtidos por meio do SB. Para a regra de voto majoritário ponderado os resultados foram inferiores em todos os casos.

Tabela 6.17: Definição do Oráculo para o conjunto de 2 classificadores (com melhor desempenho individual).

Taxa de correta classificação (%)				
Classes	Individual	*Oráculo*	$Voto\,maj_{moda}$	$Voto\,maj_{acc}$
2 classes	95,00	100	95,00	93,30
5 classes	74,82	91,37	74,82	68,34
17 classes	50,96	71,43	50,96	50,74
50 classes	30,88	43,72	30,88	26,75
200 classes	23,50	36,54	23,51	22,04

Uma comparação entre os valores estabelecidos pelo Oráculo do conjunto de 7 classificadores (Tabela 6.15), o Oráculo do conjunto dos 3 classificadores (Tabela 6.16) e o Oráculo do conjunto dos 2 classificadores (Tabela 6.17). Indica que apesar da heterogeneidade dos classificadores combinados, todos possuem relevantes contribuições, pois retirando classificadores do conjunto a taxa de correta classificação diminui. Outro forte indício para sustentar essa afirmação, deve-se a baixa variação nas taxas apontadas pelo Oráculo. Por exemplo, para o conjunto de 200 classes a variação é de 6,83%, e para o conjunto de 50 classes a variação é de 9,8%.

Por fim, as principais contribuições desse experimento residem: na combinação de classificadores que utilizam estratégias superficiais e estratégias profundas; e na constatação que as regras aplicadas para a fusão de rótulos pouco contribuem para aumentar o desempenho do sistema de classificação; apesar disso a combinação de classificadores consegue mostrar situações pontuais de ganhos e perdas de informação, ou seja, confirmam a hipótese de que existe complementaridade entre as diferentes características visuais.

6.3.0.3 Combinação de classificadores: nível de medida

A seção 6.3 apresentou os resultados para a combinação de classificadores em nível abstrato e em nível de Oráculo (Kuncheva, 2007). A seção atual apresenta um experimento adicional para verificar o comportamento da combinação de classificadores em nível de medida (Kuncheva, 2007). A Tabela 6.18 apresenta resultados em relação aos níveis de medida e Oráculo, considerando os três classificadores com melhor desempenho individual para 50 e 200 classes.

Os resultados obtidos sugerem que a combinação em nível de medida pode obter taxas de correta classificação melhores do que as obtidas por meio do nível abstrato. Cabe enfatizar que foram aproveitadas informações fornecidas pelos três classificadores individuais, mesmo com a grande variação desempenho entre eles. Por consequência da exploração de diferentes características e paradigmas de aprendizagem são gerados resultados bem

diferentes uns dos outros, de maneira que fornecem informações complementares umas às outras e podem ser exploradas por meio de diferentes regras de combinação. As regras Sum, Prod, Wsum e Wprod superam em todos os casos o desempenho do classificador individual, porém estão distantes de obterem resultados competitivos com os obtidos pelo Oráculo. A regra Prod se destacou em relação as demais pelo desempenho alcançado para o subconjunto de 50 classes e também para o conjunto completo de 200classes. Os resultados obtidos indicam que a combinação de classificadores possibilita a melhor capacidade de generalização para o problema de identificação de espécies de pássaros, assim como, resultados melhores e mais estáveis quando comparados a classificadores individuais.

Tabela 6.18: Combinação de classificadores em nível de medida.

Taxa de correta classificação (%)						
	Individual	Oráculo	Sum	Prod	wSum	wProd
50 classes	30,88	48,31	32,91	38,11	31,69	35,54
200 classes	23,50	39,13	28,35	33,44	27,60	31,79

6.4 Análise de Erros

O objetivo desta seção é apresentar uma análise das espécies incorretamente identificadas por meio de imagens. A ideia geral é analisar e compreender a similaridade entre as imagens incorretamente classificadas. O interesse está centrado em reconhecer e organizar imagens em uma estrutura. O pressuposto utilizado para fundamentar esse tipo de análise é o entendimento de que a imagem é produto da observação atenta a um cenário. Assim, propõe-se analisar essas imagens visando a identificar os elementos que constituem a estrutura das imagens incorretamente identificadas. Talvez, conhecedores destas estruturas, seja possível a criação de técnicas que possam operar de um modo significativo sobre a taxa de correta identificação.

Cabe esclarecer que a análise não é realizada pela observação da matriz de confusão, mas, sim, diretamente nas imagens em que o conjunto de sete classificadores apresentou mais dificuldade para identificação. Desta forma, foram recuperadas do conjunto de teste as 3.131 amostras que não foram corretamente identificadas por nenhum dos classificadores utilizados. A análise foi conduzida em relação aos subconjuntos de 2, 5, 17 e 50 classes, bem como o conjunto completo de 200 classes.

Foi possível chegar a algumas constatações importantes para os três conjuntos com menor número de amostras:

- Considerando o conjunto de 2 espécies (classes), 6 imagens, 3 de cada espécie, fa-

zem parte do grupo que nenhum dos classificadores utilizados conseguiu identificar. Devido à pouca quantidade de amostras, nenhuma conclusão pôde ser obtida.

- Em relação ao conjunto de 5 espécies, incluindo as duas anteriores, de um total de 139 imagens, 9 fazem parte do grupo que nenhum dos classificadores utilizados conseguiu identificar. Observa-se que 6 imagens fazem parte de uma espécie que também pertencia ao conjunto de 2 espécies (*035.Purple Finch*).

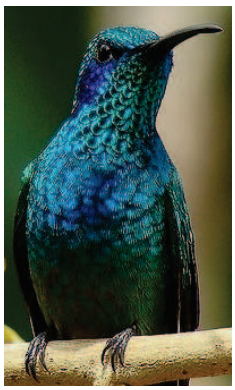
- Para o conjunto de 17 classes, incluindo as 7 anteriores, de um total de 469 imagens, 27 fazem parte do grupo que nenhum dos classificadores utilizados conseguiu identificar. Neste caso, observa-se a maior ocorrência de erros em 2 espécies (*045.Northern Fulmar e 065.Slaty backed Gull*).

Para o conjunto de 50 espécies de um total de 1480 imagens, 620 geram mais confusão aos classificadores e fazem parte do grupo que nenhum dos classificadores utilizados conseguiu identificar. A abordagem adotada é organizar a classificação de espécies de acordo com suas informações taxonômicas, na sequência: Reino, Filo, Classe, Ordem, Família, Gênero e Espécie. Observando essa organização as 50 espécies pertencem ao Reino Animalia, ao Filo Chordata, a classe Aves. A partir da Ordem é que iniciam-se as variações. Neste conjunto, 47 espécies são da Ordem Passeriformes e as outras três são respectivamente Anseriformes, Caprimulgiformes, Apodiformes.

Um detalhe importante com relação às espécies *105.Whip poor Will* da Ordem Caprimulgiformes e *070.Green Violetear* da Classe Apodiformes, é que nenhuma imagem pertencente a essas espécies foi incorretamente identificada. A espécie *088.Western Meadowlark* pertencente à Ordem Anseriformes criou confusão em apenas cinco amostras. A Figura 6.12 apresenta as três espécies em que os classificadores cometem menos erros. A Tabela 6.19 apresenta a relação de espécies em que o classificador cometeu até 5 erros de identificação.

Observando a organização por Família, foram identificadas 19 famílias: Alaudidae, Anatidae, Caprimulgidae, Cardinalidae, Emberizidae, Fringillidae, Hirundinidae, Lcteridae, Mimidae, Parulidae, Passeridae, Trochilidae, Troglodytidae, Tyrannidae, Vireonidae. A Família Vireonidae representada pelas espécies *52.Blue headed Vireo*, *54.Red eyed Vireo*, *156.White eyed Vireo* apresentaram altos números de erros de identificação. A Tabela 6.20 apresenta a relação de espécies em que o classificador cometeu 20 ou mais erros de identificação.

Os resultados descritos por meio da análise de erros indicam possibilidades que podem ser melhor investigadas. Um indício forte em relação a tal constatação deve-se principalmente à comparação com o trabalho de Branson et al. (2014).



(a) *070.Green Violetear.*

(b) *105.Whip poor Will.*

(c) *088.Western Meadowlark.*

Figura 6.12: Espécies pertencentes ao subconjunto de 50 classes em que os classificadores cometem menos erros.

Tabela 6.19: Resumo das espécies do conjunto de 50 classes com até 5 erros.

Informação do Classificador		Informação Taxonômica Reino: Animalia | Filo: Chordata | Classe: Aves	
Classe	Erros	Ordem	Família
088.Western Meadowlark	5	Passeriformes	Anatidae
057.Rose breasted Grosbeak	5	Passeriformes	Cardinalidae
161.Blue winged Warbler	5	Passeriformes	Parulidae
164.Cerulean Warbler	4	Passeriformes	Parulidae
140.Summer Tanager	3	Passeriformes	Cardinalidae
139.Scarlet Tanager	1	Passeriformes	Cardinalidae
105.Whip poor Will	0	Caprimulgiformes	Caprimulgidae
070.Green Violetear	0	Apodiformes	Trochilidae

Tabela 6.20: Resumo das espécies do conjunto de 50 classes com 20 ou mais erros.

Classificador		Informação Taxonômica Reino: Animalia | Filo: Chordata Classe: Aves | Ordem: Passeriformes
Classe	Erros	Família
116.Chipping Sparrow	23	Emberizidae
118.House Sparrow	23	Passeridae
193.Bewick Wren	23	Troglodytidae
152.Blue headed Vireo	23	Vireonidae
154.Red eyed Vireo	21	Vireonidae
156.White eyed Vireo	21	Vireonidae
162.Canada Warbler	20	Parulidae

Tabela 6.21: Resumo das espécies do conjunto de 200 classes com até 5 erros.

Classificador		Informação Taxonômica Reino: Animalia | Filo: Chordata Classe: Aves
Classe	Erros	Ordem
052.Pied billed Grebe	5	Charadriiformes
081.Pied Kingfisher	5	Podicipediformes
083.White breasted Kingfisher	5	Coraciiformes
105.Whip poor Will	5	Coraciiformes
110.Geococcyx	5	Caprimulgiformes
070.Green Violetear	5	Cuculiformes
101.White Pelican	4	Apodiformes
187.American Three toed Woodpecker	4	Pelecaniformes
007.Parakeet Auklet	4	Piciformes
046.Gadwall	3	Charadriiformes
053.Western Grebe	3	Anseriformes
089.Hooded Merganser	3	Podicipediformes
090.Red breasted Merganser	3	Anseriformes
113.Baird Sparrow	3	Anseriformes
190.Red cockaded Woodpecker	3	Passeriformes
028.Brown Creeper	3	Piciformes
005.Crested Auklet	2	Passeriformes
006.Least Auklet	1	Charadriiformes
018.Spotted Catbird	1	Charadriiformes
052.Pied billed Grebe	1	Passeriformes

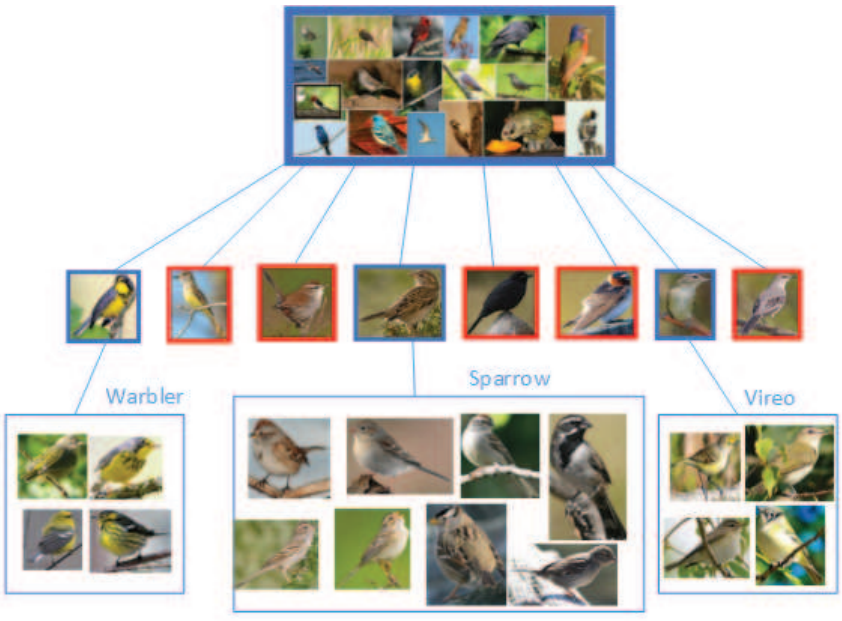


Figura 6.13: Representação visual dos principais erros de identificação para o conjunto CUB 200 2011.

Tabela 6.22: Resumo das espécies do conjunto de 200 classes com 20 ou mais erros.

Classificador		Informação Taxonômica Reino: Animalia | Filo: Chordata Classe: Aves
Classe	Erros	Ordem
115.Brewer Sparrow	28	Passeriformes
130.Tree Sparrow	28	Passeriformes
091.Mockingbird	27	Passeriformes
155.Warbling Vireo	27	Passeriformes
011.Rusty Blackbird	26	Passeriformes
114.Black throated Sparrow	26	Passeriformes
137.Cliff Swallow	26	Passeriformes
152.Blue headed Vireo	26	Passeriformes
163.Cape May Warbler	26	Passeriformes
038.Great Crested Flycatcher	25	Passeriformes
116.Chipping Sparrow	25	Passeriformes
117.Clay colored Sparrow	25	Passeriformes
119.Field Sparrow	25	Passeriformes
132.White crowned Sparrow	25	Passeriformes
161.Blue winged Warbler	25	Passeriformes
172.Nashville Warbler	25	Passeriformes

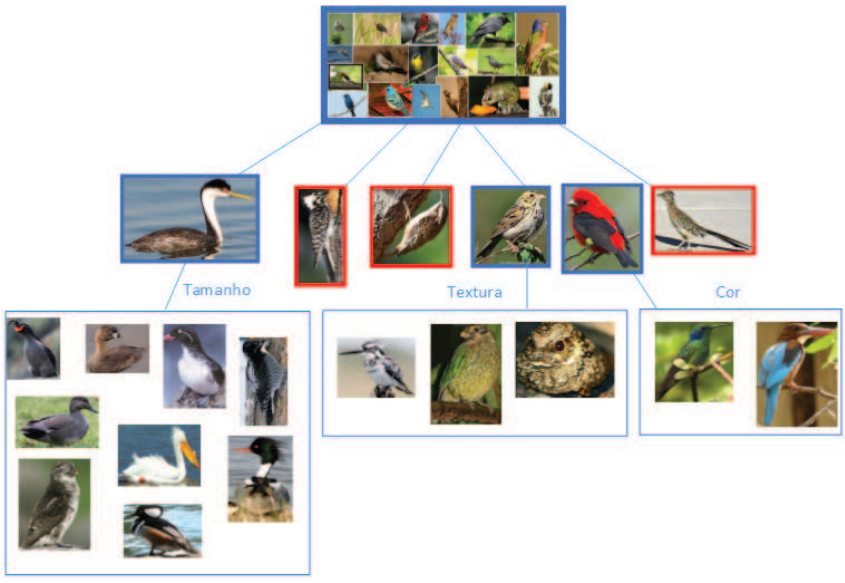


Figura 6.14: Representação visual dos principais acertos de identificação para o conjunto CUB 200 2011.

O recente trabalho de Branson et al. (2014) propõe um sistema híbrido de identificação de espécies. Desta forma, poderosos recursos de visão computacional são combinados a interação humana para a construção de um sistema de identificação para granularidade fina. Esse trabalho foi apresentado em detalhes na seção 3.2. Após os experimentos realizados na interface proposta, com um grupo de 27 usuários e 10 imagens de pássaros para identificação da espécie, algumas conclusões são descritas. Uma das conclusões obtidas refere-se aos pássaros pertencentes à Família *Sparrow*. As imagens de diferentes *Sparrows* apresentam muita similaridade aos não especialistas em pássaros, sendo indicada como uma das principais razões para que os usuários não conseguirem atingir 100% de correta classificação. O uso da informação taxonômica como ponto de referência indica a real semelhança entre as espécies (ou seja, todos vêm da Família Emberizidae, porém não compartilham do mesmo Gênero). As espécies *Sayornis* e *Gray Kingbird* também são citadas por causarem alto número de incorreta identificação devido à semelhança visual.

6.5 Discussões e considerações finais

Para análise dos resultados obtidos por meio deste trabalho em relação aos disponíveis no Estado da Arte, se faz necessária a definição dos elementos a serem comparados. A primeira comparação estabelece grupos de trabalho, cujo critério de formação constitui a abordagem geral do trabalho. As comparações estabelecidas são independentes de fatores e parâmetros, ou seja, uma comparação todos contra todos. A divisão é descrita a seguir: a) Definição de sistema de referência: Welinder et al. (2010) e sua atualização Wah et al. (2011); b) Métodos que consideram interação humana: Branson et al. (2010), Wah et al. (2011), Deng et al. (2013), Branson et al. (2014); c) Métodos que consideram segmentação: Chai et al. (2011), Chai et al. (2012), Chai et al. (2013) e Angelova e Zhu (2013); d) Métodos que consideram partes: Zhang et al. (2012), Berg e Belhumeur (2013), Zhang e Farrell (2013), Gavves et al. (2013); e) Métodos que consideram áudio: Glotin et al. (2013), Goëau et al. (2014), Lopes et al. (2011), Stowell e Plumbley (2014); e f) Métodos que consideram aprendizagem de características: Bo et al. (2013), Zhang et al. (2014).

O objetivo desta análise é comparar os grupos para identificar as abordagens com melhor desempenho acerca de taxa de correta identificação. O resultado desta busca é a constatação de que os trabalhos que utilizam métodos de identificação de partes são os que conseguem obter melhores resultados. Cabe salientar que o Estado da Arte aponta o trabalho de Zhang et al. (2014), pertencente ao grupo que utiliza aprendizagem de características, como o melhor desempenho atual. Entretanto, este trabalho combina a

identificação de partes e a aprendizagem de características. O próximo destaque de desempenho é atribuído ao grupo de métodos que consideram áudio, seguidos, respectivamente, pelos métodos que envolvem interação humana e segmentação.

Neste trabalho, a obtenção dos resultados por meio de abordagens superficiais e características de cor, textura e forma extraídas da imagem podem ser comparados aos resultados obtidos pelo grupo de sistema de referência. Em particular, quanto ao caso das características de texturas, é possível a definição de um sistema de referência para a identificação de espécies de pássaros. Até onde essa pesquisa pode alcançar não foram encontrados trabalhos que pretendem resolver o problema por meio de descritores LBP, método largamente utilizado em granularidade fina, como por exemplo, na identificação facial.

De forma geral, a comparação dos resultados traz a constatação de que os resultados de nossos experimentos não se apresentam de forma competitiva em comparação direta com os demais grupos. Em consequência disso, as demais comparações são realizadas por contrastes observados na maneira de obtenção de resultados. O primeiro contraste deve-se ao fato de este trabalho não utilizar informação de partes em nenhum dos experimentos realizados; o segundo, a troca da custosa etapa de segmentação pela utilização da imagem recortada pelas informações de caixa delimitadora; e o terceiro, a não utilização de metadados ou *ground truth*; quarto, a não utilização de ferramentas adicionais de normalização, como é o caso da normalização de poses, ou de alinhamentos em relação às imagens; quinto, e último, a informação taxonômica relacionada à espécie também não foi considerada.

Os resultados obtidos experimentalmente mostram o comportamento isolado de características importantes, em condições muito próximas de cenários reais. Exemplificando, as cores presentes em uma imagem, utilizadas de forma isolada, apresentam desempenho menor do que o esperado para a construção de uma ferramenta de identificação de espécies, que possa tornar-se uma aplicação útil. Entretanto, a combinação de diferentes informações visuais, como cores e texturas (seção 6.3), ou principalmente, a fusão de informação visual e acústica (seção 6.2.6), implica no aumento da taxa de correta classificação de um sistema de identificação automática.

Para a comparação final, os trabalhos foram organizados quanto à abordagem superficial ou profunda. As abordagens atuais para a identificação de espécies em problemas de granularidade fina, na maioria dos casos bem sucedidos, dependem de uma robusta localização de partes do objeto de interesse, para que seja possível a extração de características adequadas para a discriminação entre as classes. No entanto, a localização de partes se apresenta como uma etapa adicional, não trivial ao processo de identificação de

espécies, mas que se torna viável por isolar as sutis diferenças entre as mesmas.

Para facilitar a comparação, apresentam-se trabalhos associados a abordagens superficiais e profundas que possuem o melhor desempenho para o problema de identificação de espécies de pássaros. Alguns exemplos de trabalhos que utilizam informações de partes e indicam as melhores taxas para a identificação de espécies, utilizando o conjunto de dados CUB 200 2011 são: Os trabalhos de Zhang e Farrell (2013) (50,98%), Berg e Belhumeur (2013) (56,78%), Chai et al. (2013) (59,40%), Gavves et al. (2013) (62,7%). Embora, ainda existam poucos trabalhos que apliquem abordagens profundas em problemas de granularidade fina, o cenário atual, indica grande vantagem quando essa abordagem é utilizada e mostra que a identificação de partes é adequada para a aprendizagem de características. Os trabalhos de Bo et al. (2013) (30,3%) e Zhang et al. (2014) (76,37%) incorporam a identificação de partes do objeto de interesse em seus métodos.

Capítulo 7

Conclusão

A identificação visual tem sido uma área atuante de pesquisa em visão computacional. O reconhecimento de imagens tem sido estudado por muitos anos e possibilita a solução de inúmeros problemas no nível básico de categorização. Neste trabalho, o foco foi um problema de granularidade fina, ou também conhecido como de baixo nível de categorização (que exige o tratamento de diferenças sutis entre as classes), utilizando algoritmos de Aprendizagem de Máquina.

Para relembrar a diferença entre os dois níveis de categorização, considere uma imagem qualquer. O nível básico poderá responder, de maneira precisa, quais objetos estão presentes na imagem (uma planta, um animal, um veículo, etc.). Entretanto, baixo nível de categorização poderá responder, infelizmente não ainda com precisão adequada, que a planta é um cogumelo da espécie *Amanita Muscaria* (uma variedade tóxica que pode causar envenenamento fatal a seres humanos). Ao contrário do reconhecimento de nível básico, o baixo nível de categorização pode ser difícil para seres não especialista no domínio da aplicação. Assim, um sistema visual automatizado para esse tipo de tarefa pode ser valioso e viável em muitas aplicações reais.

Neste trabalho, abordamos um problema de granularidade fina, por meio da identificação de espécies de pássaros em imagens capturadas em ambientes sem restrições. Duas estratégias de identificação visual, no contexto de sistemas de classificação supervisionada, são comparadas: estratégias superficiais e estratégias profundas.

A forma usual de resolver essa tarefa é encontrar e extrair um conjunto de características que consiga representar a variabilidade dos dados. A etapa de extração de características é computacionalmente exigente e as características escolhidas podem impactar diretamente nos resultados obtidos. Denominamos as práticas atuais que buscam solução por meio de extração de características como estratégias superficiais.

Os conceitos de aprendizagem profunda vêm ganhando interesse na comunidade

de visão computacional. A ideia chave é ignorar a etapa de extração de características e possibilitar que elas possam ser aprendidas a partir das imagens pertencentes ao conjunto. Denominamos as práticas atuais que buscam solução por meio de aprendizagem de características como estratégias profundas. Devido ao grande número de variáveis, os modelos de aprendizagem profunda geralmente precisam ser aprendidos a partir de dados de grande escala.

Os resultados obtidos por esse trabalho buscam responder às hipóteses iniciais e estabelecer conclusões em relação aos aspectos relativos à identificação visual automática, no contexto de classificação supervisionada, para problemas de granularidade fina, em duas estratégias de representação: superficial e profunda.

O estudo das características visuais que o ser humano utiliza para identificar as espécies foi conduzido para características de cor, textura e forma, pertencentes às imagens dos subconjuntos e conjuntos de dados. Os resultados obtidos indicam que as características visuais estudadas podem auxiliar na identificação de espécies de pássaros de maneira computacional baseando-se, em Aprendizagem de Máquina. Entretanto, as características isoladas não conseguem apresentar resultados satisfatórios. A constatação da possibilidade de existir complementariedade entre diferentes características nos levou a estudar a fusão de diferentes tipos de informação.

A identificação de espécies de pássaros é um excelente cenário para a realização de experimentos em relação à fusão de informação. Isso se deve ao fato que a identificação pode ser feita a partir de imagens ou sons emitidos pelos pássaros. O método proposto para fusão a de informação visual e acústica fornecem resultados que permitem a conclusão de que a fusão visual e acústica pode auxiliar e melhorar as taxas de correta classificação de sistema de identificação automático. Em outro nível de fusão, combinação de diferentes informações visuais, como cores e texturas (por meio regras de combinação de classificadores) também apresentam impacto na taxa de correta classificação.

Os resultados de pesquisa desta tese, especialmente o modelo de fusão de informação audiovisual proposto, conseguem contribuir na perspectiva de possíveis soluções para o tratamento de diferentes tipos de problemas de identificação visual de granularidade fina. Os experimentos realizados com alguns dos algoritmos do Estado da Arte fazem a extração de características em imagens sem restrição de entrada. O vetor obtido foi submetido à aprendizagem supervisionada por meio do classificador SVM. A análise dos resultados nos leva a concluir que as abordagens disponíveis somente indicam resultados satisfatórios (cometem poucos erros de identificação) quando o número de classes é pequeno. Para os subconjuntos de 2 ou 5 classes todas as abordagens consideradas nos experimentos produzem resultados satisfatórios. Na maioria dos casos, quando conside-

rado o conjunto completo de 200 classes os resultados são drasticamente reduzidos. Essa constatação indica uma grave limitação para estender essas soluções para cenários reais que possuem uma grande variedade de espécies.

Os princípios teóricos e metodológicos utilizados no decorrer de toda a pesquisa sustentam e envolvem um tema de interesse atual e pesquisa ativa. A recentidade dos trabalhos correlatos e a ininterrupta atualização do Estado da Arte dão condições necessárias às comparações de estratégias capazes de definir e organizar formalmente abordagens que desempenham variados papéis no domínio específico de problemas de granularidade fina. A comparação de diferentes propostas para a solução do problema evidenciou características relevantes que podem ser utilizadas para a solução de problemas de granularidade fina, com significativo destaque a identificação de partes do objeto de interesse.

A associação de propostas representativas envolvendo a identificação visual foi o grande diferencial desta tese em relação às levantadas no estudo do Estado da Arte (apresentado no capítulo 3) e as estudadas no decorrer desta pesquisa. Geralmente, as propostas ou focam em aspectos mais genéricos, envolvendo a busca por altas taxas de correta classificação, mas sem contemplar estudos detalhados em relação ao impacto de determinadas características, ou oferecem descritores específicos para resolver os problemas inerentes à granularidade fina. Por fim, as contribuições propostas se destacam como uma solução abrangente para a identificação de espécies e problemas de granularidade fina por terem sido obtidas a partir de uma avaliação metódica e criteriosa realizada, por meio de diferentes abordagens, o que promoveu as condições necessárias para conclusões e produção de conhecimento apropriado ao domínio de identificação visual de espécies de pássaros.

7.1 Principais contribuições

Os resultados alcançados com a pesquisa refletem diversas contribuições referentes à utilização de características visuais, especialmente para problemas de granularidade fina, quando se propõe diferentes abordagens: superficiais e profundas. Conforme foi evidenciado no decorrer da pesquisa, existem problemas significativos no tocante ao problema de granularidade fina e identificação de espécies, especificamente para a solução por meio de imagens, principalmente porque as imagens não possuem restrições de coleta.

Desse modo, a questão central pode ser posta da seguinte forma: como identificar eficazmente espécies de pássaros por meio de imagens, dadas as inúmeras variações, particularidades e conjunturas. As principais contribuições envolvem esclarecimento de aspectos referentes à identificação visual de granularidade fina em relação às quatro hi-

póteses iniciais desta pesquisa.

Primeiro, busca-se a viabilidade do uso de informações geradas por algoritmos de extração de características e aprendizagem supervisionada, em imagens sem restrição de entrada. Os experimentos realizados com alguns dos algoritmos do Estado da Arte fazem uso de vetores obtidos nas condições descritas e submetidos à aprendizagem supervisionada por meio do classificador SVM. A análise dos resultados nos leva a concluir que as abordagens disponíveis, somente indicam resultados satisfatórios (cometem poucos erros de identificação da espécie) quando o número de classes é pequeno. Para os subconjuntos de 2 ou 5 classes todas as abordagens consideradas nos experimentos produzem resultados satisfatórios. Na maioria dos casos, quando considerado o conjunto completo de 200 classes os resultados são drasticamente reduzidos. Essa constatação indica uma grave limitação para estender essas soluções para cenários reais que possuem uma grande variedade de espécies. Além disso, nos subconjuntos a presença da granularidade fina não se apresenta de forma marcante como em um número maior de classes.

Segundo, a partir da hipótese de que o ser humano faz uso de características visuais para identificar espécies, apresenta-se um estudo detalhado de abordagens de identificação visual baseadas exclusivamente em cores, texturas e formas. O uso de características visuais que diferenciam espécies podem auxiliar na identificação de espécies de pássaros de maneira computacional, baseando-se em Aprendizagem de Máquina, mas não possuem bom desempenho quando isoladas. Uma das principais contribuições refere-se à verificação de que existe complementariedade entre diferentes características visuais. Adicionalmente, um novo método de segmentação de imagens baseado nas cores presentes na imagem foi proposto, implementado e avaliado.

Terceiro, considerando que a identificação de espécies de pássaros pode ser feita a partir de imagens ou sons emitidos pelos pássaros, um método de fusão de informações visual e acústica pode auxiliar e melhorar os resultados de um sistema de identificação automático. Um novo método para fusão audiovisual foi proposto, implementado e avaliado. Para a implementação do novo método foi construído um conjunto de dados de áudio relacionado às espécies pertencentes ao conjunto CUB 200 2011. O conjunto CUB 50 Songs possui as gravações de áudio de 50 espécies, na forma original e a geração de novas amostras manualmente segmentadas. Ambos os conjuntos podem ser explorados em novas pesquisas.

O novo método de fusão de informações visual e acústica faz uso de mecanismos de rejeição com variações no compromisso erro e rejeição. Essa alternativa se apresenta de forma viável para lidar com um problema tão peculiar, principalmente quando o número de amostras de imagens é diferente do número de amostras de áudio. A fusão de classifica-

dores, em termos abstratos e sem apoio probabilístico pode auxiliar no aumento da taxa de acerto de um sistema automático de identificação de espécies. Embora o aumento seja pequeno, a fusão fortalece a conclusão de que existe complementariedade entre diferentes características visuais.

Quarto, as contribuições anteriores aliadas à solução da terceira questão de pesquisa da tese permite responder à quarta questão: qual abordagem seria mais adequada para representar problemas de granularidade fina e os impactos desta abordagem nas taxas de identificação visual automática. Considerando que a complementariedade entre diferentes características pode não ser suficiente para identificar espécies de pássaros, novas abordagens podem ser adotadas. Desta forma, a aprendizagem de características, a fusão de informações e a combinação de classificadores podem gerar resultados satisfatórios. Destaque-se, que essas alternativas são associadas com sucesso na literatura para a solução de problemas em conjuntos de larga escala.

Finalmente, outra contribuição importante da comparação do tema da identificação de espécies por meio dos paradigmas de aprendizagem superficial e profunda é a análise de erros. Uma análise referente às imagens incorretamente classificadas, resultantes da combinação de classificadores de ambos os paradigmas de aprendizagem apresentou uma estrutura de similaridade entre as amostras. O resultado dessa análise aponta que a seguinte organização taxonômica: Ordem, Família, Gênero e Espécie reflete nos erros do classificador. Assim, as peculiaridades de uma espécie interfere diretamente na taxa de acerto do classificador. Outras contribuições na esfera social e do ponto de vista ecológico foram descritas no capítulo 1 na seção 1.4.

7.2 Publicações

Esta tese gerou trabalhos publicados em eventos nacionais e internacionais, os quais são listados a seguir. Outros trabalhos ainda se encontram em fase de redação e serão publicados futuramente.

Congressos internacionais - artigos completos

- MARINI, A. ; FACON, J. ; KOERICH, A.L. Bird Species Classification Based on Color Features. In: IEEE International Conference on Systems, Man, and Cybernetics, 2013, Manchester. Proceedings of the IEEE SMC 2013. New York: IEEE Press, 2013. p. 4336-4341.

- MARINI, A. ; TURATTI, A.J. ; BRITO, A.S. ; KOERICH, A.L. Visual and Acoustic Identification of Bird Species. In: International Conference os Acoustics, Speech

and Signal Processing, 2015, Brisbane. Proceeding of ICASSP 2015. New York: IEEE, 2015. p. 1-4.

Congresso nacional - artigo completo

- MARINI, A.; FLORIDO, I. H.; KOERICH, A. L. Classificação de Espécies de Pássaros Utilizando Características de Textura. In: Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2013, Fortaleza.

Resumos publicados

- MARINI, A.; TURATTI, A. J.; KOERICH, A.L. Identificação Automática de Espécies de Pássaros Usando Visão Computacional. In: XXI SEMIC, 2013, Curitiba. XXI Seminário de Iniciação Científica da Pontifícia Universidade Católica do Paraná (SEMIC). Curitiba: Editora Champagnat, 2013. v. 1

- TURATTI, A. J.; MARINI, A.; KOERICH, A.L. Identificação automática de espécies de pássaros por meio de fusão de informações audiovisuais. In: XXII SEMIC, 2014, Curitiba. XXII Seminário de Iniciação Científica da Pontifícia Universidade Católica do Paraná (SEMIC). Curitiba: Editora Champagnat, 2014. v. 1

7.3 Trabalhos futuros

Nesta seção são listadas algumas sugestões que podem servir como diretivas para trabalhos futuros:

- Explorar a fusão de informações audiovisuais em diferentes características otimizadas para imagens e áudio.

- Abordar a combinação de classificadores em diferentes níveis de fusão de rótulos de classes preditas por um classificador.

- Considerar metadados e informação de localização de partes em relação ao objeto de interesse.

- Estender os experimentos considerando aprendizagem de características.

- Validar as melhores práticas observadas em outros domínios de granularidade fina, além de pássaros.

- Contribuir para a organização de dados de espécies de forma automática. Uma situação crítica é a dificuldade de validação de um grande número de espécies abrangentes. A definição de *ground-truth* normalmente envolve interação manual por especialistas, o que é uma tarefa intratável. Este gargalo, de certa forma limita as abordagens que podem ser aplicadas e resultados obtidos.

7.4 Considerações finais

Este capítulo apresentou as contribuições deste trabalho em diferentes níveis a fim de se perceberem as contribuições científicas desta pesquisa. Além disso, foram apresentadas propostas de trabalhos que podem dar continuidade à pesquisa conduzida até aqui. A identificação visual em problemas de granularidade fina ainda deixa muitas questões em aberto, e podem gerar inúmeras pesquisas a fim de proporcionar soluções para esse tipo de problema.

Referências Bibliográficas

Acevedo, M. a., C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, e T. M. Aide (2009, September). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics 4* (4), 206–214. 8

Angelova, A. e S. Zhu (2013, June). Efficient Object Detection and Segmentation for Fine-Grained Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 811–818. IEEE. 39, 51, 119

Bardeli, R., D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, e K.-H. Frommolt (2010, September). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters 31* (12), 1524–1534. 9

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning 2* (1), 1–127. 29

Berg, T. e P. N. Belhumeur (2013, June). POOF: Part-Based One-vs.-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 955–962. IEEE. 39, 51, 119, 121

Bo, L., X. Ren, e D. Fox (2013, June). Multipath Sparse Coding Using Hierarchical Matching Pursuit. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 660–667. IEEE. 47, 51, 105, 119, 121

Branson, S., G. Van Horn, C. Wah, P. Perona, e S. Belongie (2014, February). The Ignorant Led by the Blind: A Hybrid HumanâMachine Vision System for Fine-Grained Categorization. *International Journal of Computer Vision 108* (1-2), 3–29. 34, 35, 51, 115, 119

Branson, S., C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, e S. Belongie

(2010). Visual Recognition with Humans in the Loop. Technical report, University of California, San Diego - California Institute of Technology. 34, 35, 51, 119

Chai, Y., V. Lempitsky, e A. Zisserman (2011). BiCoS: A Bi-level co-segmentation method for image classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2579–2586. 37, 38, 51, 58, 119

Chai, Y., V. Lempitsky, e A. Zisserman (2013, December). Symbiotic Segmentation and Part Localization for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision*, pp. 321–328. IEEE. 39, 51, 119, 121

Chai, Y., E. Rahtu, V. Lempitsky, L. V. Gool, e A. Zisserman (2012). TriCoS: a tri-level class-discriminative co-segmentation method for image classification. In *Computer Vision ECCV*, pp. 794–807. 38, 51, 119

Chang, C.-C. e C.-J. Lin (2013). LIBSVM : A Library for Support Vector Machines. Technical report. 26, 28, 66

Chatfield, K., K. Simonyan, A. Vedaldi, e A. Zisserman (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. *Computing Research Repository (CoRR) abs/1405.3531*. 15, 16

Choi, J. Y., K. N. Plataniotis, e Y. M. Ro (2010). Using colour local binary pattern features for face recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 4541–4544. 19

Ciresan, D., U. Meier, e J. Schmidhuber (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference*, pp. 3642–3649. IEEE. 29, 69, 104, 146

Costa, Y., L. Oliveira, A. Koerich, F. Gouyon, e J. Martins (2012, November). Music genre classification using LBP textural features. *Signal Processing 92*(11), 2723–2737. 19, 83

Csurka, G., C. Dance, L. Fan, J. Willamowski, e C. Bray (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, pp. 1—-22. 25, 65

Das, M. e R. Manmatha (2001). Automatic segmentation and indexing in a database of bird images. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Volume 2, pp. 351 —-358 vol.2. 37, 58

Deng, J. (2012). *Large scale visual recognition*. Ph. D. thesis. 6

Deng, J., J. Krause, e L. Fei-Fei (2013). Fine-Grained Crowdsourcing for Fine-Grained Recognition. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 580–587. 2, 6, 36, 51, 119

Duda, R. O., P. E. Hart, e D. G. Stork (2000). *Pattern Classification*. New York: John Wiley & Sons. 17

Elkan, C. (2003). Using the triangle inequality to accelerate k-means. In *ICML*, pp. 147—-153. 26, 65

Fan, R.-e., X.-r. Wang, e C.-j. Lin (2008). LIBLINEAR : A Library for Large Linear Classification. Technical report. 26, 67

Fei-Fei, L. (2013). Datasets and challenges. Disponível em: http://www.ipam.ucla.edu/publications/gss2013/gss2013_11323.pdf. Acesso em: 19 de Maio de 2014. . 6

Gavves, E., B. Fernando, C. Snoek, A. Smeulders, e T. Tuytelaars (2013, December). Fine-Grained Categorization by Alignments. *2013 IEEE International Conference on Computer Vision*, 1713–1720. 41, 51, 119, 121

Glotin, H., C. Clark, Y. Lecun, P. Dugan, X. Halkias, e J. Sueur (2013). The 1st International Workshop on Machine Learning for Bioacoustics. In ICML (Ed.), *Proc. 1st workshop on Machine Learning for Bioacoustics - ICML4B*, Volume 1, Atlanta. 8, 9, 42, 119

Goëau, H., H. Glotin, W. Vellinga, e A. Rauber (2014). LifeCLEF bird identification task 2014. *CLEF working notes*, 585–597. 42, 43, 79, 80, 119

Gonzalez, R. C. e R. E. Woods (2000). *Processamento de Imagens Digitais*. São Paulo: Blucher. 17, 18

Hafemann, L. G., L. S. Oliveira, e P. Cavalin (2014). Forest Species Recognition using Deep Convolutional Neural Networks. In *International Conference on Pattern Recognition*, Stockholm, Sweden, pp. 1103–1107. 68

Kasten, E. P., P. K. McKinley, e S. H. Gage (2010, May). Ensemble extraction for classification and detection of bird species. *Ecological Informatics 5*(3), 153–166. 8

Krizhevsky, A. (2014). Cuda convnet. Disponível em: https://code.google.com/p/cuda-convnet. Acesso em: 19 de Outubro de 2014. . 30, 71

Krizhevsky, A. e G. Hinton (2009). Learning multiple layers of features from tiny images. *Science Department, University of Toronto, Tech.*. 69, 104, 146

Krizhevsky, A., I. Sutskever, e G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural*, 1–9. 29, 31, 67, 68, 69, 104, 146

Kuncheva, L. I. (2007, May). Combining Pattern Classifiers: Methods and Algorithms. *IEEE Transactions on Neural Networks 18* (3), 964–964. 71, 113

Lecun, Y., L. Bottou, Y. Bengio, e P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86* (11), 2278–2324. 30

Lin, H. H. e W. Chen (2011). Study on Recognition of Bird Species in Minjiang River Estuary Wetland. *Procedia Environmental Sciences 10, Part C*, 2478–2483. 49

Lin, H. H. e H. Guo (2012, June). Research of wetland birds networking and method of collaborative epidemic identify. *Robotics and Applications (ISRA), 2012* (1), 717–720. 49

Liu, Y., J. Zhang, D. Tjondronegoro, e S. Geve (2007). A Shape Ontology Framework for Bird Classification. In *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on*, pp. 478–484. 48

Lopes, M. T., L. L. Gioppo, T. T. Higushi, C. A. A. Kaestner, C. N. Silla, e A. L. Koerich (2011). Automatic Bird Species Identification for Large Number of Species. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pp. 117–122. 44

Lopes, M. T., C. A. A. Kaestner, C. N. S. Jr, e A. L. Koerich (2011). Identificação Automática de Espécies de Pássaros a partir da Análise do Canto. *13o Simpósio Brasileiro de Computação Musical (SBCM2011), Vitória, Brazil ISSN 2175-*, pp.1—-4. 44, 119

Lovett, R. a. (2012, November). How birds are used to monitor pollution. *Nature* (November), 2012–2014. 9

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, Volume 2, pp. 1150 –1157 vol.2. 21

Lowe, D. G. (2004, November). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 60*(2), 91–110. vi, 21, 22, 23, 24, 25

Mäenpää, T. e M. Pietikäinen (2005). Texture analysis with local binary patterns. In: Chen CH & Wang PSP (eds) Handbook of Pattern Recognition and Computer Vision, 3rd ed, World Scientific, 197-216 (invited chapter). 19

Mäenpää, T., M. Pietikäinen, T. Maenpaa, e M. Pietikainen (2004). Classification with color and texture: jointly or separately? *Pattern recognition 37*(8), 1629–1640. 19, 100

Mäenpää, T., M. Turtinen, e M. Pietikäinen (2003). Real-time surface inspection by texture. *Real-Time Imaging*. 19

Marini, A., J. Facon, e A. L. Koerich (2013, October). Bird Species Classification Based on Color Features. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 4336–4341. IEEE. 18

Marini, A. e A. L. Koerich (2008). Análise de estratégias de rejeição para problemas com múltiplas classes. In *XXXIV Conferência Latinoamericana de Informática*, Santa Fé, pp. p. 162–171. 84

Moghimi, M. (2011). Using Color for Object Recognition. Technical report, California Institute of Technology. 38, 51

Nadimpalli, U. D., R. R. Price, S. G. Hall, e P. Bomma (2006). A Comparison of Image Processing Techniques for Bird Recognition. *Biotechnol Progress 22*, 9–13. 48

Ojala, T., T. Mäenpää, e M. Pietikäinen (2001). A generalized Local Binary Pattern operator for multiresolution gray scale and rotation invariant texture classification. In *Pattern Recognition ICAPR 2001*. 21

Ojala, T., M. Pietikäinen, e D. Harwood (1996, January). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition 29*(1), 51–59. 19

Ojala, T., M. Pietikäinen, e T. Mäenpää (2000). Gray scale and rotation invariant texture classification with local binary patterns. *Computer Vision-ECCV 2000*. 146

Ojala, T., M. Pietikäinen, e T. Mäenpää (2002, July). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24*(7), 971–987. vi, 19, 20

Ranzato, M. A. (2013). Deep Learning for Vision : Tricks of the Trade. vi, 16

Rother, C., V. Kolmogorov, e A. Blake (2004). "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. 23*(3), 309–314. 38

Sermanet, P., S. Chintala, e Y. Lecun (2012). Convolutional neural networks applied to house numbers digit classification. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3288–3291. vii, 29, 32

Sivic, J. e A. Zisserman (2003). Video Google: a text retrieval approach to object matching in videos. *Proceedings Ninth IEEE International Conference on Computer Vision* (Iccv), 1470–1477 vol.2. 26, 65

Stowell, D. e M. D. Plumbley (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *Computing Research Repository (CoRR)*. 45, 119

Tuceryan, M. e A. K. Jain (1993). Handbook of pattern recognition & computer vision. Chapter Texture an, pp. 235–276. River Edge, NJ, USA: World Scientific Publishing Co., Inc. 19

Tzanetakis, G. e P. Cook (2002, July). Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on 10*(5), 293–302. 83

Vapnik, V. (2000). *The nature of statistical learning theory*. Statistics for Engineering and Information Science. New York: Springer. 26

Vedaldi, A. e B. Fulkerson (2010). VLFeat: An open and portable library of computer vision algorithms. *Proceedings of the international conference*, 1469–1472. 22, 34, 40, 65, 103

Wah, C., S. Branson, P. Perona, e S. Belongie (2011). Multiclass recognition and part localization with humans in the loop. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2524–2531. 35, 119

Wah, C., S. Branson, P. Welinder, P. Perona, e S. Belongie (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology. x, 14, 15, 33, 51, 57, 67, 102, 119, 137

Wang, H.-C., Y.-S. Chen, e M.-Y. Wu (2010). A user-augmented object query system using color and shape features for Taiwan wild birds photos. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, Volume 5, pp. 2516–2520. 49

Welinder, P., S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, e P. Perona (2010). Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology. vi, 3, 13, 14, 33, 51, 67, 89, 119, 137

Yao, B., G. Bradski, e L. Fei-Fei (2012). A codebook-free and annotation-free approach for fine-grained image categorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3466–3473. 2, 46, 51

Zhang, N., J. Donahue, R. B. Girshick, e T. Darrell (2014). Part-based r-cnns for fine-grained category detection. *Computer Vision and Pattern Recognition abs/1407.3867*. 47, 51, 105, 119, 121

Zhang, N. e R. Farrell (2013). Deformable part descriptors for fine-grained recognition and attribute prediction. *Computer Vision (ICCV)*, 729–736. 41, 51, 119, 121

Zhang, N., R. Farrell, e T. Darrell (2012). Pose pooling kernels for sub-category recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, Number c, pp. 3665–3672. 40, 51, 103, 119

Zhu, C., C. E. Bichot, e L. Chen (2011). Color orthogonal local binary patterns combination for image region description. *Rapport technique RR-LIRIS-2011-012, LIRIS*. 19, 100, 146

Zhu, C., C.-E. C.-E. Bichot, e L. Chen (2010, August). Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition. *2010 20th International Conference on Pattern Recognition* (2), 3065–3068. 19, 100, 146

Apêndice A

Ambiente de desenvolvimento

Este capítulo apresenta as principais ferramentas utilizadas no desenvolvimento de código necessário para realização deste trabalho. As linguagens de programação, editores, bibliotecas, softwares e *frameworks* são indicados a seguir:

1. **Linguagem de programação C** - em conjunto com o editor *Microsoft Visual Studio 10*.

2. **Linguagem de programação Python** - *Python 2.7.3* em conjunto com o editor de texto multiplataforma *Geany*.

3. **Biblioteca OpenCV** - *Computer Vision Open Source* (OpenCV) oferece uma infraestrutura de desenvolvimento para aplicações acadêmicas ou comerciais que adotem Visão Computacional. As funções necessários foram acessadas inicialmente, por meio da linguagem C, em seguida pela linguagem Python.

4. **Biblioteca VLFeat** - agrupa a implementação de alguns dos principais algoritmos de Visão Computacional do Estado da Arte. As funções necessários foram acessadas por meio do Matlab.

5. **Cuda Convnet** - Ferramenta em linguagem CUDA, executada em arquitetura massivamente paralela de placa gráfica denominada GPU (*Graphics Processing Unit*). As funções necessários foram acessadas por meio da Linguagem Python.

6. **Matrix Laboratory** (MatLab) - A *Student Version* R2013 foi utilizada para extração de características, atividades de formatação, comparação e representação de resultados.

7. **Music Analysis Retrieval and SYnthesis for Audio Signals** - *Framework* utilizado para extrair características dos áudios dos cantos dos pássaros.

Apêndice B

Conjuntos de dados

Este capítulo apresenta as espécies pertencentes aos conjuntos de dados CUB 200 Welinder et al. (2010) e CUB 200-2011 Wah et al. (2011). A Tabela B.1 apresenta 100 espécies numeradas entre 1 (*001.Black footed Albatross*) e 100 (*100.Brown Pelican*). Em seguida a Tabela B.2 apresenta as demais 100 espécies numeradas entre 101 (*101.White Pelican*) e 200 (*200.Common Yellowthroat*).

As espécies que compõem os subconjuntos de imagens descritos na seção 4.1 do capítulo 4 são detalhadas:

- CUB 200 - 2 classes: fazem parte deste subconjunto as espécies aleatoriamente escolhidas: *001.Black footed Albatross* e *035.Purple Finch*.
- CUB 200 - 5 classes: fazem parte deste subconjunto as espécies aleatoriamente escolhidas: *001.Black footed Albatross*, *035.Purple Finch*, *070.Green Violetear*, *105.Whip poor Will* e *180.Wilson Warbler*.
- CUB 200 - 17 classes: fazem parte deste subconjunto as espécies aleatoriamente escolhidas: *001.Black footed Albatross*, *005.Crested Auklet*, *010.Red winged Blackbird*, *015.Lazuli Bunting*, *020.Yellow breasted Chat*, *025.Pelagic Cormorant*, *030.Fish Crow*, *035.Purple Finch*, *040.Olive sided Flycatcher*, *045.Northern Fulmar*, *050.Eared Grebe*, *055.Evening Grosbeak*, *060.Glaucous winged Gull*, *065.Slaty backed Gull*, *070.Green Violetear*, *105.Whip poor Will*, *180.Wilson Warbler*.

Para a implementação do método de informação visual e acústica o conjunto de áudios CUB 50 Songs foi construído. As espécies pertencentes a CUB 50 Songs são também pertencente a CUB 200 e CUB 200 2011. A Tabela B.3 apresenta a listagem das 50 espécies (a quantidade de amostras pertencentes a uma classes poderá variar de 15 a 22 amostras).

Tabela B.1: Listagem numerada (entre 1 e 100) das classes e espécies relativas aos conjuntos e subconjuntos de imagens.

Número da classe e nome da espécie	
001.Black footed Albatross	051.Horned Grebe
002.Laysan Albatross	052.Pied billed Grebe
003.Sooty Albatross	053.Western Grebe
004.Groove billed Ani	054.Blue Grosbeak
005.Crested Auklet	055.Evening Grosbeak
006.Least Auklet	056.Pine Grosbeak
007.Parakeet Auklet	057.Rose breasted Grosbeak
008.Rhinoceros Auklet	058.Pigeon Guillemot
009.Brewer Blackbird	059.California Gull
010.Red winged Blackbird	060.Glaucous winged Gull
011.Rusty Blackbird	061.Heermann Gull
012.Yellow headed Blackbird	062.Herring Gull
013.Bobolink	063.Ivory Gull
014.Indigo Bunting	064.Ring billed Gull
015.Lazuli Bunting	065.Slaty backed Gull
016.Painted Bunting	066.Western Gull
017.Cardinal	067.Anna Hummingbird
018.Spotted Catbird	068.Ruby throated Hummingbird
019.Gray Catbird	069.Rufous Hummingbird
020.Yellow breasted Chat	070.Green Violetear
021.Eastern Towhee	071.Long tailed Jaeger
022.Chuck will Widow	072.Pomarine Jaeger
023.Brandt Cormorant	073.Blue Jay
024.Red faced Cormorant	074.Florida Jay
025.Pelagic Cormorant	075.Green Jay
026.Bronzed Cowbird	076.Dark eyed Junco
027.Shiny Cowbird	077.Tropical Kingbird
028.Brown Creeper	078.Gray Kingbird
029.American Crow	079.Belted Kingfisher
030.Fish Crow	080.Green Kingfisher
031.Black billed Cuckoo	081.Pied Kingfisher
032.Mangrove Cuckoo	082.Ringed Kingfisher
033.Yellow billed Cuckoo	083.White breasted Kingfisher
034.Gray crowned Rosy Finch	084.Red legged Kittiwake
035.Purple Finch	085.Horned Lark
036.Northern Flicker	086.Pacific Loon
037.Acadian Flycatcher	087.Mallard
038.Great Crested Flycatcher	088.Western Meadowlark
039.Least Flycatcher	089.Hooded Merganser
040.Olive sided Flycatcher	090.Red breasted Merganser
041.Scissor tailed Flycatcher	091.Mockingbird
042.Vermilion Flycatcher	092.Nighthawk
043.Yellow bellied Flycatcher	093.Clark Nutcracker
044.Frigatebird	094.White breasted Nuthatch
045.Northern Fulmar	095.Baltimore Oriole
046.Gadwall	096.Hooded Oriole
047.American Goldfinch	097.Orchard Oriole
048.European Goldfinch	098.Scott Oriole
049.Boat tailed Grackle	099.Ovenbird
050.Eared Grebe	100.Brown Pelican

Tabela B.2: Listagem numerada (entre 101 e 200) das classes e espécies relativas aos conjuntos e subconjuntos de imagens.

Número da classe e nome da espécie	
101.White Pelican	151.Black capped Vireo
102.Western Wood Pewee	152.Blue headed Vireo
103.Sayornis	153.Philadelphia Vireo
104.American Pipit	154.Red eyed Vireo
105.Whip poor Will	155.Warbling Vireo
106.Horned Puffin	156.White eyed Vireo
107.Common Raven	157.Yellow throated Vireo
108.White necked Raven	158.Bay breasted Warbler
109.American Redstart	159.Black and white Warbler
110.Geococcyx	160.Black throated Blue Warbler
111.Loggerhead Shrike	161.Blue winged Warbler
112.Great Grey Shrike	162.Canada Warbler
113.Baird Sparrow	163.Cape May Warbler
114.Black throated Sparrow	164.Cerulean Warbler
115.Brewer Sparrow	165.Chestnut sided Warbler
116.Chipping Sparrow	166.Golden winged Warbler
117.Clay colored Sparrow	167.Hooded Warbler
118.House Sparrow	168.Kentucky Warbler
119.Field Sparrow	169.Magnolia Warbler
120.Fox Sparrow	170.Mourning Warbler
121.Grasshopper Sparrow	171.Myrtle Warbler
122.Harris Sparrow	172.Nashville Warbler
123.Henslow Sparrow	173.Orange crowned Warbler
124.Le Conte Sparrow	174.Palm Warbler
125.Lincoln Sparrow	175.Pine Warbler
126.Nelson Sharp tailed Sparrow	176.Prairie Warbler
127.Savannah Sparrow	177.Prothonotary Warbler
128.Seaside Sparrow	178.Swainson Warbler
129.Song Sparrow	179.Tennessee Warbler
130.Tree Sparrow	180.Wilson Warbler
131.Vesper Sparrow	181.Worm eating Warbler
132.White crowned Sparrow	182.Yellow Warbler
133.White throated Sparrow	183.Northern Waterthrush
134.Cape Glossy Starling	184.Louisiana Waterthrush
135.Bank Swallow	185.Bohemian Waxwing
136.Barn Swallow	186.Cedar Waxwing
137.Cliff Swallow	187.American Three toed Woodpecker
138.Tree Swallow	188.Pileated Woodpecker
139.Scarlet Tanager	189.Red bellied Woodpecker
140.Summer Tanager	190.Red cockaded Woodpecker
141.Artic Tern	191.Red headed Woodpecker
142.Black Tern	192.Downy Woodpecker
143.Caspian Tern	193.Bewick Wren
144.Common Tern	194.Cactus Wren
145.Elegant Tern	195.Carolina Wren
146.Forsters Tern	196.House Wren
147.Least Tern	197.Marsh Wren
148.Green tailed Towhee	198.Rock Wren
149.Brown Thrasher	199.Winter Wren
150.Sage Thrasher	200.Common Yellowthroat

Tabela B.3: Listagem numerada (entre 1 e 200) das 50 classes relativas ao conjunto de áudios.

039.Least Flycatcher	152.Blue headed Vireo
048.European Goldfinch	154.Red eyed Vireo
054.Blue Grosbeak	155.Warbling Vireo
057.Rose breasted Grosbeak	156.White eyed Vireo
070.Green Violetear	157.Yellow throated Vireo
076.Dark eyed Junco	159.Black and white Warbler
085.Horned Lark	160.Black throated Blue Warbler
088.Western Meadowlark	161.Blue winged Warbler
097.Orchard Oriole	162.Canada Warbler
099.Ovenbird	164.Cerulean Warbler
105.Whip poor Will	165.Chestnut sided Warbler
109.American Redstart	166.Golden winged Warbler
114.Black throated Sparrow	167.Hooded Warbler
116.Chipping Sparrow	169.Magnolia Warbler
118.House Sparrow	170.Mourning Warbler
119.Field Sparrow	171.Myrtle Warbler
120.Fox Sparrow	179.Tennessee Warbler
121.Grasshopper Sparrow	180.Wilson Warbler
129.Song Sparrow	183.Northern Waterthrush
131.Vesper Sparrow	184.Louisiana Waterthrush
136.Barn Swallow	193.Bewick Wren
139.Scarlet Tanager	195.Carolina Wren
140.Summer Tanager	196.House Wren
148.Green tailed Towhee	199.Winter Wren
149.Brown Thrasher	200.Common Yellowthroat

B.0.1 Definição dos subconjuntos de espécies

A representação adotada ao longo deste documento utiliza o conjunto de dados CUB 200 2011 completo e subconjuntos com 50, 17, 5 e 2 classes. Embora os subconjuntos não sejam utilizados como objetivo principal da análise de resultados eles podem oferecer informações complementares importantes relacionadas ao conjunto completo.

Alguns experimentos adicionais foram conduzidos para verificar se a escolha aleatória de espécies poderia impactar nos resultados dos métodos implementados. Os novos subconjuntos são descritos a seguir e organizados em três grandes grupos: a) Grupo 1 - classes aleatoriamente escolhidas (trata-se dos subconjuntos para os quais os valores são descritos ao longo deste documento); b) Grupo 2 - novos subconjuntos que consideram classes pertencentes a mesma família e que possuem alguma similaridade visual; c) Grupo 3 - novos subconjuntos de espécies que não possuem qualquer similaridade visual.

As espécies que compõem os subconjuntos CUB 200 - 2 classes:

- **Grupo 1 -** fazem parte deste grupo as duas espécies aleatoriamente escolhidas: *001.Black footed Albatross* e *035.Purple Finch*;
- **Grupo 2 -** fazem parte deste grupo as duas espécies que possuem alguma similaridade: *122.Harris Sparrow* e *123.Henslow Sparrow*;
- **Grupo 3 -** fazem parte deste grupo as duas espécies que não possuem similaridade: *091.Mockingbird* e *115.Brewer Sparrow.*

As espécies que compõem os subconjuntos CUB 200 - 5 classes:

- **Grupo 1 -** fazem parte deste grupo as cinco espécies aleatoriamente escolhidas: *001.Black footed Albatross*, *035.Purple Finch*, *070.Green Violetear*, *105.Whip poor Will*, *180.Wilson Warbler.*
- **Grupo 2 -** fazem parte deste grupo as cinco espécies que possuem alguma similaridade: emph117.Clay colored Sparrow, *122.Harris Sparrow*, *123.Henslow Sparrow*, *123.Henslow Sparrow*, *124.Le Conte Sparrow*, *125.Lincoln Sparrow.*
- **Grupo 3 -** fazem parte deste grupo as espécies que não possuem similaridade: *005.Crested Auklet*, *010.Red winged Blackbird*, *015.Lazuli Bunting*, *091.Mockingbird*, *115.Brewer Sparrow.*

As espécies que compõem os subconjuntos CUB 200 - 17 classes:

- **Grupo 1 -** fazem parte deste grupo as dezessete espécies aleatoriamente escolhidas: *001.Black footed Albatross*, *005.Crested Auklet*, *010.Red winged Blackbird*, *015.Lazuli Bunting*, *020.Yellow breasted Chat*, *025.Pelagic Cormorant*, *030.Fish Crow*, *035.Purple Finch*, *040.Olive sided Flycatcher*, *045.Northern Fulmar*, *050.Eared Grebe*, *055.Evening Grosbeak*, *060.Glaucous winged Gull*, *065.Slaty backed Gull*, *070.Green Violetear*, *105.Whip poor Will*, *180.Wilson Warbler*.

- **Grupo 2** - fazem parte deste grupo dezessete espécies que possuem alguma similaridade (todas espécies da família *Sparrow*): *113.Baird Sparrow*, *114.Black throated Sparrow*, *115.Brewer Sparrow*, *116.Chipping Sparrow*, *117.Clay colored Sparrow*, *118.House Sparrow*, *119.Field Sparrow*, *120.Fox Sparrow*, *121.Grasshopper Sparrow*, *122.Harris Sparrow* , *123.Henslow Sparrow*, *125.Lincoln Sparrow*, *126.Nelson Sharp tailed Sparrow*, *127.Savannah Sparrow*, *128.Seaside Sparrow*, *132.White crowned Sparrow*, *133.White throated Sparrow*.

- **Grupo 3** - fazem parte deste grupo as dezessete espécies que não possuem similaridade: *091.Mockingbird*, *115.Brewer Sparrow*, *005.Crested Auklet*, *010.Red winged Blackbird*, *015.Lazuli Bunting*, *039.Least Flycatcher*, *048.European Goldfinch*, *054.Blue Grosbeak*, *088.Western Meadowlark*, *097.Orchard Oriole*, *099.Ovenbird*, *139.Scarlet Tanager*, *149.Brown Thrasher*, *156.White eyed Vireo*, *184.Louisiana Waterthrush*, *195.Carolina Wren*, *200.Common Yellowthroat*.

A Tabela B.4 apresenta os resultados obtidos para classificação de espécies de pássaros utilizando características de cores (abordagem superficial) para o subconjuntos de 2, 5, e 17 classes. A organização dessas classes foi conduzida de três diferentes formas: aleatório, similar e não similar. Observa-se que os conjuntos aleatórios conseguem representar o problema de mameira adequada quando comparado a um cenário real. Os conjuntos que possuem forte similaridade apontam o pior desempenho para o método automático de identificação de espécies de pássaros. Isso pode ser explicado por meio da alta granularidade entre as espécies e toda a problemática associada quando essa situação é representada. Em seguida, quando o conjunto de espécies não similares é comparado aos demais, pode-se afirmar que a constatação anterior permanece, pois, ocorre a diminuição de representação de granularidade no problema. Neste caso, as espécies diferentes possibilitam a extração de características que podem promover maior discriminação entre as classes.

A Tabela B.5 apresenta os resultados obtidos para a mesma forma de organização dos subconjuntos utilizando uma abordagem profunda (CNN)[1] para o subconjuntos de 2,

[1]Arquitetura 1 indicada na seção C.0.6.

Tabela B.4: Resumo da classificação de espécies de pássaros utilizando características de cores para o subconjuntos de 2, 5, e 17 classes em diferentes formas de organização. Os resultados foram obtidos por meio de histograma de cores de 30 faixas, classificador SVM - RBF (c, g) otimizados em relação as imagens recortadas pelos valores de caixa delimitadora.

Taxa de correta classificação (%)			
	Escolha da espécie	Grupo	(%)
2 classes	aleatório	Grupo 1	81,66
	similar	Grupo 2	75,00
	não similar	Grupo 3	61,01
5 classes	aleatório	Grupo 1	66,18
	similar	Grupo 2	33,33
	não similar	Grupo 3	35,87
17 classes	aleatório	Grupo 1	25,79
	similar	Grupo 2	11,08
	não similar	Grupo 3	19,26

5, e 17 classes. Os resultados mostram que o conjunto aleatório consegue representar o problema de identificação automática de espécies de pássaros com menor grau de interferência do que nos casos em que a granularidade fina é evidente. Os resultados obtidos sustentam a apresentação dos demais experimentos descritos neste documento por meio dos subconjuntos de imagens indicados na seção 4.1 do capítulo 4.

Tabela B.5: Resumo da classificação de espécies de pássaros utilizando abordagem profunda para o subconjuntos de 2, 5, e 17 classes em diferentes formas de organização. Os resultados foram obtidos por meio de uma arquitetura CNN.

Taxa de correta classificação (%)			
	Escolha da espécie	Grupo	(%)
2 classes	aleatório	Grupo 1	90,00
	similar	Grupo 2	78,80
	não similar	Grupo 3	60,83
5 classes	aleatório	Grupo 1	74,82
	similar	Grupo 2	36,39
	não similar	Grupo 3	55,01
17 classes	aleatório	Grupo 1	50,96
	similar	Grupo 2	16,84
	não similar	Grupo 3	46,16

Apêndice C

Parâmetros experimentais

Este capítulo apresenta alguns dos principais experimentos realizados para definição de parâmetros experimentais: i) porcentagem de borda utilizado pelo algoritmo de segmentação; ii) imagem completa em relação a imagem recortada por valores de caixa delimitadora; iii) faixas no histograma de cor; iv) parâmetros e operadores para texturas; v) definição de arquitetura para CNN.

C.0.2 Porcentagem de borda

Para a implementação do algoritmo de segmentação por meio das cores presentes nas imagens um importante parâmetro é a porcentagem de borda considerada para separação do objeto de interesse do restante da imagem. Diferentes porcentagens foram verificadas. A Tabela C.1 apresenta os resultados obtidos para as 20 primeiras classes do conjunto CUB 200 por meio de histograma de cores de 30 faixas e imagem sem segmentação. A Tabela C.2 apresenta os resultados obtidos para o conjunto CUB 200 completo por meio de histograma de cores de 30 faixas e imagem sem segmentação.

Tabela C.1: Experimento 1: imagens sem segmentação para diferentes espaços de cores.

Taxa de correta classificação (%)		
Classificador	RGB	HSV
SVM - RBF $(c,\ g)$ otimizados	8,6%	15,5%

Tabela C.2: Experimento 2: imagens sem segmentação para diferentes espaços de cores.

Taxa de correta classificação (%)		
Classificador	RGB	HSV
SVM - RBF $(c,\ g)$ otimizados	1,7%	7,8%

A Tabela C.3 apresenta exemplos do impacto da porcentagem de borda atribuída ao algorítimo de segmentação. Ambos os espaços de cores foram considerados para verificar a taxa de correta segmentação descrita na seção 4.7.2. Observa-se que o valor de 2% apresenta o melhor compromisso para as taxas de correta segmentação. Valores inferiores a 2% não conseguem obter cores suficiente para a separação do objeto de interesse do restante da imagem. Por outro lado, valores superiores a 2%, em geral consideram muitas cores, confundindo o objeto de interesse com o fundo da imagem.

Tabela C.3: Experimento 3: variação da porcentagem de borda.

Taxa de correta classificação (%)		
Porcentagem da borda	RGB	HSV
2% da imagem	11,3%	18,2%
5% da imagem	9,36%	14,%
10% da imagem	5,7%	9,8%

C.0.3 *Caixa delimitadora*

Concluídos os experimentos com a segmentação baseada em cores e de posse das importantes conclusões discutidas na seção 6.1 e subseção 6.1.1, observou-se a necessidade de verificar o impacto do uso da imagem recortada pelos valores de caixa delimitadora em relação ao uso da imagem completa. A Tabela C.4 apresenta os resultados obtidos para o conjunto CUB 200 completo por meio de histograma de cores de 30 faixas, imagens sem segmentação e classificador SVM - RBF $(c,\ g)$ otimizados em relação as imagens recortadas pelos valores de caixa delimitadora.

Tabela C.4: Experimento 4: impacto do uso da imagem recortada pelos valores de caixa delimitadora em relação ao uso da imagem completa.

Taxa de correta classificação (%)		
Tipo da imagem	RGB	HSV
CUB 200 - original	1,5%	1,8%
CUB 200 - recorte BB	2,9%	4,3%

C.0.4 Histogramas de cores

No capítulo 4 a seção 4.4 e a subseção 4.4.1.1 indicam o uso de 30 faixas no histograma de cores (sendo 10 de cada canal do espaço de cores indicados). A Tabela C.5 apresenta a variação na taxa de correta classificação para o conjunto CUB 200 completo

Tabela C.5: Experimento 5: impacto do uso de diferentes quantidades de faixas no histograma de cores.

Taxa de correta classificação (%)		
Quantidade de faixas	RGB	HSV
30	1,7%	3,7%
300	1,6%	3,7%

por meio de histograma de cores de 30 faixas e em seguida para 300 faixas. Para esse experimento foram utilizadas imagens sem segmentação e classificador SVM - RBF (c, g) otimizados.

C.0.5 Parâmetros e operadores para texturas

Os parâmetros e operadores utilizados para extração de texturas do conjunto CUB 200 ou CUB 200-2011 são fundamentados na literatura. Os operadores $LBP^{u2}_{P,R}$, $LBP^{ri}_{P,R}$, $LBP^{riu2}_{P,R}$ são executados com os valores 8 e 12 para P e 2 e 1.5 para R (opções indicadas na literatura: (Zhu et al., 2010), (Zhu et al., 2011) e (Ojala et al., 2000)).

C.0.6 Arquiteturas CNN

Nos experimentos inicias algumas arquiteturas foram testadas para a identificação de espécies. Essas arquiteturas foram inspiradas nos trabalhos de (Krizhevsky e Hinton, 2009), (Krizhevsky et al., 2012), (Ciresan et al., 2012). A Figura 4.10 da seção 2.5 detalhou cada uma das arquiteturas a organização e quantidade de camadas.

A Tabela C.6 apresenta detalhes dos resultados obtidos durante 10 execuções, seguido da variância para imagens de 32x32 *pixels*. A Tabela C.7 para imagens de 64x64 *pixels*. O resultado é descrito por meio da *taxa de incorreta classificação* obtida para cada uma das arquiteturas. Desta forma, as menores taxas indicam a Arquitetura 1 e imagens de 64x64 *pixels* como mais adequada para solução do problema de identificação de espécies de pássaros.

Os resultados descritos neste documento são obtidos por meio da Arquitetura 1, de imagens de 64x64 pixels, considerando a utilização de *dropout*. A execução que mais se aproxime do valor médio de 10 execuções é utilizada para expressar os resultados finais das implementações das abordagens profundas. O procedimento indicado é o mesmo utilizado para os subconjuntos e também para conjunto completo de 200 classes.

Tabela C.6: Detalhamento das execuções para as 5 arquiteturas utilizando imagens de 32x32 *pixels*.

Taxa de incorreta classificação (%)											
Arquitetura	Execução 1	Execução 2	Execução 3	Execução 4	Execução 5	Execução 6	Execução 7	Execução 8	Execução 9	Execução 10	Variância
ARQUITETURA 1	87,50	87,94	87,21	87,31	87,16	88,26	87,60	87,62	87,63	87,52	0,1
ARQUITETURA 2	87,74	85,90	85,89	85,18	-	85,98	84,93	-	-	-	1,1
ARQUITETURA 3	90,85	90,12	89,87	89,68	89,60	89,88	89,93	87,35	88,94	88,94	0,3
ARQUITETURA 4	87,69	87,96	87,74	88,24	86,69	87,52	86,47	87,45	87,31	87,55	0,0
ARQUITETURA 5	87,21	87,47	87,13	86,69	86,38	87,6	86,70	87,48	87,11	86,98	0,0

Tabela C.7: Detalhamento das execuções para as 5 arquiteturas utilizando imagens de 64x64 *pixels*.

Taxa de incorreta classificação (%)											
Arquitetura	Execução 1	Execução 2	Execução 3	Execução 4	Execução 5	Execução 6	Execução 7	Execução 8	Execução 9	Execução 10	Variância
ARQUITETURA 1	85,17	85,55	84,73	85,52	84,39	84,98	85,80	85,86	84,96	85,54	0,2
ARQUITETURA 2	-	-	-	-	-	-	-	-	-	-	-
ARQUITETURA 3	87,78	87,50	87,97	87,36	87,84	86,79	86,87	87,47	87,18	87,94	0,1
ARQUITETURA 4	85,2	87,90	86,06	86,26	86,82	86,20	86,8	85,67	85,96	86,48	1,9
ARQUITETURA 5	87,58	90,00	90,53	90,01	89,39	87,47	89,38	87,72	89,14	86,87	2,5