A Convolutional Neural Network for Learning Local Feature Descriptors on Multispectral Images

Cristiano F. G. Nunes, Flávio L. C. Pádua

***Abstract*—This work presents a novel convolutional neural network, termed multispectral features network (MF-Net), for learning local feature descriptors in multispectral images. Unlike most existing solutions, which primarily handle images from the visible light spectrum, we propose a learning-based method that deals with image data acquired from different spectrum bands. To design our convolutional neural network, we introduce a new layer that incorporates Log-Gabor filters to enhance the network capability to work with nonlinear intensity variations in images captured from different electromagnetic frequencies spectrum. This layer, entitled mapping layer, can be easily integrated into different network architectures. To demonstrate the efficacy and limitations of our method, we went on experiments with two distinct datasets extensively used in previous works composed of image pairs from the visible spectrum and the infrared spectrum. Experimental results with datasets containing images obtained from visible light and infrared spectrum show that our method can accurately match features, outperforming some state-of-the-art learning-based algorithms.**

***Index Terms*—Local feature descriptor, Multispectral images, Log-Gabor filters, Convolutional neural networks.**

I. Introdução

As imagens multiespectrais desempenham um papel importante na área de Visão Computacional, pois, elas possibilitam a obtenção de informações da cena que não seriam facilmente obtidas por imagens individuais. Nesse sentido, as imagens multiespectrais, por possuírem mais informações sobre determinado objeto ou cena, possibilitam a análise mais completa em diversas aplicações, incluindo sistemas de vigilância [1], navegação guiada por imagem, monitoramento de desastres naturais [2], reconhecimento facial [3], dentre outras.

Para usar efetivamente um conjunto de imagens obtidas de distintas faixas do espectro, é necessário realizar um alinhamento espacial, comumente realizado por meio da correspondência de características locais. Uma característica local consiste em um padrão de imagem que difere de sua vizinhança imediata, como: pontos, cantos, bordas ou algum outro padrão específico [4].

A descrição das características locais ainda é um desafio devido a alguns fatores, como: diferenças de iluminação, obstruções, mudanças no ângulo de visão, presença de ruídos ou borrões nas imagens [5], [6]. Ambiguidades na cena também se tornam um fator desafiador, pois, diferentes regiões podem ter aparências similares, quando observadas sob pontos de vista distintos, dificultando assim sua discriminação e a correspondência de suas projeções em diferentes imagens [7].

Cristiano F. G. Nunes, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brasil, cfgnunes@cefetmg.br.

Flávio L. C. Pádua, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brasil, cardeal@cefetmg.br.

No contexto de imagens multiespectrais, as tarefas de detecção e descrição de características possuem um desafio ainda maior, posto que a relação entre as intensidades dos pixels das diferentes imagens pode ser não-linear [5], [8]. Para ilustrar as diferenças de intensidade dos pixels entre imagens multiespectrais, a Fig. 1 exibe os valores dessas intensidades ao longo de um segmento de reta (em vermelho) para um par de imagens obtidas em faixas distintas do espectro eletromagnético. Os segmentos de retas indicados na imagem possuem regiões correspondentes. Observa-se, por meio das taxas de variação (setas no gráfico) que, as diferenças de intensidade em imagens multiespectrais, geralmente de natureza não-linear, podem afetar a imagem e dificultar a descrição dessas regiões. Apesar dessas diferenças, as aparências das formas dos objetos (por exemplo, suas bordas ou texturas) tendem a permanecer em ambas as imagens [7].

Como as bordas e texturas da mesma região entre imagens capturadas de diferentes faixas do espectro podem definir padrões similares, é possível, portanto, obter características que as descrevam em bandas distintas, mesmo quando as variações de intensidade dos pixels são diferentes. Essa extração pode ser realizada aplicando-se funções para extrair bordas em imagens.

Nas últimas décadas, as Redes Neurais Convolucionais (CNNs) têm recebido muita atenção por seu sucesso em tarefas de Visão Computacional, como classificação de cenas [9], reconhecimento de objetos [10], segmentação de imagens [11], e recuperação de imagens [12]. Segundo [13], as Redes Neurais Convolucionais e outras técnicas de aprendizado de máquina têm sido empregadas em diversos níveis, como em detecção de características, cálculos de similaridade ou até mesmo a estimação de transformação entre imagens. No entanto, devido aos desafios relacionados às imagens obtidas de diferentes faixas do espectro, os modelos existentes ainda não funcionam adequadamente para imagens multiespectrais [14].

No contexto das imagens obtidas de diferentes faixas do

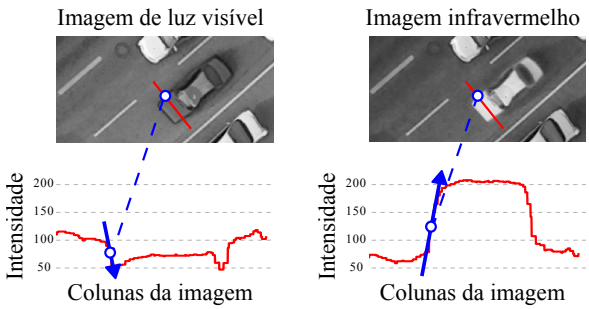

Fig. 1. Variação nos valores de intensidade dos pixels entre imagens obtidas de diferentes faixas do espectro.

espectro, uma das limitações das CNNs consiste no fato de que a distribuição espacial dos pixels na rede é essencialmente tratada de forma linear, dado que não há nenhuma camada em uma CNN típica capaz de lidar diretamente com a não-linearidade entre pixels vizinhos de uma imagem. Pois, ao se utilizar imagens obtidas de diferentes faixas do espectro, essas intensidades podem não seguir um mesmo padrão, conforme exemplificado na Fig. 1.

Com base nessa limitação e motivado pelo sucesso recente das CNNs para imagens obtidas do espectro visível, este trabalho propõe uma nova CNN projetada para aprender descritores de características locais em imagens multiespectrais.

Podemos resumir as contribuições deste trabalho da seguinte forma:

- Projetamos uma arquitetura de CNN para aprender descritores de características locais em imagens multiespectrais. Além disso, verificamos esse método comparando-o com outros métodos do estado da arte;
- Para a construção desta CNN, apresentamos uma nova camada, intitulada camada de mapeamento, para lidar com a não-linearidade dos dados;
- Avaliamos o desempenho de diferentes arquiteturas compreendendo distintas combinações de posições da camada de mapeamento, proposta neste estudo.

II. Trabalhos Relacionados

Pesquisadores desenvolveram recentemente arquiteturas de CNN para descrever características locais em imagens obtidas do espectro visível para atuar de forma similar aos métodos tradicionais, baseados em modelagem manual (do inglês, *handcrafted*), como os métodos SIFT [15] e SURF [16]. Nesta linha, destacam-se as arquiteturas DeepCompare [17], MatchNet [18], DeepDesc [19], PN-Net [20] e TFeat [21]. Essas arquiteturas foram avaliadas no estudo [22], no qual os autores concluem que métodos tradicionais para descrição de características ainda têm desempenho igual, ou superior, comparado aos métodos baseados em CNN.

Inspirados na arquitetura PN-Net [20], os autores [23] propõem uma arquitetura, denominada Q-Net, especialmente projetada para aprender características locais em imagens multiespectrais. Os autores concluem que a solução proposta obteve uma melhora de aproximadamente 3% em relação aos métodos avaliados. Uma desvantagem dessa arquitetura é o seu tempo de processamento para treinamento, sendo mais lento que a arquitetura PN-Net devido à sua complexidade.

Ainda no contexto de imagens multiespectrais, os autores [24] introduziram uma arquitetura, denominada TS-Net, motivada pela arquitetura MatchNet [18]. Os autores apresentam uma nova função de perda para fazer com que as características locais obtidas de diferentes espectros fiquem próximas umas das outras no espaço de características. No entanto, a rede proposta é limitada em termos de diferenças não-lineares. Além disso, esta rede não pode ser empregada em um sistema convencional de correspondência de características locais, dado que esta rede produz um descritor que depende de métricas de comparação personalizadas, diferente da norma L_2, usada em descritores tradicionais como o algoritmo SIFT.

As soluções baseadas em aprendizagem profunda para descrever características locais para imagens multiespectrais ainda têm desafios a serem explorados. Grande parte dos trabalhos da literatura não aborda diretamente as não-linearidades das imagens, mostrando a importância de explorar estudos nessa linha. Também, as soluções desenvolvidas para imagens de luz visível [17], [18], [19], [20], [21] ainda não possuem uma eficácia satisfatória em imagens multiespectrais.

Na literatura, também há trabalhos recentes que apresentam métodos descritores, baseados em modelagem manual (*handcrafted*), projetados exclusivamente para descrever características em imagens multiespectrais, como MFD [7], HOMPC [25], HoDM [26], DMPCLGM [27], HOSM [8], EMCM [28]. Esses trabalhos possuem uma solução semelhante para lidar com a não-linearidade das imagens, por meio do uso de informações das bordas das imagens. Em alguns desses trabalhos, as informações são extraídas por filtros de detecção de bordas, como os filtros Gabor e Log-Gabor [7], [25], [27].

Inspirado nos trabalhos aqui apresentados, propomos um método baseado em aprendizado de máquina que lida com dados de imagens adquiridas em diferentes bandas de espectros. Para lidar com a não-linearidade das imagens multiespectrais, apresentamos uma nova camada de aprendizagem profunda que incorpora o uso de filtros Log-Gabor para a extração das informações que são similares entre um par de imagens obtidas de diferentes espectros.

III. Método Proposto

O objetivo principal, ilustrado na Fig. 2, consiste em calcular um vetor descritor $D \in \mathbb{R}^k$ de uma janela de imagem $J \in \mathbb{R}^{n \times n}$. O descritor D é o vetor resultante da última camada da CNN e possui dimensionalidade k. Este descritor pode ser empregado em sistemas convencionais de correspondência de características locais por meio da utilização da norma L_2 (distância euclidiana) como critério de similaridade. A janela de imagem J representa uma característica local extraída por um método detector de características locais.

A estrutura do método proposto é inspirada na arquitetura TFeat [21], visto que esse modelo possui propriedades desejáveis, por se tratar de uma arquitetura de poucas camadas, e por se mostrar eficiente para descrever janelas de imagens por meio de CNN. No entanto, nosso método não se trata de uma simples extensão do método TFeat, pois, nós o projetamos para trabalhar exclusivamente com imagens obtidas de distintas faixas do espectro. A Fig. 3 exibe, de forma detalhada, a arquitetura de CNN proposta neste trabalho, bem como a descrição de cada camada.

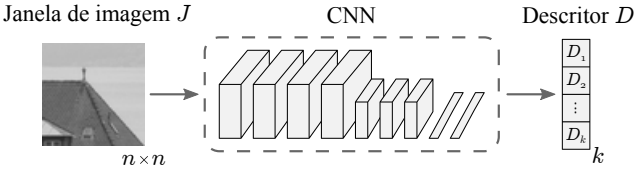

Fig. 2. Cálculo do descritor a partir de uma janela de imagem.

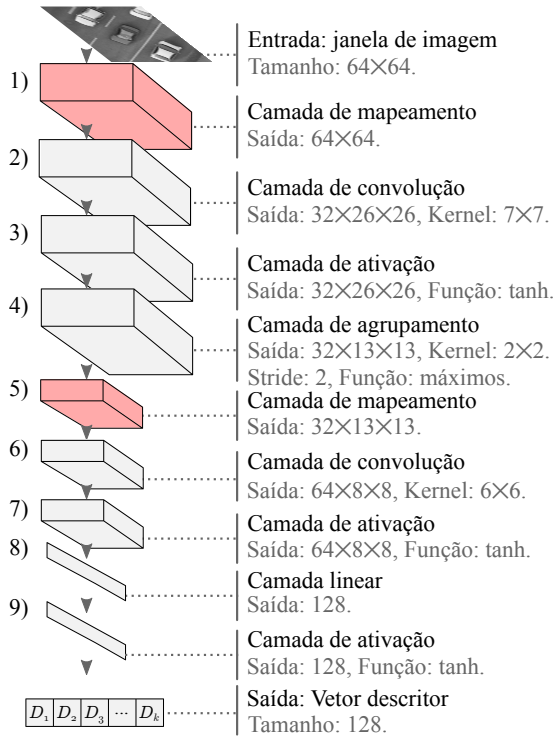

Fig. 3. Camadas da arquitetura MF-Net.

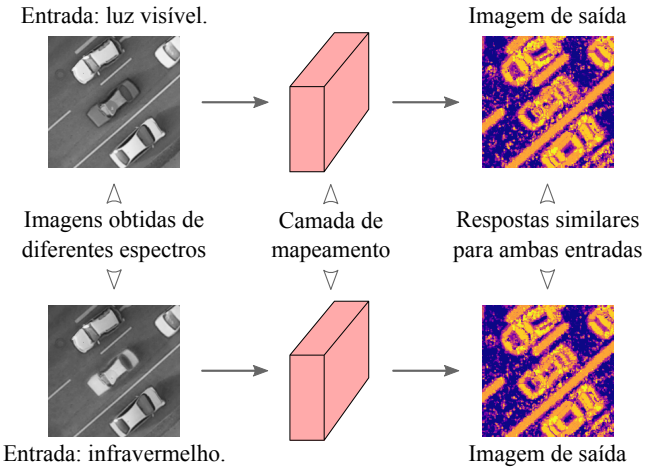

Fig. 4. Exemplo de funcionamento da camada de mapeamento na arquitetura MF-Net.

A. *Camada de Mapeamento*

Na arquitetura da nossa CNN, desenvolvemos um novo tipo de camada da CNN para lidar com a não-linearidade dos dados. Essa camada, intitulada de *Camada de Mapeamento* tem por objetivo realizar o mapeamento de entradas não-lineares, como no caso das imagens multiespectrais, de tal forma que a saída produzida tenha uma representação similar, conforme exemplificado na Fig. 4. Essa saída pode ser representada pelas bordas das imagens de entrada e, apesar de haver uma perda das informações da imagem original, as bordas tendem a permanecer similares em imagens de diferentes espectros. A camada de mapeamento incorpora os filtros Log-Gabor para realizar a tarefa de extração de características e lidar com o problema da não-linearidade das imagens multiespectrais. Nesse sentido, a camada tem o objetivo de produzir um Mapa de Respostas Absolutas (MRA), de forma semelhante aos métodos descritores tradicionais do estado da arte para imagens multiespectrais [25], [26].

Na camada de mapeamento, para uma imagem de entrada, é aplicado um banco de n filtros Log-Gabor, produzindo n imagens de resposta, conforme a Eq. 1:

$$R_n = |\text{IFFT}(\text{FFT}(I) * F_n)|, \tag{1}$$

em que "||" representa o módulo (valor absoluto) e R_n representa a resposta absoluta do filtro F_n para a imagem de entrada I. FFT representa a Transformada Rápida de Fourier e IFFT representa a sua função inversa.

Após computar todas as imagens de resposta R_n, é realizada a combinação dessas respostas e calculado o MRA. Nesse cálculo, o valor do pixel $M(x, y)$ é igual a j se o pixel $R_j(x, y)$ for o valor máximo entre todos os R_n, em que $j = 1, 2, \ldots, n$. Dessa forma, a matriz $M(x, y)$ será composta apenas pelos índices em que o pixel $R_n(x, y)$ for o maior entre todas as respostas.

A função Log-Gabor [29] no domínio da frequência utilizando coordenadas polares é dada pela Eq. 2:

$$G(r, \theta) = \exp\left(-\frac{\log\left(r/f_0\right)^2}{2\sigma_r^2}\right) . \exp\left(-\frac{(\theta - m)^2}{2}\right), \tag{2}$$

em que r e θ representam o raio e o ângulo do filtro em coordenadas polares, m consiste no ângulo de orientação do filtro, σ_r representa o desvio padrão gaussiano do filtro e f_0 representa a frequência central do filtro, dada pela Eq. 3:

$$f_0 = \lambda k^n, \tag{3}$$

em que λ consiste no comprimento de onda da menor escala do filtro, k representa uma constante multiplicativa de escala entre filtros sucessivos e n consiste na escala do filtro.

Para os filtros Log-Gabor foram definidos os valores das constantes como sugeridos no trabalho de [29] por apresentarem melhores resultados na extração de informações de texturas ao utilizar os filtros Log-Gabor em descrição de imagens. Os valores definidos foram: $\lambda = 3{,}0$, $\sigma_r = 0{,}65$ e $k = 2{,}0$ (ver equações 2 e 3 para detalhes de cada constante). Para produzir o banco de filtros, foram utilizados os mesmos valores como sugerido no método descritor MFD [7], na medida em que os autores demonstraram resultados efetivos em imagens multiespectrais. Sendo assim, foram utilizados 10 filtros, com 5 filtros de orientação (valores de m igualmente divididos de 0 até 180 graus) e 2 filtros de escala ($n = 0, 1$ da Eq. 3).

Embora os parâmetros da camada de mapeamento sejam pré-definidos (como os parâmetros da função Log-Gabor e a quantidade de filtros), o fato dessa camada estar em uma arquitetura CNN permite o aprendizado de todos os parâmetros por uma simples modificação na implementação. Assim, seria possível aprender os parâmetros da função Log-Gabor para produzir os filtros Log-Gabor de forma adaptativa. No entanto, queríamos produzir o menor modelo possível com uma menor

quantidade de parâmetros e também usar parâmetros que já foram validados na literatura, como os parâmetros definidos em [7] e [29], no contexto de extração de características para imagens multiespectrais.

B. Arquitetura MF-Net

A arquitetura da rede é inspirada na arquitetura TFeat. Para as camadas já existentes do modelo TFeat foram utilizados os mesmos parâmetros. Para a construção da arquitetura MF-Net, foram avaliadas diferentes arquiteturas, compreendendo diferentes combinações da camada de mapeamento. A camada de mapeamento foi inserida sempre antes de uma camada de convolução. Nesse sentido, foram avaliadas três arquiteturas, quais sejam: (i) arquitetura sem camadas de mapeamento (mesma arquitetura TFeat), (ii) com apenas a camada de mapeamento 1 e (iii) arquitetura com as camadas de mapeamento 1 e 5, conforme ilustrada na Fig. 5.

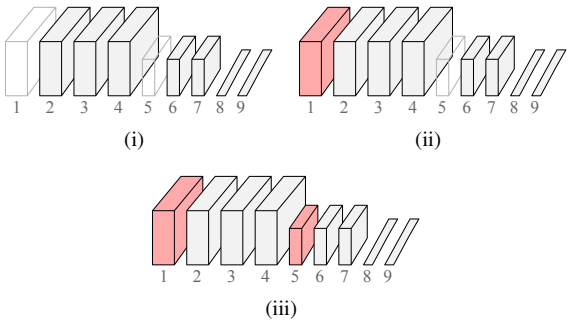

Fig. 5. Diferentes arquiteturas avaliadas: (i) sem camadas de mapeamento, mesma arquitetura TFeat, (ii) com apenas a camada de mapeamento 1 e (iii) com as camadas de mapeamento 1 e 5.

IV. Experimentos e Resultados

A. Bases de Dados

Para demonstrar a eficácia e as limitações de nosso método, realizamos experimentos em duas bases de dados distintas, amplamente utilizadas na literatura. As bases de dados estão disponíveis gratuitamente para garantir a reprodutibilidade dos resultados.

A primeira base de dados, intitulada *Potsdam*[30]. , é composta por 38 pares de imagens do espectro visível e do espectro NIR (infravermelho próximo) da vista aérea da cidade de Potsdam. Todas as imagens possuem o tamanho de 6000×6000 pixels e estão retificadas e alinhadas espacialmente. A segunda base de dados, denominada *RGB-NIR*[31], consiste em um conjunto de 477 pares de imagens do espectro visível e do espectro NIR. Todas as imagens possuem o tamanho de 1024×768 pixels e também estão retificadas e alinhadas espacialmente.

Para as duas bases de dados, extraímos janelas de imagem em torno das características locais detectadas por meio do algoritmo detector FAST [32], como sugerido em [33] e [34]. Usamos janelas de imagem de tamanho 64×64 pixels, como sugerido em [23] dado que indicaram os melhores resultados ao se utilizar abordagens de aprendizagem profunda em imagens multiespectrais.

B. Treinamento da Rede

Para o treinamento da rede, foi utilizado o mesmo protocolo definido no trabalho de [21]. Os parâmetros da rede foram otimizados por meio do algoritmo Gradiente Descendente Estocástico. O treinamento foi realizado em lotes de 128 amostras com uma taxa de aprendizagem de 0,1.

Para o processo de treinamento, foi utilizada uma estrutura tripla conforme ilustrado na Fig. 6 (mesma estrutura utilizada no treinamento da rede TFeat). Para cada par de imagens, são detectados os pontos-chave por meio do algoritmo FAST e então recortadas janelas de tamanho 64×64. Estas janelas são utilizadas no treinamento na forma $\{J_a, J_p, J_n\}$ que representa uma janela âncora J_a, uma janela positiva J_p e uma janela negativa J_n. A janela positiva consiste em uma região correspondente à janela âncora, já a janela negativa é constituída de uma região não-correspondente à mesma janela âncora. Nos experimentos, as janelas J_a são janelas do espectro visível, J_p são janelas do espectro infravermelho e J_n são janelas tomadas aleatoriamente do espectro visível ou infravermelho, como indicado no trabalho de [23], por ser a maneira mais apropriada.

A função de perda utilizada consiste na mesma função empregada pelo método TFeat, conforme a Eq. 4, definida no trabalho de [21]:

$$L(\delta_p, \delta_n) = \max(0, \mu + \delta_p - \delta_n), \tag{4}$$

em que μ representa uma constante de margem, δ_p corresponde à norma L_2 entre os descritores referentes ao par de janelas J_a e J_p, $\delta_p = \|D(J_a) - D(J_p)\|_2$ e δ_n corresponde à norma L_2 entre os descritores referentes ao par de janelas J_a e J_n, $\delta_n = \|D(J_a) - D(J_n)\|_2$.

Todas as janelas são embaralhadas de forma aleatória no início de cada época (em um total de 100 épocas), e as intensidades dos pixels de cada janela foram normalizadas (média 0 e desvio padrão 1). Assim como em [23], as janelas foram divididas em dois grupos, em que 80% foram utilizadas

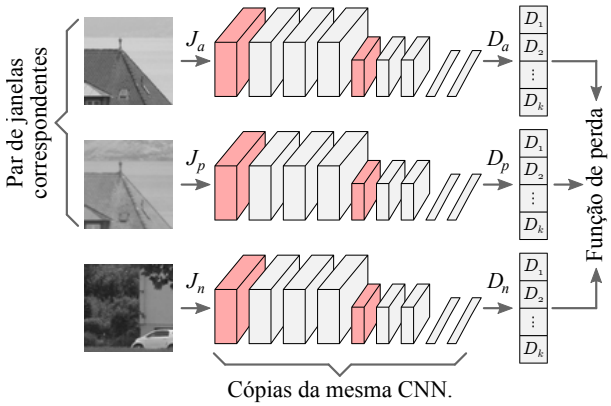

Fig. 6. Estrutura utilizada para o treinamento da arquitetura MF-Net.

para o treinamento e 20% das janelas foram separadas para teste da rede.

Cada rede foi treinada 10 vezes para suprimir os efeitos de randomização na inicialização. Nenhum ajuste de parâmetros foi utilizado, neste caso, foram usados os mesmos parâmetros fornecidos pelos autores. Os valores das sementes dos geradores de números aleatórios foram fixados com a finalidade de garantir a reprodutibilidade dos resultados, desta forma os mesmos resultados são obtidos caso os experimentos fossem executados novamente.

Assim como nos trabalhos de [21] e [20], este trabalho não utilizou nenhum tipo de técnica para aumentar a quantidade de dados, nem adicionou cópias modificadas de dados já existentes (do inglês, *data augmentation*). O objetivo dessa decisão foi tornar o processo de treinamento mais eficiente e demonstrar que as melhorias no resultado provém do método proposto e não do maior número de janelas utilizadas para treinamento.

C. Metodologia de Avaliação

Como estratégia de correspondência de características locais, utilizamos a razão da distância do vizinho mais próximo (NNDR, do inglês *Nearest Neighbor Distance Ratio*), proposta por Lowe [15]. Nesse caso, um descritor de referência D_A é correspondido a um descritor de teste D_B se a seguinte expressão for aplicável:

$$\frac{\|D_A - D_B\|_2}{\|D_A - D_C\|_2} < t, \tag{5}$$

em que t é o limiar NNDR, D_B é o primeiro e D_C é o segundo vizinho mais próximo do descritor D_A. Nesse caso, valores de t próximos de 1 retornam mais correspondências, porém, com mais correspondências incorretas e, consequentemente, menor precisão.

Para a avaliação dos métodos, foram utilizadas as métricas de precisão, revocação e medida F1. A metodologia para avaliação utilizada aqui é inspirada nos trabalhos de [33] e [35], comumente empregada na literatura para avaliação de métodos descritores de características locais.

Para cada par de imagens, é realizada a detecção e a descrição de características locais. Em seguida, é realizada a correspondência de cada descritor sobre esse par. Esta correspondência é realizada por diversas vezes, variando o limiar t (limiar NNDR, referente à Eq. 5) de 0,8 até 1,0 em intervalos de 0,05. Para cada valor do limiar, são calculados valores de precisão e revocação e então produzidas as curvas de precisão por revocação.

D. Análise Estatística

Para verificar a significância dos resultados obtidos durante a comparação dos algoritmos avaliados, nos experimentos deste trabalho é utilizado o teste t de Student unilateral superior. Como neste trabalho se deseja comparar algoritmos, utilizou-se esse teste de forma pareada, comparando as diferenças dos valores médios entre cada algoritmo avaliado. O teste estatístico nesse caso foi realizado comparando-se um determinado algoritmo com os demais algoritmos avaliados (um contra todos, dois a dois) [36].

Para todos os testes estatísticos, foi utilizado um nível de significância de $\alpha = 0{,}05$, e como se trata de múltiplas comparações, foi necessário corrigir esse valor para evitar a probabilidade de se obter uma conclusão falsa em uma série de testes de hipótese. A correção deste nível de significância foi realizada pelo método de Bonferroni, por se tratar de uma abordagem simples e conservadora [36].

E. Avaliação de Diferentes Arquiteturas

Conforme explicado na subseção III-B, foram avaliadas diferentes arquiteturas, compreendendo diferentes combinações da camada de mapeamento. Nesse sentido, foram avaliadas três arquiteturas: (i) arquitetura sem camadas de mapeamento (mesma arquitetura TFeat), (ii) com apenas a camada de mapeamento 1 e (iii) arquitetura com as camadas de mapeamento 1 e 5, conforme ilustrada na Fig. 5.

Este experimento foi realizado para a base de dados Potsdam, e a Fig. 7 exibe os valores de precisão e revocação obtidos pelas diferentes arquiteturas avaliadas nessa base de dados. Por meio deste experimento, observa-se que as arquiteturas (iii) e (ii), que possuem camadas de mapeamento, obtiveram maiores eficácias tanto na precisão quanto na revocação, comparadas à arquitetura (i). A arquitetura (iii), que possui duas camadas de mapeamento, obteve um desempenho inferior à arquitetura (ii), que possui apenas uma camada de mapeamento na entrada.

A Fig. 8 apresenta os resultados para a medida F1 dos experimentos realizados na base de dados Potsdam para diferentes limiares NNDR (ver Eq. 5). Para estes valores, foi verificada a significância dos resultados obtidos, por meio da análise estatística (explicada na subseção IV-D), e o resultado da análise indicou que, existem evidências estatísticas de que a arquitetura (ii) é superior na média das medidas F1, para cada limiar NNDR.

Em relação à construção da arquitetura MF-Net, o resultado da Fig. 7 mostra que apenas a inserção da camada de mapeamento no início da arquitetura (camada 1 da Fig. 5) é suficiente para lidar com a questão da não-linearidade dos dados de entrada. Ao adicionar mais uma camada de mapeamento no meio da arquitetura (camada 5 da Fig. 5), torna o método menos eficaz. A hipótese seria de que essa última camada realizaria uma espécie de filtro nos dados, de tal forma que o descritor se torna menos distintivo e reduz a eficácia da rede.

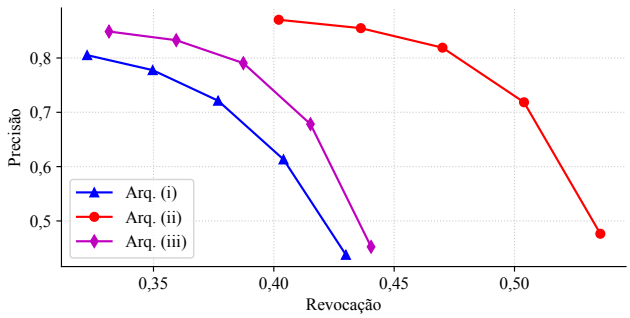

Fig. 7. Resultados de precisão e revocação para diferentes arquiteturas.

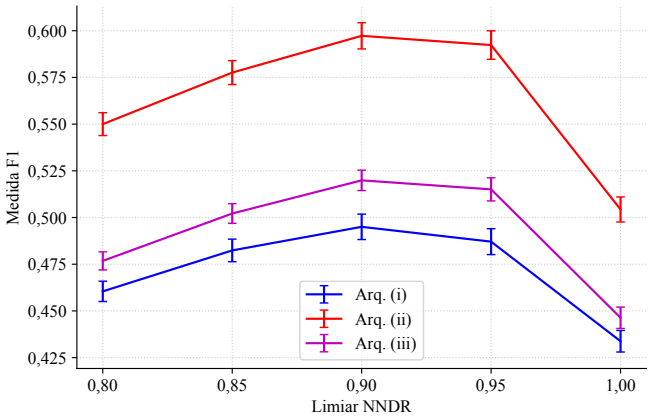

Fig. 8. Valores médios e desvio padrão para a medida F1 na base de dados Potsdam.

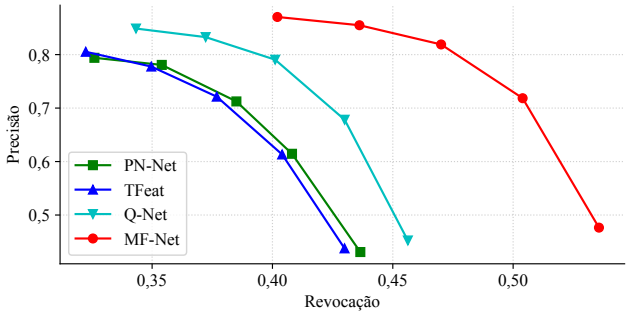

Fig. 9. Resultados de precisão e revocação para a base de dados Potsdam.

Conforme os experimentos mostrados nesta seção, recomenda-se a implementação da arquitetura MF-Net apenas com uma camada de mapeamento (camada 1 da Fig. 5), sendo esta arquitetura selecionada para os próximos experimentos.

F. Avaliação da Rede MF-Net

A arquitetura MF-Net e os descritores utilizados como referências (do inglês, *baselines*) foram avaliados em cada base de dados de forma separada, com o intuito de se obter uma melhor compreensão dos resultados dos métodos ao utilizar diferentes escopos de aplicação.

A Fig. 9 exibe os valores de precisão e revocação obtidos pelos métodos avaliados na base de dados Potsdam. Por meio deste experimento, observa-se que as arquiteturas TFeat e PN-Net obtiveram eficácia menor tanto na precisão quanto na revocação. A arquitetura MF-Net obteve maiores valores de precisão e revocação em relação aos demais métodos, incluindo a arquitetura Q-Net, que também é projetada para imagens multiespectrais.

A Fig. 10 apresenta os resultados para a medida F1 dos experimentos realizados na base de dados Potsdam para diferentes limiares NNDR. Para estes valores, foi verificada a significância dos resultados obtidos, por meio da análise estatística, e o resultado obtido indicou que há evidências estatísticas de que os valores das medidas F1 da arquitetura MF-Net é numericamente superior comparada às demais médias.

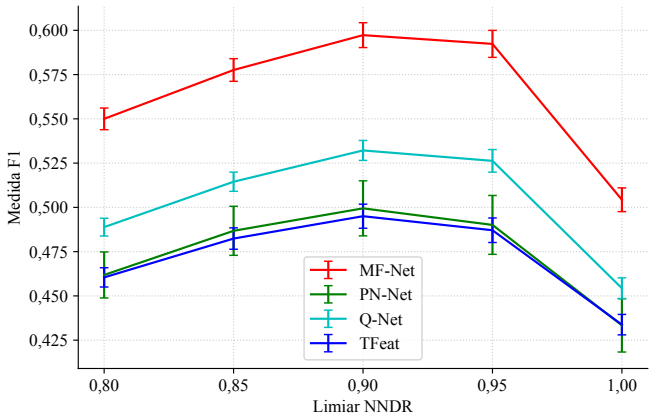

Fig. 10. Valores médios e desvio padrão para a medida F1 na base de dados Potsdam.

A Fig. 11 exibe os valores de precisão e revocação obtidos pelos métodos avaliados na base de dados RGB-NIR. Por meio deste experimento, observa-se que as arquiteturas TFeat e PN-Net obtiveram eficácia menor tanto na precisão quanto na revocação. A arquitetura MF-Net obteve maiores valores de precisão e revocação em relação aos demais métodos.

A Fig. 12 apresenta os resultados para a medida F1 dos experimentos realizados na base de dados RGB-NIR para diferentes limiares NNDR. Para estes valores, foi verificada a significância dos resultados obtidos, e o resultado indicou que há evidências estatísticas de que a média das medidas F1 da arquitetura MF-Net é numericamente superior comparada às demais médias.

Apesar das discrepâncias nos resultados entre as bases de dados, não é possível supor que os métodos avaliados tenham uma aplicação mais apropriada para um determinado cenário, pois, poderia haver outras circunstâncias que afetam os valores de precisão e revocação, incluindo diversidades de resolução e presença de ruído nas imagens (que não foram avaliadas neste trabalho). No entanto, os experimentos mostram que a eficácia do MF-Net é superior aos demais métodos comparados.

V. Limitações

O método proposto neste trabalho apresenta algumas limitações que devem ser destacadas. Primeiramente, as imagens de resposta aos filtros Log-Gabor podem, em alguns casos,

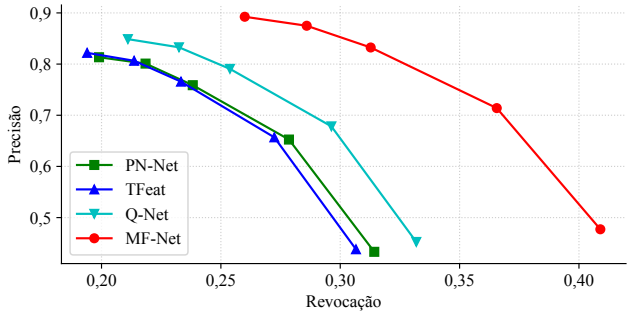

Fig. 11. Resultados de precisão e revocação para a base de dados RGB-NIR.

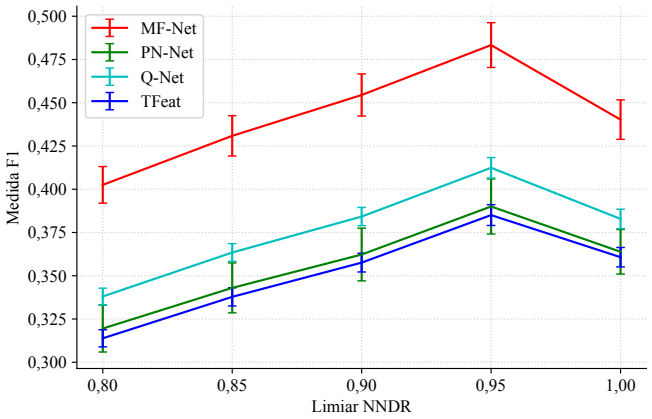

Fig. 12. Valores médios e desvio padrão para a medida F1 na base de dados RGB-NIR.

ter resultados diferentes entre imagens obtidas de espectros diferentes na camada de mapeamento. Como exemplo, tanto no espectro infravermelho quanto no espectro visível podem ocorrer casos em que as bordas das imagens são ambíguas devido às diferenças na reflexão da luz em diferentes faixas do espectro eletromagnético. Esse fenômeno, portanto, pode influenciar na extração de informações de bordas entre imagens de diferentes espectros.

Também, assim como os métodos PN-Net, TFeat e Q-Net, o método proposto, por sua construção, não trata diferentes escalas de características locais, conforme feito em alguns métodos projetados para o espectro visível, como os algoritmos SIFT e SURF. No entanto, cabe ainda avaliar o comportamento do método nessas situações.

VI. Conclusão

Neste trabalho foi apresentado um novo método descritor de características locais em imagens multiespectrais para ser aplicado em problemas de correspondência de características. A construção do método proposto, intitulado MF-Net, foi motivada pelo recente sucesso de abordagens que utilizam aprendizagem profunda.

As soluções propostas na literatura para a descrição de características locais, baseadas em aprendizagem profunda, ainda possuem limitações, dado que grande parte dos trabalhos não tratam a questão da não linearidade das imagens multiespectrais. Sendo assim, foi proposta uma nova camada de aprendizagem profunda, intitulada camada de mapeamento, para lidar com a não-linearidade dos dados. Essa camada foi inspirada na estrutura do método descritor MFD, que utiliza filtros Log-Gabor para realizar a tarefa de extração de características e lidar com as especificidades das imagens multiespectrais. Portanto, a arquitetura proposta baseia-se na extração de informações de imagens no domínio do espaço e da frequência, empregando os filtros mencionados acima.

Para a construção da arquitetura do método proposto, foi realizado um experimento para verificar o comportamento da camada de mapeamento em diferentes posições, e selecionar a arquitetura que resultasse um maior valor de precisão e revocação. Conclui-se nesse experimento que a inserção da camada de mapeamento no início da arquitetura é suficiente para se obter um melhor desempenho de descrição.

O método MF-Net foi comparado ao método do estado da arte Q-Net, também baseado em aprendizagem profunda e projetado para imagens multiespectrais. Também foram utilizados para comparação os métodos PN-Net e TFeat. Foi verificado, que o método MF-Net possui uma eficácia superior ao método Q-Net. Foi constatado também que o método descritor MF-Net supera os métodos PN-Net e TFeat, em termos de precisão e revocação para realizar a descrição de características locais em imagens multiespectrais.

Nós acreditamos que a camada de mapeamento idealizada e construída neste trabalho possa ser útil em abordagens de aprendizagem profunda que trabalham com outros tipos de dados não-lineares, visto que é relativamente simples inserir e adaptar essa camada em outras arquiteturas existentes.

A. *Trabalhos Futuros*

Com base nos resultados obtidos neste trabalho, podem ser realizados os seguintes estudos futuros:

- Comparar o método MF-Net com abordagens tradicionais (*handcrafted*) para descrição de pontos-chave em imagens multiespectrais;
- Avaliar o método MF-Net sob prováveis transformações que possam ocorrer em casos reais, como ruídos, mudanças de escala e rotação;
- Analisar e avaliar nosso método em imagens obtidas de outras bandas do espectro eletromagnético;
- Modificar a camada de mapeamento proposta para aprender os parâmetros da função Log-Gabor, de forma adaptativa;
- Avaliar a camada de mapeamento proposta com outros tipos de dados não-lineares, como mapas de profundidade.

Agradecimentos

Os autores agradecem ao CEFET-MG.

References

[1] H. Le, C. Smailis, L. Shi, and I. Kakadiaris, "EDGE20: A cross spectral evaluation dataset for multiple surveillance problems," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2020, DOI: 10.1109/wacv45572.2020.9093573.

[2] Y. Quan, X. Zhong, W. Feng, G. Dauphin, L. Gao, and M. Xing, "A novel feature extension method for the forest disaster monitoring using multispectral data," *Remote Sensing*, vol. 12, no. 14, p. 2261, jul 2020, DOI: 10.3390/rs12142261.

[3] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1025–1037, may 2020, DOI: 10.1109/tpami.2019.2961900.

[4] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2007, DOI: 10.1561/0600000017.

[5] C. Fan, H. Jin, F. Wang, G. Zhang, and Y. Li, "Combining and matching keypoints and lines on multispectral images," *Infrared Physics & Technology*, vol. 96, pp. 316–324, jan 2019, DOI: 10.1016/j.infrared.2018.12.004.

[6] C. Leng, H. Zhang, B. Li, G. Cai, Z. Pei, and L. He, "Local feature descriptor for image matching: A survey," *IEEE Access*, vol. 7, pp. 6424–6434, 2019, DOI: 10.1109/access.2018.2888856.

[7] C. F. G. Nunes and F. L. C. Padua, "A local feature descriptor based on log-gabor filters for keypoint matching in multispectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1850–1854, oct 2017, DOI: 10.1109/lgrs.2017.2738632.

[8] T. Ma, J. Ma, and K. Yu, "A local feature descriptor based on oriented structure maps with guided filtering for multispectral remote sensing image matching," *Remote Sensing*, vol. 11, no. 8, p. 951, apr 2019, DOI: 10.3390/rs11080951.
[9] M. A. Dede, E. Aptoula, and Y. Genc, "Deep network ensembles for aerial scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 732–735, may 2019, DOI: 10.1109/lgrs.2018.2880136.
[10] L. Ren, J. Lu, J. Feng, and J. Zhou, "Uniform and variational deep learning for RGB-d object recognition and person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4970–4983, oct 2019, DOI: 10.1109/tip.2019.2915655.
[11] L. Ngo, J. Cha, and J.-H. Han, "Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images," *IEEE Transactions on Image Processing*, vol. 29, pp. 303–312, 2020, DOI: 10.1109/tip.2019.2931461.
[12] Q. Qi, Q. Huo, J. Wang, H. Sun, Y. Cao, and J. Liao, "Personalized sketch-based image retrieval by convolutional neural network and deep transfer learning," *IEEE Access*, vol. 7, pp. 16 537–16 549, 2019, DOI: 10.1109/access.2019.2894351.
[13] K. Kuppala, S. Banda, and T. R. Barige, "An overview of deep learning methods for image registration with focus on feature-based approaches," *International Journal of Image and Data Fusion*, vol. 11, no. 2, pp. 113–135, jan 2020, DOI: 10.1080/19479832.2019.1707720.
[14] Y. Dong, W. Jiao, T. Long, L. Liu, G. He, C. Gong, and Y. Guo, "Local deep descriptor for remote sensing image feature matching," *Remote Sensing*, vol. 11, no. 4, p. 430, feb 2019, DOI: 10.3390/rs11040430.
[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, nov 2004, DOI: 10.1023/b:visi.0000029664.99615.94.
[16] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, 2006, pp. 404–417, DOI: 10.1007/11744023_32.
[17] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015, DOI: 10.1109/cvpr.2015.7299064.
[18] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015, DOI: 10.1109/cvpr.2015.7298948.
[19] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015, DOI: 10.1109/iccv.2015.22.
[20] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "Pn-net: Conjoined triple deep network for learning local image descriptors," *arXiv preprint arXiv:1601.05030*, 2016.
[21] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Association, 2016, DOI: 10.5244/c.30.119.
[22] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017, DOI: 10.1109/cvpr.2017.736.
[23] C. Aguilera, A. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," *Sensors*, vol. 17, no. 4, p. 873, apr 2017, DOI: 10.3390/s17040873.
[24] S. En, A. Lechervy, and F. Jurie, "TS-NET: Combining modality specific and common features for multimodal patch matching," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, oct 2018, DOI: 10.1109/icip.2018.8451804.
[25] Z. Fu, Q. Qin, B. Luo, H. Sun, and C. Wu, "HOMPC: A local feature descriptor based on the combination of magnitude and phase congruency information for multi-sensor remote sensing images," *Remote Sensing*, vol. 10, no. 8, p. 1234, aug 2018, DOI: 10.3390/rs10081234.
[28] B. Fang, K. Yu, J. Ma, and P. An, "EMCM: A novel binary edge-feature-based maximum clique framework for multispectral image matching,"
[26] Z. Fu, Q. Qin, B. Luo, C. Wu, and H. Sun, "A local feature descriptor based on combination of structure and texture information for multispectral image matching," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2018, DOI: 10.1109/lgrs.2018.2867635.
[27] X. Liu, J.-B. Li, and J.-S. Pan, "Feature point matching based on distinct wavelength phase congruency and log-gabor filters in infrared and visible images," *Sensors*, vol. 19, no. 19, p. 4244, sep 2019, DOI: 10.3390/s19194244.
Remote Sensing, vol. 11, no. 24, p. 3026, dec 2019, DOI: 10.3390/rs11243026.
[29] E. Walia and V. Verma, "Boosting local texture descriptors with log-gabor filters response for improved image retrieval," *International Journal of Multimedia Information Retrieval*, vol. 5, no. 3, pp. 173–184, apr 2016, DOI: 10.1007/s13735-016-0099-2.
[30] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label potsdam.html.
[31] https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html.
[32] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, 2006, pp. 430–443, DOI: 10.1007/11744023_34.
[33] T. Mouats, N. Aouf, D. Nam, and S. Vidas, "Performance evaluation of feature detectors and descriptors beyond the visible," *Journal of Intelligent & Robotic Systems*, vol. 92, no. 1, pp. 33–63, feb 2018, DOI: 10.1007/s10846-017-0762-8.
[34] S. Saleem, A. Bais, R. Sablatnig, A. Ahmad, and N. Naseer, "Feature points for multisensor images," *Computers & Electrical Engineering*, vol. 62, pp. 511–523, aug 2017, DOI: 10.1016/j.compeleceng.2017.04.032.
[35] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, oct 2005, DOI: 10.1109/tpami.2005.188.
[36] F. Campelo and E. F. Wanner, "Sample size calculations for the experimental comparison of multiple algorithms on multiple problem instances," *Journal of Heuristics*, vol. 26, no. 6, pp. 851–883, aug 2020, DOI: 10.1007/s10732-020-09454-w.

Cristiano F. G. Nunes received the BS degree in computation engineering from the Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brazil, in 2014. He received the MS degree in Mathematical and Computational Modeling from the Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brazil, in 2017. He's a doctoral degree student and has been a system analyst at the same institution. His research interests include computer vision, pattern classification and content-based image and video retrieval.

Flávio L. C. Pádua received the BS degree in electrical engineering and the MS and PhD degrees in computer science from the Universidade Federal de Minas Gerais (UFMG), Brazil, in 1999, 2002, and 2005, respectively. He has been an associate professor of computer engineering at the Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) since 2005. His research interests include computer vision, pattern classification and content-based image, and video retrieval.