Image Inspection of Railcar Structural Components: An approach through Deep Learning and Discrete Fourier Transform

Rafael. L. Rocha[1,2], Cleison D. Silva[1], Ana Claudia S. Gomes[2], Bruno V. Ferreira[2], Eduardo C. Carvalho[2], Ana Carolina Q. Siravenha[3] e Schubert R. Carvalho[3]

[1] Universidade Federal do Pará, Brazil
rafael.rocha@itec.ufpa.br, cleison@ufpa.br
[2] Instituto SENAI de Inovação em Tecnologias Minerais, Brazil
claudia.isi@senaipa.org.br, bruno.isi@senaipa.org.br,
eduardo.isi@sesipa.org.br
[3] Instituto Tecnológico Vale, Brazil
schubert.carvalho@itv.org, ana.siravenha@pq.itv.org

Abstract.

Railcar components inspection is one of the most critical tasks in railway maintenance. The use of image processing, coupled with machine learning, has emerged as a solution for replacing current standard methodologies. The spectral analysis gives the frequency representation of a signal and has been largely used in signal processing tasks. In this sense, this work proposes the evaluation of the use of the Discrete Fourier Transform (DFT) in addition to the spatial representation image of railcar components for an automatic detector of defective parts performed by Convolutional Neural Network (CNN) classification. The results are given in measures of accuracy, precision, recall, and F1-score metrics in addition to the accuracy boxplot, and showed that the use of the DFT increase in 1.04% the CNN classification accuracy.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning

Keywords: railcar inspection, convolutional neural network, discrete Fourier transform, image classification

1. INTRODUÇÃO

O vagão ferroviário é um dos ativos mais importantes de uma ferrovia, uma vez que pode ser empregado tanto no transporte de cargas quanto no transporte de pessoas. Tamanha importância é refletida na necessidade de uma severa inspeção dos componentes que os formam, a fim de, principalmente, prevenir acidentes [Macucci et al. 2016].

Em grande parte das empresas do setor, a inspeção de componentes é realizada de forma visual por um técnico operacional posicionado ao longo da ferrovia. Ainda que venha demonstrando eficiência na maioria dos casos, a inspeção visual está suscetível a erros causados, por exemplo, por uma distração do inspetor. Contribui para falhas na inspeção o ambiente insalubre ao qual esse trabalhador está inserido [Hart et al. 2008; Park et al. 1996].

A utilização de imagens para inspeção de componentes defeituosos na área ferroviária é comumente encontrada na literatura. O Aprendizado de Máquina (*Machine Learning*), através do classificador de Máquina de Vetores de Suporte (*Support Vector Machine*, SVM), é empregado na classificação

de imagens para detecção de presença ou ausência da chave de contenção em trens de carga após a extração de características fundamentais das imagens, como mostra [Liu et al. 2016]. Em especial, modelos de Aprendizado Profundo (do inglês, *Deep Learning*) como em [Gibert et al. 2017], classifica a imagem do fixador do trilho ferroviário em relação a suas possíveis condições em campo: ausente, normal ou danificado. Estudos como este empregam Redes Neurais Convolucionais (*Convolutional Neural Network*, CNN) na tarefa de classificação valendo-se de sua robustez em tratar imagens obtidas em condições ambientais adversas; diferenças de luminosidade, sujeira na lente da câmera e obstruções, são exemplos dos efeitos aos quais essas imagens estão sujeitas [Park et al. 2016; Ravikumar et al. 2011].

O presente trabalho apresenta técnicas baseadas em aprendizado profundo aplicadas à inspeção automática de componentes de vagões ferroviários. O objetivo é, a partir de uma CNN, reconhecer padrões associados a defeitos em componentes de vagões a partir de imagens dos domínios espacial (imagem original) e da frequência, que é obtida a partir da Transformada Discreta de Fourier (*Discrete Fourier Transform*, DFT). A rede deverá ser capaz de indicar se há ou não a presença de defeitos no componente conhecido como *pad*, um dos elementos responsáveis por suavizar o atrito entre a roda e o vagão [IWnicki 2006]. Os estados possíveis a serem identificados são (1) *pad* normal, i.e., a peça está em perfeito estado, (2) *pad* danificado, quando há indicação de avaria que pode gerar algum transtorno, e (3) *pad* ausente. Em particular, a ausência do *pad* advém da diferença entre os projetos de vagões de fabricantes diferentes que percorrem essas vias.

A metodologia apresentada analisará a influência da informação frequencial da imagem no processo classificatório. Propõe-se então, que a imagem a ser inserida na rede neural seja formada pela composição de três elementos: (1) a imagem no domínio espacial, em escala de cinza, (2) a informação de magnitude e (3) a informação de fase, ambas obtidas a partir da transformada de Fourier da imagem original. A avaliação do desempenho da rede será dada pela análise das métricas de acurácia, precisão, *recall* e *F1-score*, além do *boxplot* das acurácias.

Este artigo está organizado da seguinte forma: A Seção 2 apresenta a fundamentação teórica do trabalho, com os conceitos relacionados ao aprendizado profundo e redes neurais convolucionais, assim como, aqueles inerentes à transformada discreta de Fourier. A Seção 3 apresenta a descrição do componente analisado e da base de dados utilizada no trabalho, além das configurações dos experimentos realizados. Na Seção 4, os resultados dos experimentos são apresentados e discutidos. Finalmente, a Seção 5 expõe a conclusão do trabalho e a sugestão para novas investigações.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Aprendizado profundo e redes neurais convolucionais

O aprendizado profundo tem como objetivo a aquisição de conhecimento através da percepção da hierarquia de conceitos, ou seja, aprender conceitos complexos (como o reconhecimento de um objeto) por meio de conceitos mais simples (como bordas, cantos e contornos) [Goodfellow et al. 2016].

Redes neurais convolucionais destacam-se como uma importante metodologia para reconhecimento de padrões e classificação de imagens. As CNNs exploram a correlação espacial local impondo um padrão de conectividade local entre os neurônios das camadas adjacentes. Além disto, possuem propriedades essenciais no reconhecimento de padrões: um elevado grau de invariância à translação, escala, inclinações e outras distorções no padrão a ser classificado [Haykin 2009].

A estrutura básica de uma CNN é formada por sucessivas camadas convolucionais e de *max-pooling*, seguidas de uma ou mais camadas totalmente conectadas (perceptron de múltiplas camadas) as quais efetivam a classificação dos padrões. A extração de características dos padrões é realizada na camada convolucional, onde os valores de saída da operação são limitados por uma função de ativação, tipicamente através da função Unidade Linear Retificada (*Rectified Linear Unit*, ReLU). As camadas

max-pooling são responsáveis por reduzir (através do máximo local) o tamanho da entrada, mantendo apenas os elementos mais relevantes.

2.2 Transformada discreta de Fourier

A série de Fourier é uma das principais contribuições do matemático Jean Baptiste Joseph Fourier, que estabelece que qualquer função periódica pode ser representada através da soma de senos e/ou cossenos de variadas frequências harmonicamente relacionadas [Gonzalez and Woods 2006]. Com relação às funções não periódicas e que tenham área sobre a curva finita, é possível escrevê-las como uma integral de senos e/ou cossenos multiplicada por uma função ponderada, essa caracterização é intitulada de transformada de Fourier [Gonzalez and Woods 2006].

Considerando que esse trabalho utiliza imagens digitais, a DFT adotada nesse caso é a bidimensional, descrita na Equação 1, na qual a imagem de tamanho MxN é definida pela função $f(x, y)$, onde x e y são coordenadas espaciais, $F(u, v)$ é a representação da imagem no domínio da frequência, que também pode ser chamado de domínio de Fourier, enquanto u e v são as coordenadas no domínio da frequência. Tanto os valores das coordenadas espaciais quanto os das coordenadas na frequência variam de $0, 1, 2, ..., M-1$ para x e u, e de $0, 1, 2, ..., N-1$ para y e v.

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi(\frac{ux}{M} + \frac{vy}{N})} \tag{1}$$

Considerando que o resultado da transformada discreta de Fourier tem quantidades complexas como resposta, torna-se conveniente trabalhar $F(u, v)$ em termos de coordenadas polares, que são obtidas a partir da parte real $R(u, v)$ e da parte imaginária $I(u, v)$. Desse modo, é possível trabalhar a DFT de duas maneiras, através da magnitude ou espectro (Equação 2), e fase ou espectro de fase (Equação 3).

$$|F(u,v)| = \left[R^2(u,v) + I^2(u,v)\right]^{\frac{1}{2}} \tag{2}$$

$$\phi(u,v) = tan^{-1}\left[\frac{I(u,v)}{R(u,v)}\right] \tag{3}$$

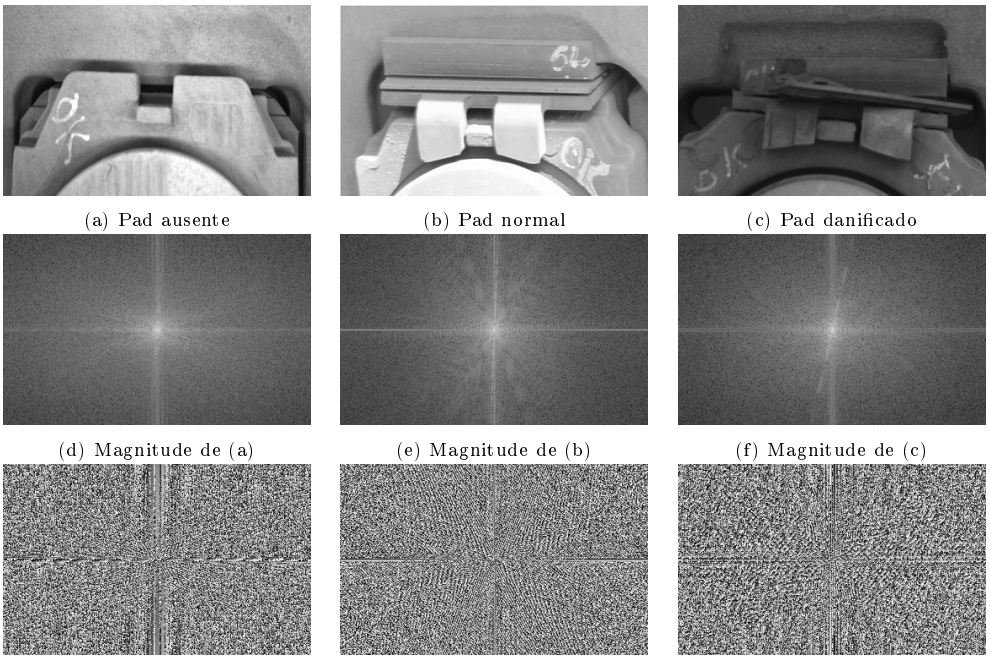
As Figuras 1a-1i apresentam as respostas da DFT nas imagens do componente analisado neste trabalho. As Figuras 1a, 1b e 1c representam as imagens originais em escala de cinza (domínio espacial). A magnitude da DFT das imagens originais é retratada nas Figuras 1d, 1e e 1f, enquanto a fase da DFT pode ser encontrada nas Figuras 1g, 1h e 1i.

3. MATERIAIS E MÉTODOS

3.1 Base de dados

O componente estrutural do vagão ferroviário analisado neste trabalho é o *pad*, um polímero inserido sobre o adaptador de rolamento, que desempenha um importante papel na dinâmica dos vagões, funcionando como suspensão primária [IWnicki 2006]. Qualquer tipo de defeito neste componente pode comprometer suspensão e vir a ocasionar acidentes, logo é um componente de imprescindível inspeção.

A imagem do componente investigado é obtida através de uma câmera localizada ao longo da ferrovia. As imagens capturadas pela câmera apresentam diferentes níveis de intensidade, iluminação e contraste, já que as imagens foram adquiridas em diferentes períodos do dia, condições climáticas,



(a) Pad ausente (b) Pad normal (c) Pad danificado

(d) Magnitude de (a) (e) Magnitude de (b) (f) Magnitude de (c)

(g) Fase de (a) (h) Fase de (b) (i) Fase de (c)

Fig. 1: Imagens que compõe as três classes utilizadas representando os possíveis estados do *pad*: ausente (a), normal (b) e danificado (c). Magnitude e fase da transformada discreta de Fourier do *pad* ausente ((d) e (g)), do *pad* normal ((e) e (h)) e do *pad* danificado, ((f) e (i)).

além de estar sujeita à poeira e lama do local de captura. A partir de cada imagem capturada, a região ao redor do *pad* é segmentada para formar base de dados utilizada neste trabalho.

Os três possíveis estados em que o componente pode ser encontrado caracterizam as três classes que são respectivamente: *pad* ausente (Figura 1a) (classe 1); *pad* normal (Figura 1b) que é o *pad* com ausência de defeito (classe 2); por fim, o *pad* danificado (Figura 1c) com ruptura ou deslocamento em relação à posição esperada (classe 3).

A base de dados utilizada para fins de classificação possui um total de 1976 imagens com resolução de 32×64, sendo 651 da classe 1, 644 da classe 2 e 681 da classe 3. Nenhum aumento artificial de dados foi implementado, assim como nenhum pré-processamento das imagens.

3.2 Modelo de aprendizagem e configuração dos experimentos

A arquitetura da CNN utilizada nos experimentos é mostrada na Figura 2, onde a camada de entrada possui dimensão $a \times l \times p$, onde $a = 32$, $l = 64$ e $p = 1, 2, 3$. A profundidade da rede muda de acordo com o teste em execução, no qual testadas as seguintes entradas: (a) imagem original em tons de cinza, (b) magnitude da imagem original, (c) fase da imagem original, (d) a combinação entre a magnitude e a fase, (e) a combinação entre a imagem original e sua magnitude, (f) a combinação entre a imagem original e sua fase, e (g) a combinação entre a imagem original e sua magnitude e fase. Assim, a dimensão das entradas a, b e c é $32 \times 64 \times 1$, das entradas d, e e f, $32 \times 64 \times 2$ e, finalmente, a dimensão da entrada g é igual a $32 \times 64 \times 3$.

A primeira camada convolucional possui um *kernel* de tamanho 2×2 e 32 filtros de saída aplicados à função de ativação ReLU. A segunda camada convolucional possui 64 filtros de mesmo tamanho. A camada *max-pooling* possui ambos *pool* e passo de tamanho 2×2. A camada totalmente conectada possui 128 neurônios regularizada em sua saída. Uma camada do tipo *softmax* mapeia os

mapas de características em termos de probabilidade entre as três classes. O treinamento da rede é feito através do otimizador denominado Descida do Gradiente Estocástica (SGD, *Stochastic Gradient Descent*) [Bishop 2006].

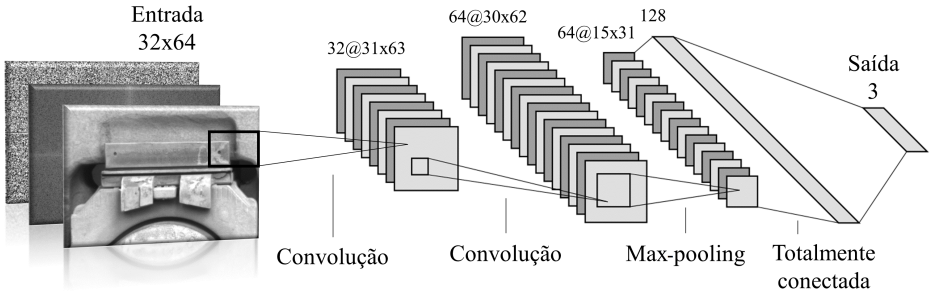


Fig. 2: Arquitetura da rede neural convolucional com seis camadas utilizada nos experimentos. A entrada de rede exemplifica o uso da imagem do componente (a frente) e da magnitude (meio) e fase (atrás) da transformada discreta de Fourier.

Para a realização dos experimentos, 80% das imagens foram destinadas ao treinamento enquanto 20% foram separadas para realização do teste do modelo, o que equivale à 1580 e 396 imagens, respectivamente. Os resultados apresentados pelos experimentos, obtidos após 50 épocas de treinamento, representam os resultados médios de 100 treinamentos da rede para cada tipo de entrada.

Os testes com a rede foram feitos em duas etapas. A primeira etapa utilizou como entrada as imagens a, b, c e d, e avaliou a contribuição das informações frequenciais em relação ao dado espacial. A segunda etapa avaliou a influência das combinações entre a imagem no domínio espacial e suas respectivas frequenciais (e, f e g).

4. RESULTADOS E DISCUSSÕES

A identificação de peças defeituosas em vagões de trens é uma das tarefas sensíveis em uma operação ferroviária. Desenvolver um sistema que realize essa tarefa automaticamente, para substituir os modelos manuais de inspeção atuais, é uma tarefa desafiadora. Além dos fatores ambientais em que o sistema de captura esta inserido, a acurácia do sistema de identificação deve ser suficientemente alta para que suporte importantes decisões como, por exemplo, retirar um vagão de operação para a troca de um componente.

Neste trabalho foi analisada a contribuição da informação frequencial para aumentar a acurácia da classificação de um dos componentes da suspensão do vagão, o *pad*. Os resultados são avaliados pelas medidas de acurácia global, precisão, *recall* e F1-score. Formalmente, as métricas são definidas pelas equações:

$$acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{4}$$

$$precisão = \frac{VP}{VP + FP} \tag{5}$$

$$recall = \frac{VP}{VP + FN} \tag{6}$$

$$F1 = \frac{2 * precisão * recall}{precisão \ + \ recall} \tag{7}$$

Onde VP e VN representam as amostras verdadeiramente classificadas para cada classe P (positiva) e N (negativa). FP e FN representam as amostras que foram falsamente atribuídas às classes (erros). A medida de acurácia avalia o comportamento global do sistema, o *recall* avalia o quão frequente a rede classifica como classe P quando realmente uma amostra é da classe P; a precisão avalia daquelas amostras classificadas como corretas, quantas efetivamente eram daquela classe; a medida F1-score também mede a qualidade geral do modelo e sendo robusta inclusive quando a base de dados é desbalanceada.

Table I: Performance da classificação realizada utilizando somente a magnitude e/ou fase da DFT da imagem do componente como entrada da rede em comparação ao uso das imagens originais. As métricas de avaliação são acurácia, precisão, *recall* e F1-score da base de dados de teste.

Imagem	Acurácia	Precisão	*Recall*	F1-score
a	**0,9343**	**0,9363**	**0,9348**	**0,9344**
b	0,8481	0,8621	0,8491	0,8459
c	0,7915	0,7954	0,7922	0,7924
d	0,8789	0,8825	0,8793	0,8794

Na Tabela I são comparados os resultados dos testes cujas entradas foram as imagens dos tipos a, b, c e d. Os testes estão identificados segundo as imagens de entrada para melhor entendimento. A classificação a partir das imagens originais em tons de cinza foi a que obteve melhor resultado, alcançando uma acurácia de $93,43\%$. Na sequência, a combinação dos produtos da DFT (magnitude e fase) atingiram acurácia de aproximadamente 87%, indicando que a informação frequencial pode ser fonte de padrões suficientes para identificar os objetos desejados. As métricas de precisão, *recall* e *F1-score* indicam o mesmo comportamento mostrado pela acurácia.

Isoladamente, a informação de fase foi a que obteve pior performance entre os testes, com um erro médio de aproximadamente 83 das 396 imagens testadas. De fato, a informação de fase da DFT em imagens é uma fonte de informação bastante complexa para a identificação de padrões de forma isolada.

Da magnitude podem-se identificar, por exemplo, a presença ou não de bordas na imagem. Das Figuras 1d-1f vê-se que as imagens dos *pads* possuem componentes de todas as frequências em níveis diferentes, e que, em geral, a magnitude torna-se menor nas altas frequências. As imagens também mostram que há direções dominantes na imagem que perpassam o centro da imagem transformada, representando a existência de padrões regulares na imagem original.

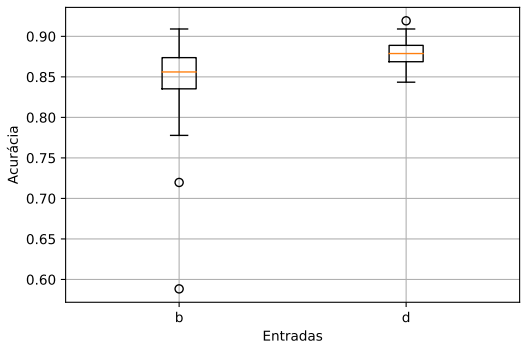


Fig. 3: *Boxplots* comparativos com o resultado de 100 realizações de experimentos com imagens dos tipos b e d.

O valor de cada ponto das Figuras 1g-1i determina a fase da frequência correspondente. Apesar de ser possível identificar as linhas verticais e horizontais correspondentes aos padrões na imagem

original, a imagem da fase não produz muitas novas informações sobre a estrutura da imagem do domínio espacial.

A Figura 3 ilustra a distribuição das amostras de teste dada as entradas do tipo b e d, aquelas que resultaram em acurácias maiores que 80% a partir de dados da DFT. Em relação ao experimento com imagens d, os resultados indicaram uma menor dispersão entre suas execuções, assim como a ocorrência de um único *outlier* que excede o limite superior $(0,9191)$ do *boxplot*. Já o experimento com imagens b apresentou maior variabilidade, além de dois *outliers*, um destes com a acurácia mais baixa de $0,5883$.

A Tabela II apresenta os resultados dos testes com imagens combinadas conforme descrição na Seção 3.2 em comparação, novamente, com a imagem original. Em geral, a união das informações frequenciais com a imagem original superou a acurácia dos testes com a magnitude ou fase, isoladamente ou combinadas.

Table II: Performance da classificação das imagens combinadas entre informações espaciais e frequenciais. Para fins comparativos, os resultados do teste com as imagens originais foram adicionados à tabela.

Imagem	Acurácia	Precisão	Recall	F1-score
a	0,9343	0,9363	0,9348	0,9344
e	**0,9441**	**0,9455**	**0,9444**	**0,9443**
f	0,9332	0,9348	0,9337	0,9337
g	0,9436	0,9446	0,9439	0,9437

Os testes usando imagens do tipo e e g apresentaram acurácia média acima de 94%, sendo que a combinação da imagem original com a sua magnitude atingiu a máxima acurácia global $(94,41\%)$. Consistentemente, as outras avaliações seguem a mesma tendência. A distância entre o melhor e o pior resultado nessa tabela recai na classificação errônea de aproximadamente 4 amostras.

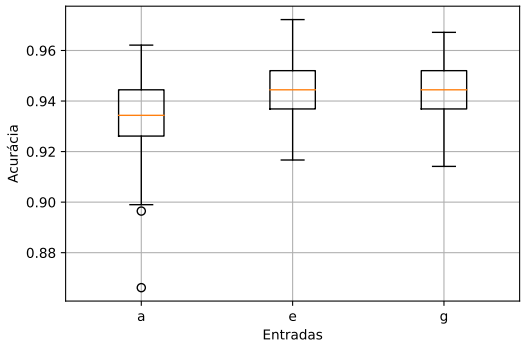


Fig. 4: *Boxplots* comparativos com o resultado de 100 realizações dos experimentos com melhores desempenhos na Tabela II. Os três *boxplots* ilustram os testes com as imagens originais (a), combinação das imagens originais e a magnitude (e) e a combinação entre as imagens originais, suas magnitudes e fases (g).

Os *boxplots* da Figura 4 ilustram os resultados dos experimentos com imagens dos tipos a, e e g. É possível observar a similaridade entre as respostas obtidas durante as execuções dos experimentos e e g que não apresentam *outliers*. Além disso, o valor de mediana (linha laranja) é bastante similar entre elas.

Dos valores de acurácia possíveis, a utilização dos dados das imagens originais e suas magnitudes variam de pouco menos de 92% até pouco mais de 97%, isso mostra que ainda existem ajustes que podem ser feitos para que essa performance seja melhorada.

5. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propôs a utilização da transformada discreta de Fourier como parte da informação aplicada como entrada de uma rede neural convolucional empregada na classificação do componente do vagão ferroviário conhecido como *pad*. Os experimentos exploraram o poder da identificação dos padrões a partir de sete testes onde foram variadas as entradas de uma CNN.

A CNN implementada possui duas camadas convolucionais seguidas de uma camada totalmente conectada. A camada de saída mapeava os pesos da rede entre três classes, *pad* ausente, *pad* normal e *pad* defeituoso. Os resultados mostram que isoladamente as informações da DFT não contribuem para a melhoria da classificação das amostras em comparação ao uso da imagem no domínio espacial. A combinação dessa imagem com sua magnitude, no entanto, eleva a acurácia da classificação em aproximadamente $1,0\%$, o que equivale em média a quatro imagens a mais classificadas corretamente, representando um ganho significativo em um sistema em tempo real de inspeção do *pad*.

Como parte das oportunidades de extensão desta metodologia, visualiza-se a inclusão de novas amostras com a aquisição de mais imagens *in loco*, a implementação de técnicas de melhoramento das imagens, a inclusão de modelos de aumento de dados (*data augmentation*) como forma de tornar o modelo mais robusto às condições adversas possíveis em uma aplicação real e a comparação com outros classificadores para realização da inspeção do *pad*. Espera-se, também, aumentar a robustez do modelo perante imagens que apresentem oclusão parcial, como simulação de eventos extremos em que somente imagens parciais precisem ser usadas para a detecção de problemas nessa peça.

Agradecimentos

Os autores agradecem o apoio financeiro do Instituto SENAI de Inovação em Tecnologias Minerais (ISI-TM), assim como o apoio logístico do Instituto Tecnológico Vale (ITV) e da Vale S.A. para obtenção da base de dados utilizada por este trabalho.

REFERENCES

Bishop, C. M. *Pattern recognition and machine learning*. Springer, 2006.

Gibert, X., Patel, V. M., and Chellappa, R. Deep multitask learning for railway track inspection. *IEEE Transactions on Intelligent Transportation Systems* 18 (1): 153–164, 2017.

Gonzalez, R. C. and Woods, R. E. *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2006.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Hart, J., Resendiz, E., Freid, B., Sawadisavi, S., Barkan, C., and Ahuja, N. Machine vision using multi-spectral imaging for undercarriage inspection of railroad equipment. In *Proceedings of the 8th World Congress on Railway Research, Seoul, Korea*, 2008.

Haykin, S. *Neural networks and learning machines*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

IWnicki, S. *Handbook of Railway Vehicle Dynamics*. CRC Press, 2006.

Liu, L., Zhou, F., and He, Y. Automated visual inspection system for bogie block key under complex freight train environment. *IEEE Transactions on Instrumentation and Measurement* 65 (1): 2–14, 2016.

Macucci, M., Di Pascoli, S., Marconcini, P., and Tellini, B. Derailment detection and data collection in freight trains, based on a wireless sensor network. *IEEE Transactions on Instrumentation and Measurement* 65 (9): 1977–1987, 2016.

Park, B., Chen, Y., Nguyen, M., and Hwang, H. Characterizing multispectral images of tumorous, bruised, skin-torn, and wholesome poultry carcasses. *Transactions of the ASAE* 39 (5): 1933–1941, 1996.

Park, J.-K., Kwon, B.-K., Park, J.-H., and Kang, D.-J. Machine learning-based imaging system for surface defect inspection. *International Journal of Precision Engineering and Manufacturing-Green Technology* 3 (3): 303–310, Jul, 2016.

Ravikumar, S., Ramachandran, K. I., and Sugumaran, V. Machine Learning Approach for Automated Visual Inspection of Machine Components. *Expert Syst. Appl.* 38 (4): 3260–3266, 2011.