Recognition of Pantanal Animal Species using Convolutional Neural Networks

Diogo Nunes Gonçalves, Mauro dos Santos de Arruda, Lucas Abreu da Silva,
Reinaldo Felipe Soares Araujo, Bruno Brandoli Machado, Wesley Nunes Gonçalves
Universidade Federal de Mato Grosso do Sul - CPPP
Ponta Porã, MS, 79907-414
{dnunesgoncalves,maurosantosarruda}@gmail.com, {lucas10df, felipeiex}@hotmail.com,
{bruno.brandoli, wesley.goncalves}@ufms.br

***Resumo*—Pantanal is one of the most important biomes of the world, with a large number of animal species. To reconcile the demand of development and biodiversity conservation, it is important to catalog the species in the region and the impact of development on the population of animals. However, the task of identifying the species of animals is time-consuming and depends on manual inspection of the images. To overcome this issue, this paper proposes a methodology for animal specie recognition using convolutional neural networks. Experimental results were performed using a database with 14.547 images divided into 13 animals species showing a regognition rate larger than 92% when combining Resnet-152 and SVM classifier.**

I. Introdução

O Pantanal é uma região natural localizada principalmente dentro dos estados brasileiros de Mato Grosso e Mato Grosso do Sul, mas se estende até a Bolívia e Paraguai. Esta região é a maior área úmida tropical do mundo, com área de aproximadamente 138.183 km^2. O nome Pantanal vem da palavra "Pântano", que significa zonas úmidas, porque 80% das planícies do Pantanal estão submersas durante as estações chuvosas.

Pantanal é um dos ecossistemas mais ricos do mundo, que abriga um grande número de animais que vivem em equilíbrio ecológico. Junk et al. [1] lista 263 espécies de peixes, 96 espécies de répteis, 40 espécies de anfíbios, 390 espécies de aves e 130 espécies de mamíferos. O número de espécies diverge em vários trabalhos, por exemplo, o número de espécies de aves geralmente é dado como algo entre 600 e 700 espécies [2], embora apenas 390 foram confirmados [1]. Devido à sua importância e diversidade ecológica, o Pantanal é considerado pela UNESCO como Patrimônio Mundial e Reservas da Biosfera.

Como resultado do uso insustentável da terra, o Pantanal sofre muita perda de biodiversidade e dos seus habitats naturais associados. Nessa região existem muitas espécies raras e ameaçadas de extinção, como o tatu-canastra (Priodontes maximus), cervo-pantanal (Blastocerus dichotomus), tamanduá-bandeira (Myrmecophaga tridactyla), cachorro-do-mato (Speothos venaticus), entre outras. O grande desafio para o Pantanal é conciliar a crescente demanda para o desenvolvimento social e econômico (por exemplo a pecuária, gado, agricultura, turismo), com a conservação da biodiversidade [3].

Desta forma, conhecer a biodiversidade é essencial para os esforços dos investigadores que trabalham para proteger o Pantanal e as espécies ameaçadas de extinção. A identificação de espécies de animais no Pantanal leva tempo devido ao ambiente de zonas úmidas e aos hábitos alimentares dos animais, o que torna difícil estimar de forma eficaz as espécies e suas densidades por km^2. Para acompanhar e monitorar a população, os pesquisadores passam dias no Pantanal observando as espécies com a ajuda de tecnologias, como a armadilha com câmera, as imagens aéreas, etc. O processo de identificação atual é extremamente demorado, requer treinamento especial e depende de uma inspeção manual das imagens.

Este artigo propõe uma metodologia para o reconhecimento de espécies do pantanal usando redes neurais convolucionais (*convolutional neural network* - CNN). Para isso, a imagem é pré-processada e passa como entrada de uma CNN. A saída da penúltima camada da CNN é utilizada como vetor de características. Os vetores de características alimentam o treinamento de classificadores supervisionados, como as máquinas de vetores de suporte. Experimentos foram realizados em uma base com 14.547 imagens divididas em 13 espécies importantes. Quatro arquiteturas de CNN e dois classificadores foram avaliados obtendo porcentagem de classificação correta de 97.24%.

Este artigo está descrito em 6 seções. A Seção II apresenta uma revisão de literatura sobre reconhecimento de espécies de animais em visão computacional e redes neurais convolucionais. A metodologia proposta é descrita em detalhes na Seção III. A Seção IV apresentas os experimentos e resultados enquanto que a Seção V apresenta as conclusões e os trabalhos futuros.

II. Revisão da Literatura

A. Reconhecimento de Espécies de Animas em Visão Computacional

O reconhecimento de espécies animais é um grande desafio na área de visão computacional, principalmente pela característica do ambiente em que o animal se encontra e pela sua movimentação constante. Algumas abordagens têm sido propostas para superar esses desafios. Um método de captura muito usado é a armadilha fotográfica que basicamente é um sensor especialmente visual que grava as imagens dos animais que se movem através de seu campo de visão [4], [5], [6], elas são usadas em conjunto com métodos de visão computacional para a identificação das espécies animais.

Mauro et al. [7] usou um método de segmentação usando superpixels e componentes conexos em imagens térmicas para obter regiões da imagem contendo animais silvestres. Aguzzi et al. [4] usou as armadilhas fotográficas na captura das imagens, e aplicou o descritor de Fourier juntamente com o método de k vizinhos mais próximos para a identificação das espécies através de contornos dos animais na imagem. Yu et al [5] propôs uma identificação automatizada de espécies para as imagens capturadas por armadilhas fotográficas usando SIFT (transformada de características invariante a escala) e pirâmide espacial. Eles testaram o método em um banco de imagens com 18 espécies, obtendo uma precisão média de 82%.

Algumas aplicações mais relacionadas a este trabalho foram de Chen et al. [8] e Gómez et al. [9], que propuseram métodos de identificação de espécies de animais usando redes neurais convolucionais. Chen et al. [8] usou um banco de imagens com 20 espécies comuns na América do Norte, chegando a 38% de precisão com um CNN de apenas 6 camadas. Por outro lado, Gómez et al. [9] alcançou 98% de precisão em um banco de imagens com 26 espécies da Tanzânia. Eles testaram seis arquiteturas CNNs, dentre elas a AlexNet [10], VGGNet [11] e GoogLeNet [12].

B. Redes Neurais Convolucionais (Convolutional Neural Networks - CNN)

As redes neurais convolucionais consistem em camadas convolucionais, na qual é incorporado o algoritmo de *back-propagation* que aprende os parâmetros de cada camada. Desde a sua criação, a CNN tem sido caracterizada por três propriedades básicas, as conexões locais, o compartilhamento de peso e o *pooling* local. As duas primeiras propriedades permitem que o modelo aprenda os padrões visuais locais importantes com menos parâmetros ajustáveis que o modelo totalmente conectado, e a terceira propriedade prepara a rede para possuir invariância à translação.

Uma das primeiras aplicações de redes neurais convolucionais é a rede LeNet-5 descrita por LeCun et al. [13] para o reconhecimento óptico de caracteres. Comparado às redes convolucionais profundas atuais, a rede foi relativamente modesta devido aos recursos computacionais limitados da época e aos desafios do treinamento de algoritmos para redes maiores. Embora houvesse muito potencial em redes convolucionais mais profundas, só recentemente elas se tornaram predominantes, seguindo o aumento do poder computacional atual, da quantidade de dados para treinamento disponíveis na Internet e da necessidade de desenvolvimento de métodos mais eficazes para a formação de tais modelos.

Um exemplo recente e notável do uso de redes convolucionais profundas para classificação de imagens é o desafio Imagenet [10] em que uma CNN obteve um erro consideravelmente menor comparado com o erro das abordagens tradicionais de visão computacional (usando SIFT e Máquinas de Vetores de Suporte). Redes neurais convolucionais também obtiveram recentemente sucesso para diferentes aplicações, incluindo estimação de pose humana [14], análise de faces [15], detecção de pontos-chave facial [16], reconhecimento de voz [17] e classificação de ação [18].

III. METODOLOGIA PROPOSTA

A metodologia proposta para reconhecimento de espécies de animais pode ser descrita em três etapas principais: i) pré-processamento da imagem de entrada; ii) extração de características com redes neurais convolucionais; e iii) treinamento supervisionado. As três etapas são descritas em detalhes nas subseções abaixo.

A. Pré-processamento da Imagem de Entrada

A primeira etapa consiste em pré-processar as imagens de entrada. Como as imagens estão em diferentes tamanhos e resoluções, uma imagem I é redimensionada para um tamanho fixo ($W \times H \times 3$). O tamanho fixo deve corresponder aos requerimentos de entrada de cada arquitetura das redes neurais convolucionais (e.g., $227 \times 227 \times 3$ para AlexNet e $224 \times 224 \times 3$ para Resnet). O segundo pré-processamento consiste em subtrair uma cor média μ de todos os pixels da imagem I:

$$I(x, y, R) = I(x, y, R) - \mu_R \quad (1)$$

$$I(x, y, G) = I(x, y, G) - \mu_G \quad (2)$$

$$I(x, y, B) = I(x, y, B) - \mu_B \quad (3)$$

onde μ é a cor média de todas as imagens usadas para treinar a rede neural convolucional.

B. Extração de Características

A arquitetura das CNNs contém milhões de parâmetros (e.g., 60 milhões para AlexNet), tornando inviável e problemático o treinamento em uma base com poucas imagens. Para contornar esse problema, Oquab et al. [19] propuseram o uso das camadas internas de CNN treinadas em grandes bases de imagens como extrator de características. Para usar uma CNN pré-treinada, a última camada que consiste em um classificador softmax foi desconsiderada e as camadas restantes foram utilizadas como um extrator de características. Nos experimentos foram utilizadas quatro arquiteturas de CNNs treinadas na ImageNet ILSVRC com 1.28 milhões de imagens.

Para ilustrar a extração de características, considere a CNN conhecida como AlexNet proposta por Krizhevsky et al. [10]. A AlexNet é composta por cinco camadas convolucionais $C1 \ldots C5$ seguida por três camadas totalmente conectadas $FC6 \ldots FC8$. A entrada da rede consiste em uma imagem $Y_0 = I \in R^{227 \times 227 \times 3}$. Em cada camada convolucional, a entrada da camada Y_{k-1} é convoluída com um conjunto de filtros treinados. Em seguida, aplica-se a unidade ReLU (*Rectified Linear Unit*) que consiste em pegar o máximo entre a convolução e 0, conforme:

$$Y_k = \sigma(conv(Y_{k-1}, W_k, B_k)) \quad (4)$$

onde (W_k, B_k) é o conjunto de filtros treinados e $\sigma(Y_k) = max(0, Y_k)$ é a unidade ReLU.

As três últimas camadas totalmente conectadas calculam $Y_6 = \sigma(W_6 Y_5 + B_6)$, $Y_7 = \sigma(W_7 Y_6 + B_7)$ e $Y_8 = \psi(W_8) Y_7 + B_8$, onde Y_k denota a saída da k-ésima camada e $\psi(X)[i] = e^{X[i]} / \sum_j e^{X[j]}$ é o classificador softmax. Para obter as características, nós descartamos a última camada que classifica as características Y_7 nas classes em que a rede foi

treinada. Dessa forma, as características consistem na saída da penúltima camada $Y_7 \in \Re^{4096}$ para a rede AlexNet.

As demais arquiteturas utilizadas neste trabalho são descritas abaixo. Para todas as camadas, a saída da penúltima camada é utilizada como o vetor de características.

- **VGGNet** [11]: a principal contribuição da VGGNet foi mostrar que CNN com mais camadas pode melhorar a acurácia. Os autores apresentaram e treinaram duas arquiteturas, uma com 16 camadas e outra com 19 camadas. A penúltima extrai vetores com 4096 valores.
- **GoogLeNet** [12]: esta arquitetura apresenta o módulo *Inception* que reduz o número de parâmetros da CNN (e.g., de 60M da AlexNet para 4M). Dessa forma, os autores apresentaram uma CNN mais profunda com 27 camadas. Vetores com 1024 valores são extraídos na penúltima camada.
- **ResNet** [20]: esta arquitetura propôs o aprendizado residual para treinar redes cada vez mais profundas. Os autores forneceram experimentos usando 50, 101 e 152 camadas, que são substancialmente mais profundas do que as arquiteturas descritas acima. A penúltima camada extrai vetores com 2048 valores.

C. Treinamento Supervisionado

Dados os vetores de características extraídos com uma CNN e suas respectivas classes, um classificador pode ser treinado. Os classificadores tradicionais, tais como máquina de vetores de suporte e vizinho mais próximo, foram utilizados nesta etapa. Portanto, essa etapa da metodologia proposta substitui a última camada das redes que são treinadas especificamente para as classes da base de imagens.

IV. Experimentos e Resultados

Nesta seção são descritos os experimentos, a base de imagens e os resultados obtidos. Para a validação da metodologia proposta, os experimentos foram realizados com imagens obtidas da Imagenet [21]. A Imagenet é um conjunto de imagens organizado de acordo com a hierarquia *WordNet*. Neste conjunto existem mais de 100.000 classes com média de 1.000 imagens para cada classe obtidas da internet. Portanto, não existe qualquer controle com relação a resolução e o dispositivo de captura. Desse conjunto, foram selecionadas 13 espécies de animais do pantanal conforme apresentado na Tabela I. A base de imagens possui um total de 14.547 imagens em que cada espécie possui de 621 a 1.430 imagens em diferentes resoluções, ambientes e dispositivos de captura, iluminação, etc. A Figura 1 apresenta quatro exemplos de cada espécie em que é possível observar a variação intra-classe.

Para a classificação, utilizaram-se dois classificadores bem conhecidos na literatura: Vizinho mais próximo (*Nearest Neighbor* - NN) [22] e a Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) [23] com núcleo polinomial. Para o treinamento e testes, as imagens foram divididas seguindo a validação cruzada em 10 dobras. Dessa forma, o conjunto de imagens foi particionado aleatoriamente em 10 subconjuntos com tamanho aproximadamente iguais e mantendo a proporção entre o número de imagens para cada

Tabela I. Base de imagens utilizada.

Nome Científico	Nome Popular	N. imagens
Ardea occidentalis	Garça Branca	1.231
Caiman yacare	Jacaré-do-Pantanal	1.430
Caracara plancus	Carcará	1.133
Cariama cristata	Seriema	834
Dasypus novemcinctus	Tatu-galinha	975
Eunectes murinus	Sucuri	1.229
Hydrochoerus hydrochaeris	Capivara	1.365
Jabiru mycteria	Tuiuiú	1.049
Myrmecophaga jubata	Tamanduá-bandeira	1.015
Panthera Onca	Onça-pintada	1.514
Pyrocephalus rubinus	Príncipe	1.029
Rhea americana	Ema	1.122
Tapirus terrestris	Anta	621

Tabela II. Resultados experimentais para arquiteturas de CNN e classificadores.

CNN	NN	SVM
AlexNet	86.75(±1.08)	92.35(±0.58)
VGGNet-16	91.10(±0.72)	95.55(±0.42)
VGGNet-19	91.61(±0.59)	95.52(±0.60)
GoogLeNet	92.69(±0.70)	95.61(±0.39)
Resnet-50	94.47(±0.53)	96.67(±0.27)
Resnet-101	94.95(±0.37)	96.93(±0.41)
Resnet-152	95.58(±0.60)	97.24(±0.38)

classe. Dos 10 subconjuntos, 9 subconjuntos são utilizados para treinar o classificador enquanto que o subconjunto restante é utilizado para avaliação. O processo é repetido 10 vezes tomando um subconjunto para avaliação exatamente uma vez. As porcentagens de classificação correta são utilizadas para calcular a média e o desvio padrão.

A. Resultados

A Tabela II apresenta os resultados obtidos pela abordagem proposta usando diferentes CNNs e os dois classificadores. Para todas as CNNs, o SVM obteve melhores resultados se comparado ao NN. A maior diferença entre os classificadores foi para a AlexNet, i.e., de 86.75%(±1.08) com NN para 92.35%(±0.58) com SVM. Os melhores resultados obtidos para os classificadores NN e SVM foram respectivamente 95.58%(±0.60) e 97.24%(±0.38) ambos para a Resnet com 152 camadas. Esses resultados são extremamente promissores devido às dificuldades impostas pelas imagens.

Com relação as CNNs, podemos observar que os melhores resultados foram obtidos pela AlexNet, VGGNet, GoogLeNet e Resnet, nessa ordem. Com relação a Resnet, os resultados mostraram que quanto mais camadas a CNN possui, melhor é o seu resultado. Por exemplo, a Resnet com 152 camadas (Resnet-152) obteve 97.24%(±0.38) com o classificador SVM, enquanto que a mesma CNN com 50 e 101 camadas (Resnet-50 e Resnet-101) obteve 96.67%(±0.27) e 96.93%(±0.41), respectivamente.

Para avaliar detalhadamente o desempenho das CNNs, a Figura 2 apresenta as matrizes de confusão para AlexNet, VGGNet-16, GoogLeNet e Resnet-152 com SVM. Em uma matriz de confusão, os acertos são armazenados na diagonal principal. Como podemos observar, as espécies com maior acerto em todas as CNNs foram o Príncipe (Figura 1(b)) devido a sua coloração específica e a onça-pintada (Figura 1(i)) devido a sua textura. Por outro lado, o Tuiuiú (Figura 1(f)) obteve as menores taxas de 88.37%, 91.99%, 93.33% e

(a) Ema (b) Príncipe (c) Carcará (d) Jacaré-do-Pantanal (e) Sucuri

(f) Tuiuiú (g) Garça Branca (h) Seriema (i) Onça-pintada (j) Capivara

(k) Anta (l) Tatu-galinha (m) Tamanduá-bandeira

Figura 1. Quatro exemplos para cada uma das 13 espécies que compõem o banco de imagens.

94.85% para AlexNet, VGGNet-16, GoogLeNet e Resnet-152, respectivamente. Os erros foram obtidos devido à similaridade do Tuiuiú com a Garça Branca. Na Resnet-152, por exemplo, 3.43% das imagens do Tuiuiú foram classificadas como Garça Branca.

B. Custo Computacional

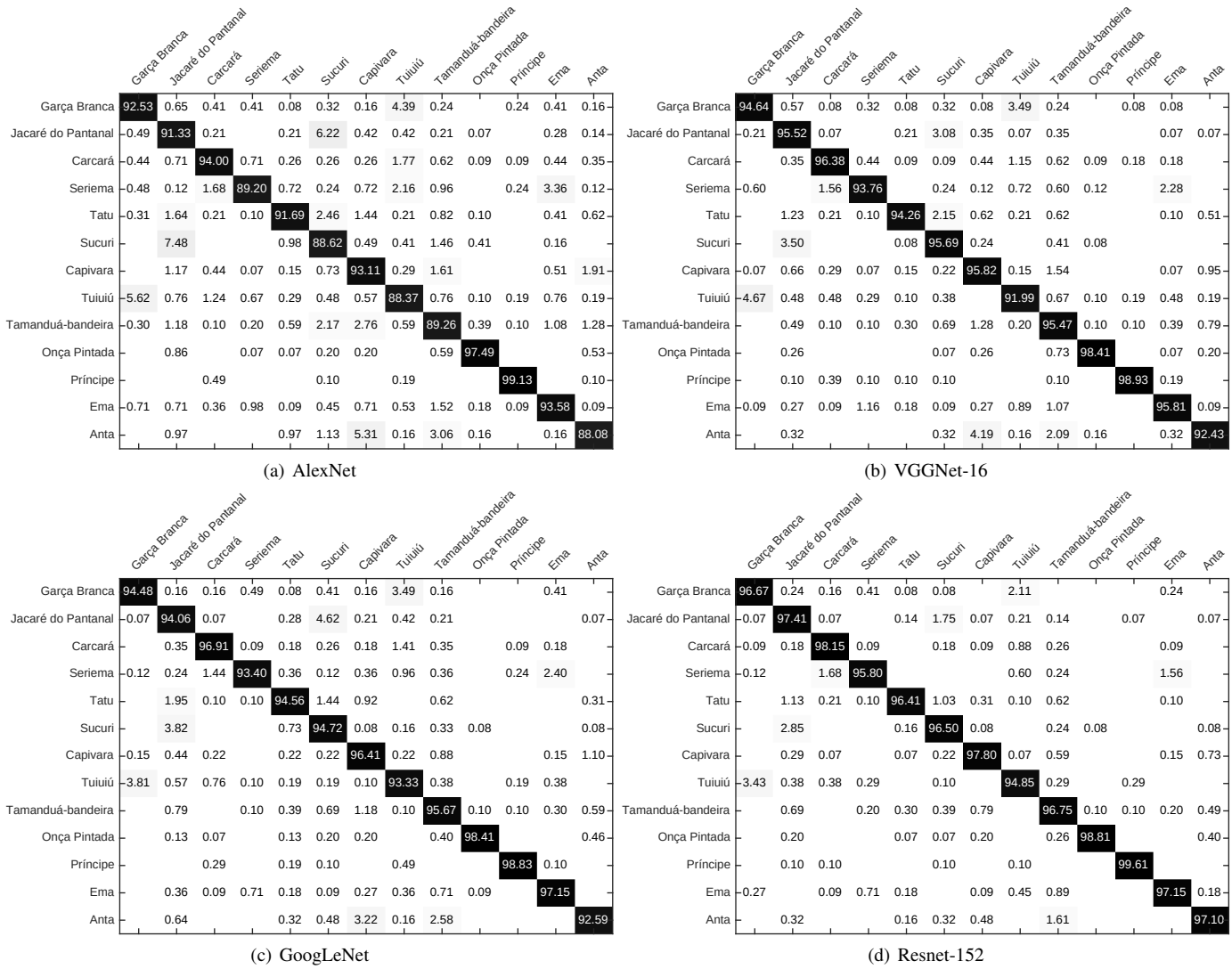
Para medir o tempo para a aplicação da metodologia proposta, experimentos foram executados em um computador com processador Intel i5 1.6GHz e 8GB de memória RAM. Vale ressaltar que todos esses resultados foram obtidos rodando apenas na CPU, isto é, não foi utilizada nenhuma GPU que deve acelerar ainda mais o processo.

A Tabela III apresenta o tempo médio em segundos para as etapas: pré-processamento da imagem de entrada (E1), extração das características pela CNN (E2), classificação com o SVM (E3) e o tempo total para as três etapas. Os tempos de pré-processamento (E1) e classificação com o SVM (E3) são baixos e similares para todas as CNNs. Como esperado, a AlexNet possui o melhor tempo de processamento, pois possui menos camadas. O tempo médio total para reconhecer uma imagem usando a AlexNet foi de 0.055s, tornando possível a sua aplicação em tempo real. As demais CNNs também tiveram tempos totais aceitáveis, como a GoogLeNet com 0.182s e a VGGNet-16 com 0.399s. Por fim, a Resnet com 152 camadas (CNN mais profunda avaliada) possui um tempo médio total de 0.585s.

C. Classificação de Quadros do Youtube

Para verificar o poder de generalização da metodologia, vídeos do youtube contendo diferentes espécies foram obtidos, seus quadros foram extraídos e classificados usando a Resnet-152 e o classificador SVM. Para treinar a metodologia, todas as imagens da ImageNet foram utilizadas.

A Figura 3 ilustra alguns quadros dos vídeos contendo as espécies: anta, carcará, tamanduá-bandeira, onça-pintada,



(a) AlexNet

(b) VGGNet-16

(c) GoogLeNet

(d) Resnet-152

Figura 2. Matrizes de confusão para diferentes CNNs e classificador SVM.

Tabela III. TEMPO MÉDIO PARA CADA UMA DAS ETAPAS DA METODOLOGIA PROPOSTA: PRÉ-PROCESSAMENTO DA IMAGEM DE ENTRADA - E1, EXTRAÇÃO DE CARACTERÍSTICAS PELA CNN - E2 E CLASSIFICAÇÃO COM O SVM - E3.

CNN	E1	E2	E3	Tempo Total
AlexNet	0.012(±0.01)	0.043(±0.002)	0.001(±0.00)	0.055(±0.012)
VGGNet-16	0.013(±0.01)	0.386(±0.003)	0.001(±0.00)	0.399(±0.014)
VGGNet-19	0.012(±0.01)	0.457(±0.003)	0.001(±0.00)	0.469(±0.014)
GoogLeNet	0.012(±0.01)	0.170(±0.002)	0.000(±0.00)	0.182(±0.013)
Resnet-50	0.012(±0.01)	0.229(±0.004)	0.001(±0.00)	0.241(±0.014)
Resnet-101	0.012(±0.01)	0.396(±0.006)	0.001(±0.00)	0.408(±0.015)
Resnet-152	0.012(±0.01)	0.572(±0.007)	0.001(±0.00)	0.585(±0.015)

seriema e tuiuiú. No topo de cada quadro, é apresentada a espécie mais provável classificada pela metodologia e sua porcentagem obtida pelo classificador SVM. Apesar dos desafios (e.g., escalas e ponto de vista), a metodologia apresenta uma excelente acurácia para as espécies.

V. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi proposta uma nova metodologia para reconhecimento de espécies de animais por meio da obtenção de característica de uma rede neural convolucional. Os resultados obtidos através desta nova metodologia demonstraram ser possível uma caracterização robusta em uma base de imagens bastante complexa. O melhor resultado de 97.24% foi obtido usando a Resnet com 152 camadas e o classificador SVM.

Os trabalhos futuros incluem a aplicação da abordagem em outras bases de imagens, e utilizar outros classificadores para avaliar os resultados. Além disso, pretendemos combinar os vetores de características extraídos de diferentes CNNs.

AGRADECIMENTOS

Os autores agradecem a Fundação Universidade Federal de Mato Grosso do Sul, ao programa PET - Fronteira, Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul - Fundect pelo apoio financeiro e o espaço cedido para a realização das pesquisas e experimentos.

Figura 3. Quadros de vídeos do youtube classificados pela Resnet-152 e SVM. Na parte superior do vídeo é apresentada a espécie mais provável seguida pela porcentagem.

REFERÊNCIAS

[1] W. J. Junk, C. N. da Cunha, K. M. Wantzen, P. Petermann, C. Strüssmann, M. I. Marques, and J. Adis, "Biodiversity and its conservation in the pantanal of mato grosso, brazil," *Aquatic Sciences*, vol. 68, no. 3, pp. 278–309, 2006. [Online]. Available: http://dx.doi.org/10.1007/s00027-006-0851-4

[2] F. A. Swarts, *The Pantanal of Brazil, Paraguay and Bolivia: Selected Discourses on the Worlds Largest Remaining Wetland System*. Hudson MacArthur Publishers, January 2000.

[3] C. Alho and J. Sabino, "A conservation agenda for the Pantanal's biodiversity," *Brazilian Journal of Biology*, vol. 71, pp. 327 – 335, 04 2011.

[4] J. Aguzzi, C. Costa, Y. Fujiwara, R. Iwase, E. Ramirez-Llorda, and P. Menesatti, "A novel morphometry-based protocol of automated video-image analysis for species recognition and activity rhythms monitoring in deep-sea fauna," *Sensors*, vol. 9, no. 11, pp. 8438–8455, 2009.

[5] X. Yu, J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang, "Automated identification of animal species in camera trap images," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–10, 2013. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2013-52

[6] R. Kays, S. Tilak, B. Kranstauber, P. A. Jansen, C. Carbone, M. J. Rowcliffe, T. Fountain, J. Eggert, and Z. He, "Monitoring wild animal communities with arrays of motion sensitive camera traps," *ArXiv e-prints*, September 2010.

[7] M. dos Santos de Arruda, B. B. Machado, W. N. Gonçalves, L. Cullen, J. H. P. Dias, C. C. Garcia, and J. F. R. Jr, "Thermal image segmentation in studies of wildlife animals," in *Workshop de Visão Computacional*, 2015, pp. 204–209.

[8] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," in *IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 858–862.

[9] A. Gomez, A. Salazar, and F. Vargas, "Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks," *ArXiv e-prints*, mar 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv e-prints*, September 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[13] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 2016/05/28 1989. [Online]. Available: http://dx.doi.org/10.1162/neco.1989.1.4.541

[14] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1653–1660.

[15] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 2480–2487.

[16] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3476–3483.

[17] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1725–1732.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1717–1724. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.222

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[22] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995, cited By 12145.