Research, Society and Development, v. 11, n. 14, e383111436631, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i14.36631

Registro automático, preciso e robusto de imagens com RANSAC (Random Sample Consensus) adaptado para o descritor SIFT (Scale Invariant Feature Transform)

Automatic, accurate and robust image registration with adapted RANSAC (Random Sample Consensus) for SIFT (Scale Invariant Feature Transform) descriptor

Registro automático, preciso y robusto de imágenes con RANSAC (Random Sample Consensus) adaptado al descriptor SIFT (Scale Invariant Feature Transform)

Recebido: 18/10/2022 | Revisado: 24/10/2022 | Aceitado: 25/10/2022 | Publicado: 30/10/2022

Rubens Antonio Leite Benevides
ORCID: https://orcid.org/0000-0003-2605-451X
Universidade Federal do Paraná, Brasil
E-mail: rubensleite11@gmail.com
Kalima Pitombeira
ORCID: https://orcid.org/0000-0003-1970-0894
Universidade Federal do Paraná, Brasil
E-mail: kalimapitombeira@hotmail.com
Jorge Centeno
ORCID: https://orcid.org/0000-0002-2669-7147
Universidade Federal do Paraná, Brasil
E-mail: jorgcenteno@gmail.com
Paulo Rodrigo Simões
ORCID: https://orcid.org/0000-0002-6789-4497
Universidade Federal do Paraná, Brasil
E-mail: prsimoes@gmail.com

Resumo

O registro de imagens é um problema comum na visão computacional com diversas aplicações que consiste em encontrar a correta transformação entre pares de imagens que se sobrepõem. Neste trabalho objetiva-se apresentar um modelo automático e preciso para registro de imagens utilizando o descritor SIFT e o método de estimação RANSAC adaptado. O registro ocorre através da estimativa da homografia entre os pares de imagens, que utilizam as correspondências pontuais dadas pelo descritor SIFT. As correspondências são classificadas usando um limiar de erro estimado dinamicamente. A análise considera a dispersão do resíduo em vários limites de erro e adota aquele que minimiza a dispersão e a magnitude do erro. O modelo é testado com 8 pares heterogêneos de imagens divididos em dois grupos: 4 pares obtidos com uma câmera profissional e 4 pares obtidos com uma câmera comum. Devido à alta qualidade das imagens do primeiro grupo, poucas iterações do modelo são necessárias para a estimativa da homografia correta. No segundo grupo, o modelo mostrou que é capaz de construir mosaicos entre pares de imagens com sobreposição inferior a 20%, encontrando correspondências exatas entre pares de imagens independentemente do método de aquisição. Além disso, foi capaz de lidar com até 65% de corrupção entre as correspondências, com um tempo total de execução de alguns segundos.

Palavras-chave: Registro de imagens; SIFT; Homografia; RANSAC.

Abstract

Image registration is a common problem in computer vision with several applications which consists of finding the correct transformation between pairs of overlapping images. This work aims to present an automatic and accurate model for image registration using the SIFT descriptor and the adapted RANSAC estimation method. The registration occurs through the estimation of the homography between the pairs of images, which use the point correspondences given by the SIFT descriptor. Matches are classified using a dynamically estimated error threshold. The analysis considers the dispersion of the residual over various error thresholds and adopts the one that minimizes the dispersion and the magnitude of the error. The model is tested with 8 heterogeneous pairs of images divided into two groups: 4 pairs obtained with a professional camera and 4 pairs obtained with a common camera. Due to the high quality of the images in the first group, few iterations of the model are necessary for a good estimate of the correct homography. In the second group, the model showed that it is capable of building mosaics between pairs of images with an overlap of less than 20%, finding exact correspondences between pairs of images regardless of the acquisition method.

Furthermore, it was able to handle up to 65% corruption between matches, with a total execution time of a few seconds.
Keywords: Image registration; SIFT; Homography; RANSAC.

Resumen
El registro de imágenes es un problema común en visión artificial con varias aplicaciones que consiste en encontrar la transformación correcta entre pares de imágenes superpuestas. Este trabajo tiene como objetivo presentar un modelo automático y preciso para el registro de imágenes utilizando el descriptor SIFT y el método de estimación RANSAC adaptado. El registro ocurre a través de la estimación de la homografía entre los pares de imágenes, que utilizan las correspondencias puntuales dadas por el descriptor SIFT. Las coincidencias se ordenan utilizando un umbral de error estimado dinámicamente. El análisis considera la dispersión del residuo en varios límites de error y adopta el que minimiza la dispersión y la magnitud del error. El modelo se prueba con 8 pares heterogéneos de imágenes divididas en dos grupos: 4 pares obtenidos con una cámara profesional y 4 pares obtenidos con una cámara común. Debido a la alta calidad de las imágenes del primer grupo, son necesarias pocas iteraciones del modelo para estimar la homografía correcta. En el segundo grupo, el modelo demostró que es capaz de construir mosaicos entre pares de imágenes con una superposición de menos del 20%, encontrando coincidencias exactas entre pares de imágenes independientemente del método de adquisición. Además, pudo manejar hasta un 65% de corrupción entre partidos, con un tiempo de ejecución total de unos pocos segundos.
Palabras clave: Registro de imagen; SIFT; Homografia; RANSAC.

1. Introdução

A automação do registro de imagens para construção de mosaicos é um tema bastante estudado por pesquisadores das áreas de Fotogrametria, Visão Computacional, Processamento Digital de Imagens (PDI) e Computação Gráfica. Isto se deve ao fato das variadas aplicações que podem ser realizadas com mosaicos, como a confecção de ortomosaico para fins de mapeamento, construção de imagens panorâmicas, determinação da posição relativa de novas imagens, reconstrução 3D, entre outros (Paul & Pati, 2021).

Um mosaico de imagens consiste em um conjunto de fotos com sobreposição entre si, as quais são transformadas geometricamente para um mesmo sistema de referência de coordenadas no espaço-imagem. Com a transformação ideal aplicada, as fotos são deformadas, possibilitando que a região de sobreposição seja mesclada entre duas ou mais imagens. Assim, uma única imagem pode ser obtida contemplando toda a área visível no conjunto de fotos (Kumar et al., 1995); (Mills & Dudek, 2009). Existem vários métodos para obter mosaicos de imagens, atualmente é comum o uso de modelos treinados por *deep learning* (Fu, et al. 2020) mas o paradigma constante dos passos é resumido em (Wang & Watada, 2015):

1) Correspondência de feições: encontrar pontos homólogos nas imagens originais;
2) Correspondência de imagens: estimar a transformação geométricas entre as imagens;
3) Alinhamento do panorama: aplicar a rotação 3D relativa entre as posições das câmeras no instante de tomada das fotos;
4) Equalização das imagens: balancear a iluminação e saturação entre as imagens;
5) União das imagens: mesclagem da imagem, corrigindo o efeito de paralaxe causado pelo movimento da câmera.

Para que seja realizado o processo de mosaicagem é necessário o conhecimento de medidas aproximadas da posição e orientação das câmeras em um referencial global ou de correspondência de feições na imagem (Mills & Dudek, 2009). No processo que utiliza observações da orientação exterior das câmeras a principal vantagem encontra-se na ausência de problemas para detecção e correspondência de feições, uma vez que a transformação entre as imagens é obtida com o Método dos Mínimos Quadrados (MMQ) aplicado a posição e atitude relativa entre as câmeras. Por outro lado, esse método requer uma boa aproximação inicial da orientação exterior das imagens, além de exigir que as rotações entre os planos das imagens sejam pequenas, para garantir que a solução irá convergir corretamente e a estimação dos parâmetros não recaia em um mínimo local (Yang & Guo, 2008).

No caso do método por correspondência de feições, a transformação entre as fotos pode ser calculada quando existe

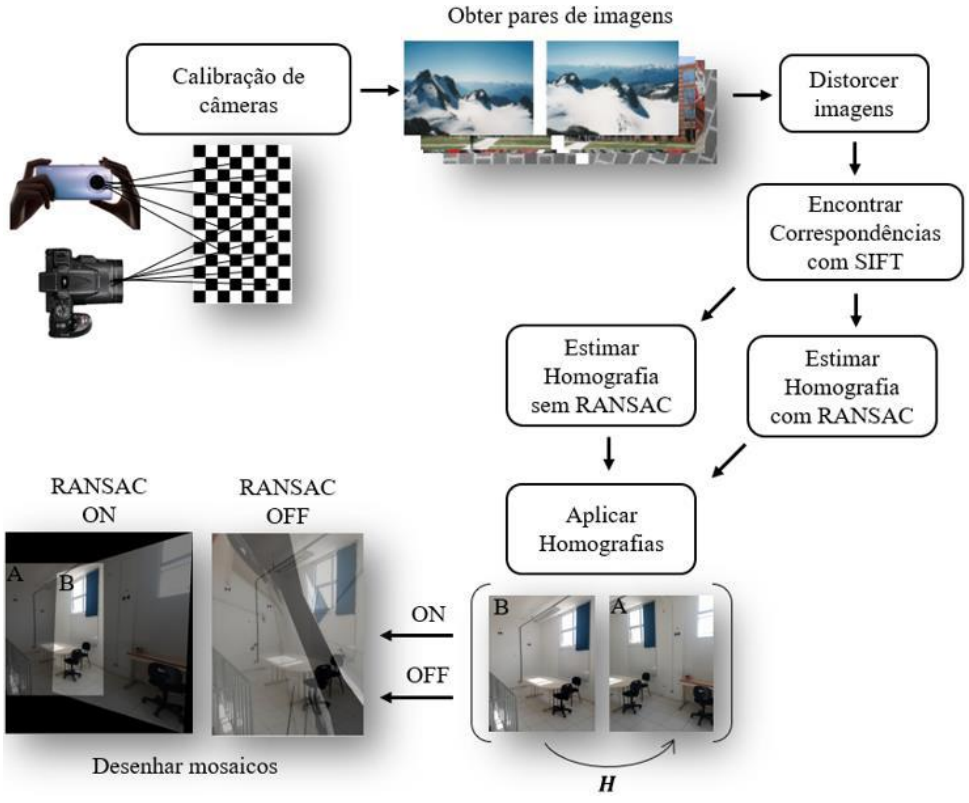
sobreposição entre as imagens e a superfície fotografada é heterogênea, com pontos fotoidentificáveis. Nesse processo, como são utilizadas observações no espaço-imagem, não é necessário ter conhecimento inicial da posição e orientação das câmeras. Contudo, a estimação dos parâmetros de transformação é sensível a existência de erros grosseiros nas correspondências de pontos homólogos, por isso, em um processo automatizado de medidas de pontos nas fotos, devido à ruídos e oclusões, é comum a existência de falsas correspondências entre os pontos (Yang & Guo, 2008).

Tendo em vista essas considerações sobre os dois métodos, o referente estudo tem o objetivo de analisar o desempenho do processo de mosaicagem entre duas imagens utilizando a transformação de Homografia, cujos parâmetros serão estimados robustamente por correspondência de pontos homólogos. A detecção e correspondência de feições será realizada por meio do descritor SIFT (Scale Invariant Feature Transform), projetado por (Lowe, 2004), e para evitar falsas correspondências será aplicado o modelo de estimação robusta RANSAC (RANdom Sample Consensus), elaborado por (Fischler & Bolles, 1981) com adaptações específicas para que se autodetermine. Desse modo, uma análise comparativa entre a elaboração do mosaico com e sem o modelo RANSAC modificado será conduzida, verificando a influência que as correspondências espúrias causam no processo de estimação da matriz de Homografia e como isto se traduz na elaboração do mosaico de imagens.

2. Metodologia

Nesta seção apresenta-se o detalhamento das etapas do trabalho. Adota-se a notação matemática padrão: escalares são representados por letras minúsculas, vetores são representados por letras minúsculas em negrito e matrizes por letras maiúsculas em negrito. Multiplicações entre elementos não comutativos são feitas da esquerda para a direita. O esquema simplificado dos passos seguidos neste trabalho é apresentado na Figura 1.

Figura 1 - Metodologia simplificada do trabalho.



Fonte: Autores.

Calibração geométrica das câmeras

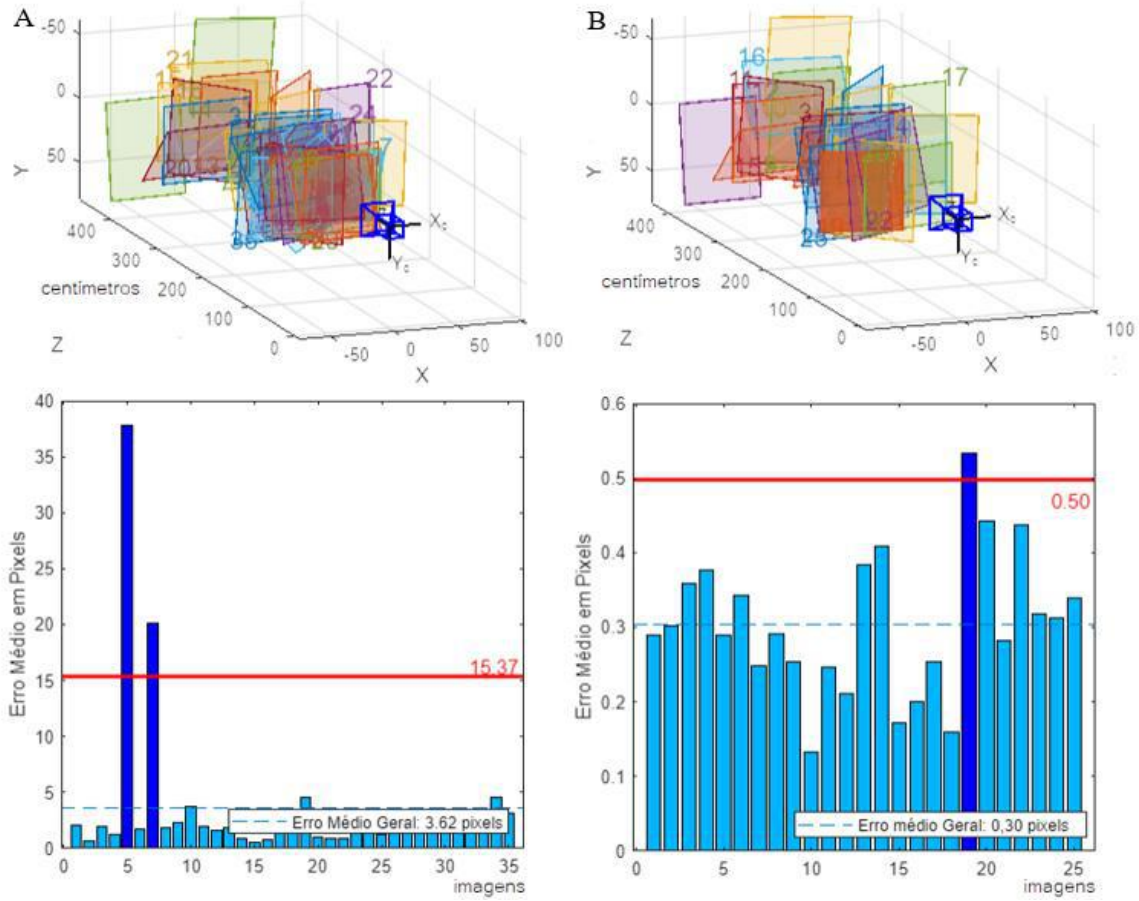
A calibração geométrica de uma câmera consiste em estimar os parâmetros envolvidos nas distorções das imagens obtidas com a câmera, estas distorções estão relacionadas aos defeitos de fabricação das partes da câmera, que não são perfeitas. Fisicamente, as distorções acontecem por causa de diversos fatores, os principais envolvidos são: (1) os desvios de curvatura nas lentes do conjunto ótico, que não são perfeitamente esféricas; (2) o desalinhamento do eixo ótico com o centro geométrico do sensor fotoelétrico; e (3) a não perpendicularidade entre o plano de formação da imagem (CCD ou CMOS) e o eixo ótico do conjunto de lentes, circunstância responsável por induzir uma escala alterada nas direções principais da imagem.

Cada efeito acima pode ser modelado matematicamente por um conjunto de parâmetros, conhecendo-os é possível corrigir quase totalmente as distorções das imagens. Diversas aplicações de beneficiam desta etapa, como a determinação da posição da câmera no espaço, a recuperação da trajetória do sensor – odometria monocular – e a obtenção de medidas do mundo real através das imagens com frameworks de *Structure From Motion* e *Multi View Stereo* (Schonberger & Frahm, 2016); (Pons et al., 2007).

Neste trabalho utilizou-se duas câmeras distintas, uma câmera profissional Nikon-P600, usada nos levantamentos de cavernas e abrigos arqueológicos (espeleologia) do laboratório de Geodésia Aplicada à Engenharia (GEENG) da Universidade Federal do Paraná (UFPR); e uma câmera simples, do smartphone LG-K51s, sem nenhuma característica especial.

A calibração das câmeras se deu de forma semiautomática por meio do Camera Calibrator Toolbox, aplicativo disponível no programa Matlab® 2022 e gratuitamente em (Bouguet, 2008). Neste propósito, 40 fotos do padrão xadrez de calibração do laboratório foram obtidas por ambas as câmeras, na maior resolução disponível, em diversas posições e orientações. O recomendado pelo software MATLAB são 20 cenas, assim, 25 cenas foram utilizadas, pois as 15 de pior qualidade foram excluídas. Com as fotos do padrão foi possível obter a matriz intrínseca $\boldsymbol{K}$ com os parâmetros de calibração de ambas as câmeras. Na Figura 2, em A), apresentam-se as imagens de calibração e os erros associados, em B) tem-se a mesma visualização após a filtragem.

Figura 2 - Erro de reprojeção no processo de calibração. A) Pré-filtragem. B) Pós-filtragem.



Fonte: Autores.

Figura 3 - Resultados da calibração das câmeras.

LG-K51s	
Dist. focal	1,928 mm
Pto. principal	O_x = 2091,1 pixels O_y = 1556,7 pixels
Tamanho da imagem	3120 × 4160 pixels
Distorção radial	0,1477 − 0,3647
Distorção tangencial	− 0,0014 0,0007
Erro de reprojeção méd.	0. 9680 pixels
$\boldsymbol{K} = \begin{bmatrix} f_x & 0 & 0 \\ skew & f_y & 0 \\ c_x & c_y & 1 \end{bmatrix}$	$\begin{bmatrix} 3236{,}2 & 0 & 0 \\ 0{,}27 & 3237{,}4 & 0 \\ 2091{,}1 & 1556{,}7 & 1 \end{bmatrix}$

Nikon-P600	
Dist. focal	4,0 mm
Pto. principal	O_x = 2287,5 pixels O_y = 1701,1 pixels
Tamanho da imagem	3456 × 4608 pixels
Distorção radial	0,0033 0,0071
Distorção tangencial	− 0,0032 − 0,0022
Erro de reprojeção méd.	0. 3033 pixels
$\boldsymbol{K} = \begin{bmatrix} f_x & 0 & 0 \\ skew & f_y & 0 \\ c_x & c_y & 1 \end{bmatrix}$	$\begin{bmatrix} 3304{,}7 & 0 & 0 \\ -0{,}51 & 3306{,}0 & 0 \\ 2287{,}5 & 1701{,}1 & 1 \end{bmatrix}$

Fonte: Autores (2022).

Conjunto de dados

O conjunto de dados utilizado neste estudo compreende 8 pares de fotografias, quatro obtidos com a câmera do smartphone LG-K51s e quatro adquiridas com a câmera Nikon-P600. A Figura 4 A) apresenta os pares de fotografias com o

LG-K51s. A Figura 4 B) apresenta as imagens do sítio arqueológico Pedra Pintada, no município Barão de Cocais. Os pares foram obtidos com a câmera Nikon-P600. Todas as imagens estão corrigidas das distorções citadas na seção anterior utilizando a matriz intrínseca K correspondente, além disso, foram amostradas para uma resolução máxima de 2000 pixels na maior dimensão da imagem utilizando o interpolador bicúbico. A menor dimensão da imagem é apropriadamente reduzida por um fator que preserve a proporção sem distorções.

Figura 4 - Pares de imagens para teste do modelo de registro. A) Pares de imagens obtidos com o LG-K51s. B) Pares de imagens obtidos com a com a câmera Nikon-P600.

Par_1 Par_2 Par_3 Par_4

Par_5 Par_6 Par_7 Par_8

Fonte: Autores.

Na literatura de Visão Computacional, a referência global costuma ser o da primeira tomada. Como tem-se apenas pares de fotografias, uma delas é a origem global, a imagem de referência (*target image*) e a outra é a imagem que será transformada, a imagem de pesquisa (*source image*).

Detecção de correspondências entre imagens por SIFT

Entre os algoritmos que permitem detectar feições homólogas entre imagens, o SIFT (Scale Invariant Feature

Transform), desenvolvido por (Lowe, 2004), é amplamente utilizado para aplicações de PDI e Visão Computacional. Isso se deve ao fato do descritor ser consideravelmente robusto a variações na escala, na rotação das imagens, e na mudança de iluminação e de perspectiva. As etapas do SIFT podem ser subdividas na seguinte sequência: detectar bordas e quinas, localizar pontos notáveis (keypoints) em quinas e bordas, atribuir orientação e desenvolver um descritor dos pontos notáveis.

Para o primeiro processo presente no SIFT são detectadas bordas estáveis que sejam invariantes à mudança de escala na imagem. Isto é feito por meio de uma função contínua que percorre várias escalas possíveis, denominada de função espaço escala. O espaço escala é definido então pela função $L(x, y, \sigma)$, a qual é obtida a partir de uma imagem $\boldsymbol{I}(x, y)$ e da convolução gaussiana $\boldsymbol{G}(x, y, \sigma)$. Sendo * a operação de convolução, tem-se a equação 1 como segue (Lowe, 2004):

$$L(x, y, \sigma) = G(x, y, \sigma) * \boldsymbol{I}(x, y) \quad (1)$$

onde:

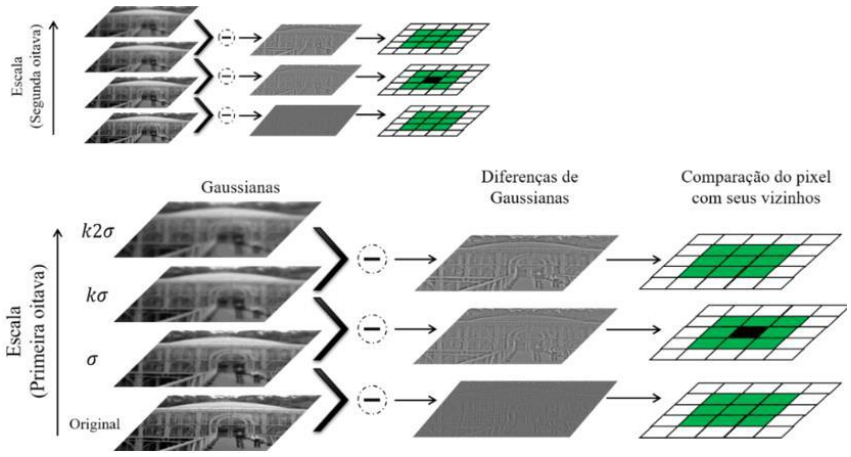
$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

Assim, para a detecção de pontos no espaço escala, (Lowe, 2004) propõe detectar extremos locais (máximos e mínimos) na função Diferença de Gaussiana $D(x, y, \sigma)$. Como a função gaussiana é uma solução da equação diferencial de difusão, o resultado de sua convolução é atenuar as altas frequências da imagem por meio do desvio-padrão σ adotado no filtro gaussiano. Desta forma, são obtidas imagens sequencialmente convolucionadas por meio de uma constante k que multiplica o valor de σ, onde $k = 2^{1/s}$, e s é o nº inteiro de intervalos entre as Diferenças de Gaussianas, isto é, o número de Diferenças de Gaussianas localizada entre outras duas (Berveglieri, 2014). A eq. 3 apresenta a função Diferença de Gaussianas:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (3)$$

O conjunto que inclui a imagem original, as imagens sucessivamente convolucionadas, e as Diferenças de Gaussianas, formam a primeira oitava na escala original da imagem. Em seguida, a imagem é subamostrada pela metade para ser utilizada na próxima oitava, e assim sucessivamente, de modo que o conjunto das oitavas formam o espaço-escala. Posteriormente, para que os extremos locais possam ser detectados nas Diferenças de Gaussianas, cada pixel é comparado com seus oito vizinhos na imagem (considerando uma matriz 3x3) e nove vizinhos nas Diferenças de Gaussiana anterior e posterior. Assim, são selecionados apenas os pontos com nível de cinza máximo ou mínimo quando comparados com seus 26 pixels vizinhos (Lowe, 2004). A Figura 5 apresenta estes processos contidos na primeira etapa do SIFT, considera-se $k = 2$ e $s = 1$, de forma que apenas uma Diferença de Gaussiana (central) de cada oitava será analisada em relação a sua vizinhança.

Figura 5 - Detecção de quinas e bordas no SIFT.



Fonte: Autores.

Os pontos encontrados na detecção de quinas tornam-se, então, candidatos à pontos notáveis (keypoints). Considerando esses candidatos, a próxima etapa do SIFT consiste em uma análise da vizinhança para localização, escala e raio dos componentes principais. Neste processo, a localização do ponto candidato é recalculada com maior exatidão, possibilitando que sejam eliminados pontos mal localizados ao longo de uma borda e pontos com baixo contraste, que são sensíveis a ruídos.

Esse método, desenvolvido por (Brown & Lowe, 2002), consiste em ajustar uma função quadrática tridimensional para interpolar a localização do máximo da função a nível subpixel. Assim, em cada ponto analisado é realizada uma expansão de Taylor de grau dois na função de Diferença de Gaussianas, com a origem transladada para o ponto candidato. Tendo-se $D(x,y)$ e suas duas primeiras derivadas no ponto de interesse, e $x = (x,y,\sigma)^T$ o deslocamento em relação a esse ponto, a função é dada por:

$$D(x) = D + \frac{\partial D^T}{\partial x}x + \frac{1}{2}x^T\frac{\partial^2 D}{\partial x^2}x \tag{4}$$

a nova localização do ponto $\hat{x}$ é obtida por meio da derivada da função em relação a x, para isto iguala-se a derivada a zero soluciona-se o modelo por mínimos quadrados, conforme a equação 5:

$$\hat{x} = \frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x} \tag{5}$$

a função da Diferença de Gaussiana é interpolada, então, para a nova localização $\hat{x}$, de modo que a função $D(\hat{x})$ é dada por:

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\hat{x} \tag{6}$$

Considera-se que os valores de nível de cinza estão normalizados no intervalo $[0,1]$, onde é estabelecido o limiar $|D(\hat{x})| < 0{,}03$ para eliminação de pontos com baixo contraste. A Diferença de Gaussianas é uma função que retorna muitos valores de pontos de máximo/mínimo ao longo de arestas, mesmo que não estejam bem localizados. Portanto, para garantir a estabilidade dos candidatos a serem selecionados como pontos notáveis, é necessário realizar uma análise de componentes principais. No caso de um ponto mal localizado em bordas, a função Diferença de Gaussiana vai apresentar grande variação de gradiente na direção perpendicular à borda, mas pequena variação ao longo desta. Logo, para definir apenas pontos de quina, podem ser determinados os autovalores da matriz Hessiana (eq. 7), definida na localização e escala deste ponto, visto que os

autovalores de $\boldsymbol{H}$ são proporcionais às componentes principais da Diferença de Gaussianas na vizinhança do ponto analisado.

$$\boldsymbol{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{xx} \end{bmatrix} \quad (7)$$

Com os autovalores, sendo λ_1 e λ_2 os autovalores de maior e menor magnitude, respectivamente, a soma desses autovalores corresponde ao traço da matriz Hessiana, e seu produto corresponde ao determinante da matriz. Assim, para que não seja necessário calcular os autovalores, o traço e o determinante da matriz H são relacionados conforme a equação 8, onde r é o raio entre λ_1e λ_2:

$$\frac{Tr(\boldsymbol{H})^2}{Det(\boldsymbol{H})} = \frac{(\lambda_1 + \lambda_2)^2}{\lambda_1 + \lambda_2} = \frac{(r+1)^2}{r} \quad (8)$$

logo, quando os autovalores são iguais ($r = 1$) o resultado de $(r+1)^2/r$ é mínimo. Portanto, é possível analisar a estabilidade do ponto levando em conta apenas a razão entre os autovalores, (Lowe, 2004) estabelece um valor empírico de $r \leq 10$ para considerar o ponto analisado como notável (keypoint), isto é, os demais pontos são eliminados por serem mal localizados ao longo de bordas.

Posteriormente, com base nas propriedades locais da imagem, a terceira etapa do SIFT realiza a atribuição de orientação para cada ponto notável. Desse modo o descritor pode ser orientado de maneira relativa, considerando as propriedades locais dos pontos, o que possibilita ao SIFT a invariância às rotações das imagens. Neste sentido, a imagem com filtro gaussiano $L(x, y, \sigma)$ que mais se aproxima da escala do ponto notável é utilizada. Calcula-se, então, a magnitude $m(x, y)$ e orientação $\theta(x, y)$ do gradiente em cada pixel em relação a sua vizinhança, como mostra a eq. 9 e 10:

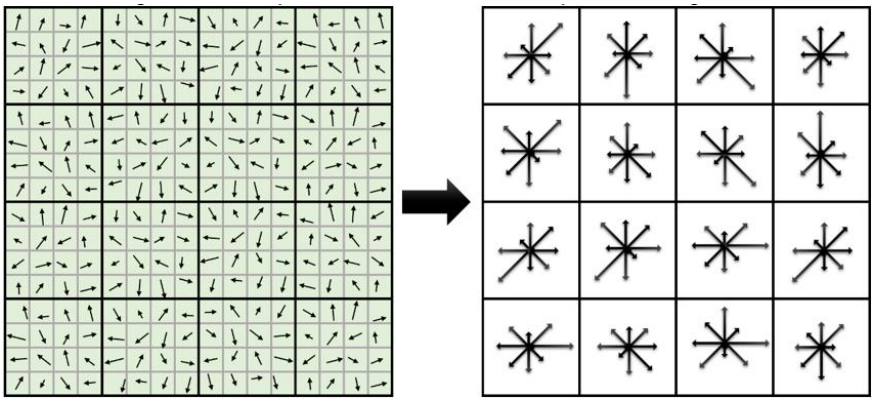
$$m(x,y) = \sqrt{\big(\boldsymbol{L}(x+1,y) - \boldsymbol{L}(x-1,y)\big)^2 + (\boldsymbol{L}(x,y+1) - \boldsymbol{L}(x,y-1))^2} \quad (9)$$

$$\theta(x,y) = tan^{-1}\left(\frac{\boldsymbol{L}(x,y+1) - \boldsymbol{L}(x,y-1)}{\boldsymbol{L}(x+1,y) - \boldsymbol{L}(x-1,y)}\right) \quad (10)$$

Com o intuito de analisar a magnitude e orientação dos gradientes determinados, calcula-se um histograma para a vizinhança do ponto analisado. Idem (2004), propõe que os 360° de rotação sejam divididos em 36 intervalos de 10°. Cada valor de gradiente adicionado no histograma é ponderado pela magnitude do gradiente e por uma janela gaussiana circular com desvio 1,5 vezes maior que o σ do ponto analisado. Assim, o intervalo com maior valor no histograma corresponde a direção dominante dos gradientes na vizinhança do ponto, sendo atribuída essa orientação ao ponto notável (keypoint). Caso exista outro intervalo no histograma com 80% do valor do maior, são criados dois pontos com mesma localização e atribuídas orientações distintas.

A quarta etapa do SIFT consiste na determinação de um descritor local do ponto que seja significativamente distinto e o mais invariante possível às mudanças de iluminação ou de perspectiva. Para isso, seleciona-se a imagem filtrada $L(x, y)$ com desvio padrão σ correspondente à oitava do ponto, e a vizinhança local é rotacionada de acordo com a orientação do ponto determinada anteriormente. Em seguida, é definida uma janela $k \times k$ de pixels em cada região $n \times n$ em torno do ponto (Belo, 2006). Para cada região será calculada um novo histograma local para 8 direções (45° de intervalo). (Lowe, 2004) sugere empiricamente que os valores para a melhor estabilidade dos pontos notáveis são $k = n = 4$, nesse caso os histogramas de regiões 4x4 subdividas em 4x4 pixels, considerando 8 direções, resultam em 128 elementos (4x4x8) que descrevem a vizinhança do ponto, como mostra a Figura 6. Esses elementos são armazenados em um vetor que, por fim, é normalizado para reduzir os efeitos de iluminação.

Figura 6 - Informação salva no descritor SIFT: 8 direções de 16 histogramas.



Fonte: Autores.

Na última etapa, para encontrar pontos correspondentes, o melhor candidato a ponto homólogo se refere ao vizinho mais próximo dentre todos os pontos notáveis da outra imagem, isto é, a correspondência será realizada com o ponto cujo vetor referente ao descritor SIFT apresentar menor distância euclidiana em relação ao ponto da imagem de referência.

Homografia

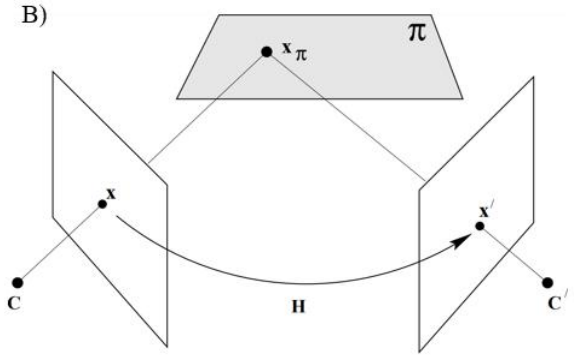
A Homografia consiste em uma transformação projetiva entre dois planos imersos no espaço 3D. Por operar no espaço 3D os pontos em planos têm as coordenadas representadas por três dimensões, garantindo vantagens sobre a transformação no espaço euclidiano $\mathbb{R}^2$ ao trabalhar no espaço projetivo $\mathbb{P}^2$ (Hartley & Zisserman, 2003). Assim, rotações ${}^A_B\boldsymbol{R}$ de um sistema B para A e translações ${}^A_B\boldsymbol{t}$ de um sistema B para A, podem ser codificadas em uma única matriz ${}^A_B\boldsymbol{H}_{3\times3}$ chamada de homografia. A Figura 7 A) apresenta uma tabela com as características de operação da transformação, em B) tem-se a interpretação geométrica, onde duas câmeras observam o mesmo plano $\boldsymbol{\pi}$. C e C' são os centros de convergência dos raios de cada câmera.

Figura 7 - A) Diferenças de operação entre os espaços $\mathbb{R}^2$ e $\mathbb{P}^2$. B) Vista do plano π imerso em $\mathbb{R}^3$ contendo um ponto x que é projetado no plano das duas câmeras $\mathbf{C}$ e $\mathbf{C}'$. A homografia fornece a transformação entre os planos das câmeras, isto é, dadas as coordenadas do ponto x em uma imagem, permite saber as coordenadas desse ponto na outra imagem.

A)

Geometria	Transformação 2D	Tipo	Coordenadas do pto.
Euclidiana $\mathbb{R}^2$	$\boldsymbol{p^A} = s({}^A_B\boldsymbol{R})\boldsymbol{p^B} + {}^A_B\boldsymbol{t}$	não linear	(x, y)
Projetiva $\mathbb{P}^2$	$\boldsymbol{p^A} = ({}^A_B\boldsymbol{H})\boldsymbol{p^B}$	linear	(x, y, z)

B)

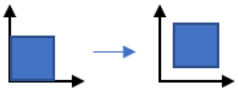
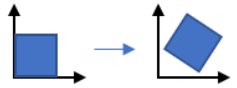
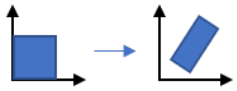
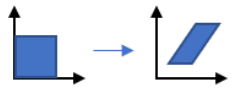
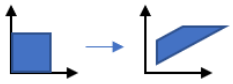


Fonte: adaptado de (Hartley & Zisserman, 2003).

A Figura 7 B) retrata o fato de que as câmeras projetam pontos do espaço em planos, portanto, a projeção homográfica 2D é exatamente o tipo de transformação necessária para relacionar duas câmeras que observam o mesmo plano.

A homografia é uma transformação linear geral que preserva apenas as linhas retas entre os pares de imagens. A Tabela 1 abaixo contextualiza a homografia em relação a outras transformações mais restritivas do espaço $\mathbb{P}^2$.

Tabela 1 - Transformações no espaço projetivo $\mathbb{P}^2$.

Transformação	Matriz	Graus de lib.	Preserva	Visualização
Translação	$\begin{bmatrix} \boldsymbol{I}_{3\times3} & \boldsymbol{t} \\ \boldsymbol{0}_{1\times3} & 1 \end{bmatrix}$ $\boldsymbol{I}$ = matriz identidade	2	Orientação + todas abaixo	
Corpo rígido 2D	$\begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}_{1\times3} & 1 \end{bmatrix}$	3	Comprimento + todas abaixo	
Similaridade	$\begin{bmatrix} s\boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}_{1\times3} & 1 \end{bmatrix}$	4	Ângulos + todas abaixo	
Afim	$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix}$	6	Paralelismo, linha no infinito + todas abaixo	
Homografia	$\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$	8	linhas retas	

Fonte: Autores.

Há pelo menos duas formas de estimar uma transformação de homografia entre um par de imagens, a primeira é conhecendo-se os Parâmetros de Orientação Interiores (POI) e Exteriores (POE) de uma câmera em relação a outra, a segunda é por meio de pares de pontos correspondentes, sob a restrição de estarem no mesmo plano. Desta forma, para duas imagens que se sobrepõem, sempre existe uma transformação homográfica que as relaciona. Considere um par de imagens A e B com sobreposição. Seja $\boldsymbol{p}_A = [x_A, y_A, z_A]^T$ um ponto na imagem A em coordenadas homogêneas, e $\boldsymbol{p}_B = [x_B, y_B, z_B]^T$ este mesmo ponto na imagem B. Então $\boldsymbol{p}_A = \boldsymbol{H}\boldsymbol{p}_B$, como segue:

$$\begin{bmatrix} x_A \\ y_A \\ z_A \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_B \\ y_B \\ z_B \end{bmatrix} \tag{11}$$

Os vetores $\boldsymbol{p}_A^T$ e $\boldsymbol{H}\boldsymbol{p}_B^T$ intersectam o mesmo ponto no plano, mas não são iguais, suas magnitudes diferem por um fator de escala homogêneo $\lambda \neq 0$ (dos Santos & Rocha, 2012). Para eliminar este fator de escala faz-se o produto externo ($\times$) com $\boldsymbol{p}_A$, como segue:

$$\boldsymbol{p}_A^T \times [\lambda \boldsymbol{p}_A^T] = \boldsymbol{p}_A^T \times [\boldsymbol{H}\boldsymbol{p}_B^T] \tag{12}$$

o lado esquerdo da eq. 12 é nulo, pois os vetores têm a mesma direção, logo:

$$\boldsymbol{p}_A^T \times [\boldsymbol{H}\boldsymbol{p}_B^T] = \boldsymbol{0}^T \tag{13}$$

onde $\boldsymbol{0}^T = [0,0,0]^T$. O produto matriz-vetor na equação 11 pode ser lido como o produto interno ($\cdot$) da i-ésima linha $\boldsymbol{h}^i$ da matriz $\boldsymbol{H}$, por $\boldsymbol{p}_B^T$, como segue:

$$\boldsymbol{H}\boldsymbol{p}_B = \begin{bmatrix} h_{11}x_B + h_{12}y_B + h_{13}z_B \\ h_{21}x_B + h_{22}y_B + h_{23}z_B \\ h_{31}x_B + h_{32}y_B + h_{33}z_B \end{bmatrix} \Rightarrow \boldsymbol{H}\boldsymbol{p}_B = \begin{bmatrix} \boldsymbol{h}^1 \cdot \boldsymbol{p}_B^T \\ \boldsymbol{h}^2 \cdot \boldsymbol{p}_B^T \\ \boldsymbol{h}^3 \cdot \boldsymbol{p}_B^T \end{bmatrix} \tag{14}$$

o produto externo na equação 13 pode ser reescrito na forma matricial, para isto se escreve o vetor $\boldsymbol{p}_A^T$ como uma matriz

antissimétrica (*skew-symmetric*), como segue:

$$[\boldsymbol{p}_{A\times}^{T}] \times [\boldsymbol{H}\boldsymbol{p}_{B}^{T}] = \begin{bmatrix} 0 & -z_A & y_A \\ z_A & 0 & -x_A \\ -y_A & x_A & 0 \end{bmatrix} \times \begin{bmatrix} \boldsymbol{h}^1 \cdot \boldsymbol{p}_B^T \\ \boldsymbol{h}^2 \cdot \boldsymbol{p}_B^T \\ \boldsymbol{h}^3 \cdot \boldsymbol{p}_B^T \end{bmatrix} \quad (15)$$

como o produto interno é comutativo, é possível isolar os elementos $\boldsymbol{h}^i$ em um vetor e definir um sistema de equações homogêneas $\boldsymbol{A}_i\boldsymbol{h}^T = \boldsymbol{0}$:

$$\begin{bmatrix} \boldsymbol{0} & -z_A\boldsymbol{p}_B & y_A\boldsymbol{p}_B \\ z_A\boldsymbol{p}_B & \boldsymbol{0} & -x_A\boldsymbol{p}_B \\ -y_A\boldsymbol{p}_B & x_A\boldsymbol{p}_B & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{h}^1 \\ \boldsymbol{h}^2 \\ \boldsymbol{h}^3 \end{bmatrix} = \boldsymbol{0}_{9x1} \Rightarrow \boldsymbol{A}_i\boldsymbol{h}^T = \boldsymbol{0} \quad (16)$$

para manter a consistência os elementos nulos são expandidos para vetores nulos $\boldsymbol{0}$, logo, a matriz $\boldsymbol{A}_i$ tem 3×9 elementos, que são funções das coordenadas de um par i de pontos correspondentes ($x_B^i, y_B^i, z_B^i, x_A^i, y_A^i, z_A^i$). Onde $\boldsymbol{h}$ é o vetor coluna formado pelos 9 elementos da matriz homografia $\boldsymbol{h} = [h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{34}]$.

Embora se tenham 3 equações em $\boldsymbol{A}_i$, apenas duas delas são linearmente independentes, a terceira linha é obtida com a soma de x_A vezes a primeira linha e y_A vezes a segunda linha. Então, cada par de correspondências fornece duas equações envolvendo os parâmetros de $\boldsymbol{H}$, sendo usual omitir a última linha e escrever o sistema de equações como (Hartley & Zisserman, 2003):

$$\begin{bmatrix} \boldsymbol{0}_3 & -z_A\boldsymbol{p}_B & y_A\boldsymbol{p}_B \\ z_A\boldsymbol{p}_B & \boldsymbol{0}_3 & -x_A\boldsymbol{p}_B \end{bmatrix} \begin{bmatrix} \boldsymbol{h}^1 \\ \boldsymbol{h}^2 \\ \boldsymbol{h}^3 \end{bmatrix} = \boldsymbol{0} \quad (17)$$

então, um par de pontos correspondentes determina um sistema de duas equações a nove incógnitas, onde $\|\boldsymbol{h}\| = 1$, pois o fator homogêneo de escala da homografia é arbitrário, de forma que apenas 8 elementos da matriz precisam ser determinados. Para $n = 4$ o sistema admite apenas uma solução. Para $n > 4$, a presença de correspondências inexatas torna a solução para o sistema $\boldsymbol{A}\boldsymbol{h} = \boldsymbol{0}$ inexistente, portanto, uma solução aproximada é desejável. É claro que $\boldsymbol{h}$ poderia ser 0, mas não se deseja esta solução trivial. O problema é formalmente definido estabelecendo-se que $\|\boldsymbol{A}\boldsymbol{h}\|$ deve ser minimizado. O vetor $\boldsymbol{h}_{min}$ é obtido com a decomposição em autovalores e autovetores de $\boldsymbol{A}^T\boldsymbol{A}$. O autovetor $\boldsymbol{v}_i$ associado ao menor autovalor λ_i será a solução desejada, tal que $\|\boldsymbol{v}\| = 1$ e $\|\boldsymbol{A}\boldsymbol{v}\| = min$ (Hartley & Zisserman, 2003).

Logo, dados $n = 4$ pares de pontos correspondentes, é possível obter um sistema $\boldsymbol{A}\boldsymbol{h} = \boldsymbol{0}$ de solução exata, onde $\boldsymbol{A}$ é a matriz 8×9 formada pela concatenação de matrizes $\boldsymbol{A}_i$, com i variando de 1 até n. Caso se utilize o sistema descrito na equação 16, a matriz $\boldsymbol{A}$ assume dimensão 12×9, mas em ambos os casos seu rank será 8 e a solução será dada pelo espaço nulo de $\boldsymbol{A}$.

Random Sample Consensus (RANSAC)

O algoritmo RANSAC, proposto por (Fischler & Bolles, 1981), é uma abordagem robusta para a estimativa dos parâmetros de um modelo matemático. O algoritmo obtém sucesso na estimação mesmo com grande proporção de erros grosseiros nos dados (outliers), algo comum na busca de pares de pontos correspondentes entre imagens. Para isto, em vez de utilizar a totalidade dos dados para obter uma solução inicial, o RANSAC usa o menor conjunto de dados possível e repete esse procedimento ranqueando os modelos segundo o ajuste do modelo aos dados (Fischler & Bolles, 1981). O algoritmo é descrito por cinco etapas:

1) Dado um conjunto P de dados, seleciona-se aleatoriamente um subconjunto S que contém o número mínimo n de elementos necessários para estimar os parâmetros de um modelo;
2) Os parâmetros do modelo são estimados com esse conjunto S de n elementos;
3) A partir de uma tolerância de erro pré-estabelecida ε, verifica-se quantos pontos do conjunto P são consistentes com o modelo estabelecido. Os pontos consistentes formam um subconjunto S^* chamado consenso, que é utilizado para obter um novo vetor de parâmetros do modelo.
4) Se o número de pontos do conjunto S^* (inliers) exceder um limiar τ pré-determinado, então os parâmetros do modelo são estimados utilizando S^*.
5) Caso contrário, as etapas (a), (b) (c) e (d) são repetidas k vezes até que um conjunto consenso S^* com pelo menos τ membros seja encontrado.

Deste modo, o algoritmo RANSAC estima os parâmetros de um modelo diversas vezes, e em cada vez o modelo é ranqueado de acordo com o ajuste aos dados. Há três parâmetros não especificados no RANSAC: a tolerância de erro ε, que distingue os pontos entre inliers e outliers, o critério de parada k, que estabelece o máximo de iterações/amostragens, e o critério de parada τ, que estabelece se dado modelo é bom. A seguir, discute-se como definir apropriadamente cada um destes parâmetros. Há muitos trabalhos neste contexto e diversas estratégias de otimização podem ser consultadas em (Raguram et al., 2008) e (Chum & Matas, 2008).

Otimização do RANSAC para determinação do parâmetro ε

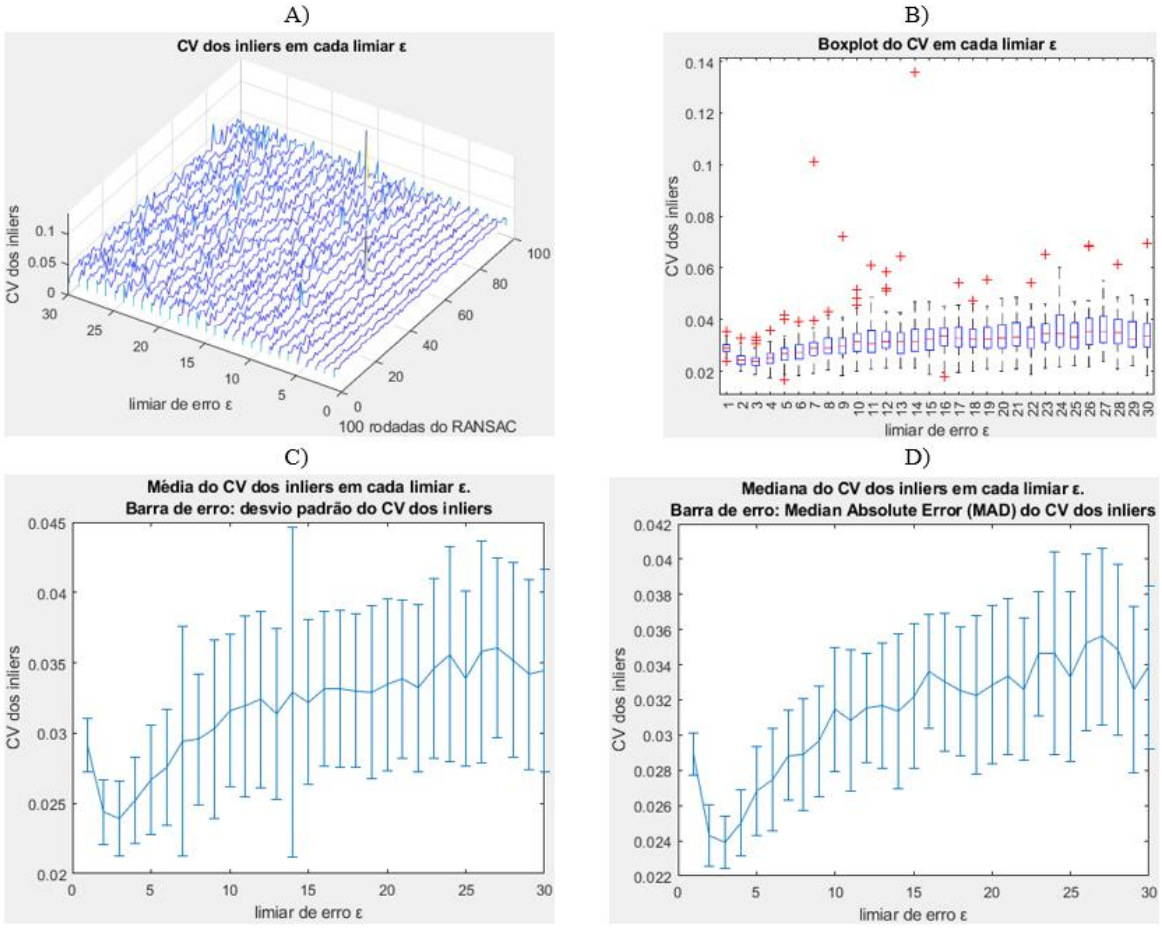
Para o parâmetro ε, que define o limiar entre inliers e outliers, não há resposta geral que possa ser adotada (Zuliani, 2009), pois ele é avaliado em função da distribuição dos dados. Uma prática comum é assumir que qualquer observação distante da média é um outlier, mas neste contexto diversas medidas de dispersão e centralidade podem ser consideradas. Caso se considere que os erros são normalmente distribuídos, a dispersão Qui-quadrado $\chi^2_{GL,\alpha}$ pode ser empregada, desde que se saiba os Graus de Liberdade (GL) do modelo e se estabeleça uma confiança desejada. O problema com este método é que os erros entre correspondências não são adequadamente modelados com a curva normal.

O RANSAC tem grande sensibilidade ao limiar que separa inliers de outliers, diferentes valores de ε determinam diferentes distribuições de erros, e ε menor não implica em menos inliers. Para contornar isto existe uma variação do RANSAC chamada Maximum Likelihood Estimator Sample Consensus (MLESAC) proposto por (Torr & Zisserman, 2000). O MLESAC modela a distribuição dos inliers com a normal e a distribuição dos outliers por um modelo uniforme, calculando um parâmetro de mistura em cada iteração.

Neste trabalho também se adota um critério com base na distribuição dos inliers segundo a tolerância de erro ε, mas o critério é definido empiricamente, com análise estatística. A instabilidade da classificação com um limiar muito restritivo ocorre porque poucos inliers são selecionados em cada iteração. Por outro lado, aumentar ε também torna a distribuição dos inliers esparsa, embora mais uniforme, pois muitos pares são selecionados. Esta dispersão segundo ε pode ser quantificada se associada a uma medida de centralidade, logo, é possível adotar um limiar ε ótimo que minimiza a dispersão dos inliers.

Para definir o limiar ε, calcula-se o Coeficiente de Variação (CV = σ/μ) do desvio padrão da média ($\sigma_\mu = \sigma/\sqrt{n}$) dos inliers. Faz-se isso ao longo de 30 limiares de erro ε, variando de 1 a 30 pixels, e executa-se o RANSAC 100 vezes em cada limiar. Fez-se isto para contornar a natureza estocástica do RANSAC. O CV tem a vantagem de ser uma medida de dispersão adimensional. Um CV próximo de zero implica em uniformidade da amostra, um CV próximo de 1 implica em grande dispersão. Toma-se o desvio padrão da média, no cálculo do CV, devido ao fato de que poucos inliers são obtidos em limiares restritivos. Por outro lado, muitos inliers são obtidos quando ε cresce. Assim, o termo $\sqrt{n}$, normaliza o valor do desvio padrão de acordo com a cardinalidade da amostra, permitindo comparar dispersões de erros obtidas em amostragens diferentes independentemente do tamanho de S^* (Bruce & Bruce, 2019). As Figuras 8 A), B), C) e D) apresentam o sumário estatístico das 100 rodadas de teste do RANSAC em 100 limiares de erros diferentes.

Figura 8 - Sumário dos testes do RANSAC para diferentes limiares de erro ε.



Fonte: autores (2022).

Na Figura 8 A) apresentam-se todos os dados, isto é, os valores do CV_{σ_μ} segundo o limiar ε em cada rodada do RANSAC. Em B) tem-se o boxplot do CV_{σ_μ} segundo o limiar de erro ε, nele é possível ver a mediana e sua dispersão por meio da amplitude interquartílica (25% e 75%). Em C) tem-se a média do CV_{σ_μ} e seu respectivo desvio padrão em cada limiar ε. Em D) tem-se a mediana do CV_{σ_μ} e o Desvio Absoluto da Mediana (Median Absolute Deviation - MAD), para cada limiar ε. O MAD é um índice robusto calculado como a mediana dos desvios absolutos da mediana (Bruce & Bruce, 2019).

Na Figura 8 percebe-se o que foi discutido antes, a dispersão dos inliers se torna instável conforme o limiar de erro se torna menos restritivo, por outro lado, também é instável no extremo esquerdo dos gráficos B), C) e D), quando $\varepsilon \to 1$. Observe que o CV_{σ_μ} mede a dispersão do erro, quanto menor for essa dispersão, mais os inliers estão concentrados em torno de zero. Nas barras de erro está a dispersão do CV_{σ_μ}, o que pode ser interpretado como a reprodutibilidade do RANSAC naquele erro. Percebe-se que os erros se tornam mais instáveis conforme ε cresce, pois a dispersão da média e a dispersão da mediana em torno do CV_{σ_μ} aumentaram.

Poderia ser dito que isto se trata de um overfitting do modelo RANSAC aos dados específicos deste trabalho, todavia a dispersão do erro dos inliers independe da acurácia do SIFT, isto é, da sensibilidade do descritor diante da variação de ε. Logo, esta etapa se trata de um fine-tunning do RANSAC para o descritor em uso, e não de overfitting do modelo. Em geral, o limiar ideal está entre 3 e 5 pixels.

Otimização do RANSAC na determinação do parâmetro k

Para determinar a quantidade máxima de iterações k, a pergunta a ser respondida pelo RANSAC é: quantas vezes k um modelo deve ser estimado, a partir de um conjunto de N dados tomados em amostras de n elementos, até que se obtenha um bom modelo? Define-se bom modelo como 99 % de confiança que os parâmetros foram estimados com dados livres de erros grosseiros (outliers). A resposta para a pergunta pode ser calculada caso a proporção de inliers/pares seja conhecida previamente. Caso contrário, escolhe-se um valor arbitrariamente grande para k.

Seja ω a razão inliers/pares, se n pares são selecionados aleatoriamente para estimar um modelo, então ω^n é a probabilidade de que todos os n pares sejam inliers. Logo, $1-\omega^n$ é a probabilidade de que pelo menos um, dos n pares, seja outlier, caso que levaria a um modelo corrompido. Esta probabilidade elevada a k tentativas, isto é: $(1-\omega^n)^k$, é a probabilidade de que após k seleções, com reposição, o algoritmo não encontre nenhum conjunto de pares inliers. Seja de ρ a probabilidade complementar desta, de que apenas conjuntos de inliers sejam selecionados, então $1-\rho$ deve ser igual a $(1-\omega^n)^k$:

$$1-\rho = (1-\omega^n)^k \tag{18}$$

ao isolar k na eq. 18 se obtém a quantidade mínima de amostragens necessárias para garantir a confiança adotada ρ:

$$k_{p,n,\omega} \geq \frac{\log(1-p)}{\log(1-\omega^n)} \tag{19}$$

por exemplo: para obter 99 % de confiança na estimação de um modelo que utiliza $n = 4$ pares, se 26% dos pares são inliers, então k deve ser maior ou igual a:

$$k_{(p=0{,}99;\, n=4;\, \omega=0{,}26)} \to k \geq 1006 \text{ iterações} \tag{20}$$

Perceba que quanto menor for a taxa ω, mais iterações o modelo exigirá para lidar com aquela proporção de outliers. Neste sentido, existem algumas estratégias de otimização para a escolha de k. A adaptação que será feita neste artigo se baseia no seguinte: a proporção de inliers ω não é conhecida previamente para um par de imagens, portanto, k é desconhecido a priori. Contudo, ω pode ser estimado grosseiramente em cada iteração do algoritmo com o conjunto consenso S^*. Caso o

algoritmo selecione uma amostra que induza alta quantidade de outliers ($\omega \cong 0$), k será automaticamente calculado com esse novo valor. Como $k \to \infty$ quando $\omega \to 0,$ o algoritmo infere que deve rodar muitas vezes para achar um bom modelo entre dados corrompidos. A cada vez que um bom conjunto consenso for encontrado, a quantidade total de iterações calculadas passa a ser menor. Em resumo, k é estimado adaptativamente em função de ω em cada iteração (Hartley & Zisserman, 2003), o que permite tempo ótimo de execução em dados com baixa quantidade de corrupção e tempo longo em dados muito corrompidos. Na prática, os dados corrompidos não podem ultrapassar 75% do total, nestes casos o RANSAC ainda funciona, mas exige uma quantidade de tempo impraticável devido ao enorme valor de k.

O último critério de parada a ser definido para o RANSAC é a contagem máxima de inliers do conjunto consenso, perceba que isto é automaticamente definido a partir da taxa ω. Seja τ a quantidade de inliers, o máximo de inliers que pode haver em S^* é $\omega\tau$, mas como ω é obtido iterativamente a partir de τ, não faz sentido calcular novamente τ em função de ω, na mesma iteração. Desta forma, a determinação automática de ω torna desnecessário estabelecer este outro critério de parada τ.

Homografia + RANSAC

Dado um conjunto de pares P de possíveis correspondências obtidas com o descritor SIFT, onde $P = \{p_A, p_B\}$, com $p_A = \left\{p_a^{(1)}, p_a^{(2)}, \dots, p_a^{(i)}\right\}$ sendo pontos da imagem A, e $p_B = \left\{p_b^{(1)}, p_b^{(2)}, \dots, p_b^{(i)}\right\}$ sendo pontos da imagem B, se objetiva estimar a transformação de homografia $\boldsymbol{H}$ com o máximo de inliers em P. Como mostrado em (Hartley & Zisserman, 2003), um dos passos fundamentais para a solução correta é que as coordenadas dos pontos sejam normalizadas, isto impede erros de mau condicionamento do sistema matricial. A normalização é feita transladando as coordenadas para o centroide dos pontos e depois escalonando-os de forma que a distância média dos pontos até a origem seja aproximadamente $2^{1/2}$.

Após a normalização aplica-se o RANSAC adaptativo na estimação do modelo de homografia. Em cada iteração são selecionados quatro pares aleatórios de P, com estes pares um sistema matricial é montado de acordo com a eq. 16 e então é solucionado por SVD. Seja $\boldsymbol{H}^{(k)}$ a matriz de homografia resultante desta amostragem k, a qualidade de $\boldsymbol{H}^{(k)}$ é calculada aplicando-a nos pontos da imagem B e contando-se o total de pontos que se mantêm dentro da tolerância de erro ε. Os pontos de B para A são transformados como segue:

$$\boldsymbol{p}_A^{(k)} = \boldsymbol{H}^{(k)}\boldsymbol{p}_B \tag{24}$$

as coordenadas x e y em $p_A^{(k)}$ são então normalizadas pelo fator de escala z da homografia:

$$\begin{aligned} \boldsymbol{p}_{A_x}^{(k)} &= \boldsymbol{p}_{A_x}^{(k)} / \boldsymbol{p}_{A_z}^{(k)} \\ \boldsymbol{p}_{A_y}^{(k)} &= \boldsymbol{p}_{A_y}^{(k)} / \boldsymbol{p}_{A_z}^{(k)} \end{aligned} \tag{25}$$

em seguida, se verificam as distâncias entre os pontos transformados e os originais:

$$\boldsymbol{v} = \sqrt{\left(\boldsymbol{p}_{A_x}^{(k)} - \boldsymbol{p}_{A_x}\right)^2 + \left(\boldsymbol{p}_{A_y}^{(k)} - \boldsymbol{p}_{A_y}\right)^2} \tag{26}$$

adota-se $\varepsilon = 5$ pixels. Pontos em B transformados para o sistema em A que estejam fora desse limiar são considerados outliers. k é inicializado arbitrariamente alto para lidar com uma quantidade arbitrária de corrupção. Em cada iteração soluciona-se a homografia, calcula-se a taxa ω e com a eq. 19 calcula-se uma nova quantidade de iterações. Caso a nova quantidade de iterações k seja menor que a restante, atualiza-se k para este valor. Após a última iteração do RANSAC todos os pontos

encontrados como inliers são utilizados para solucionar uma nova Homografia por SVD.

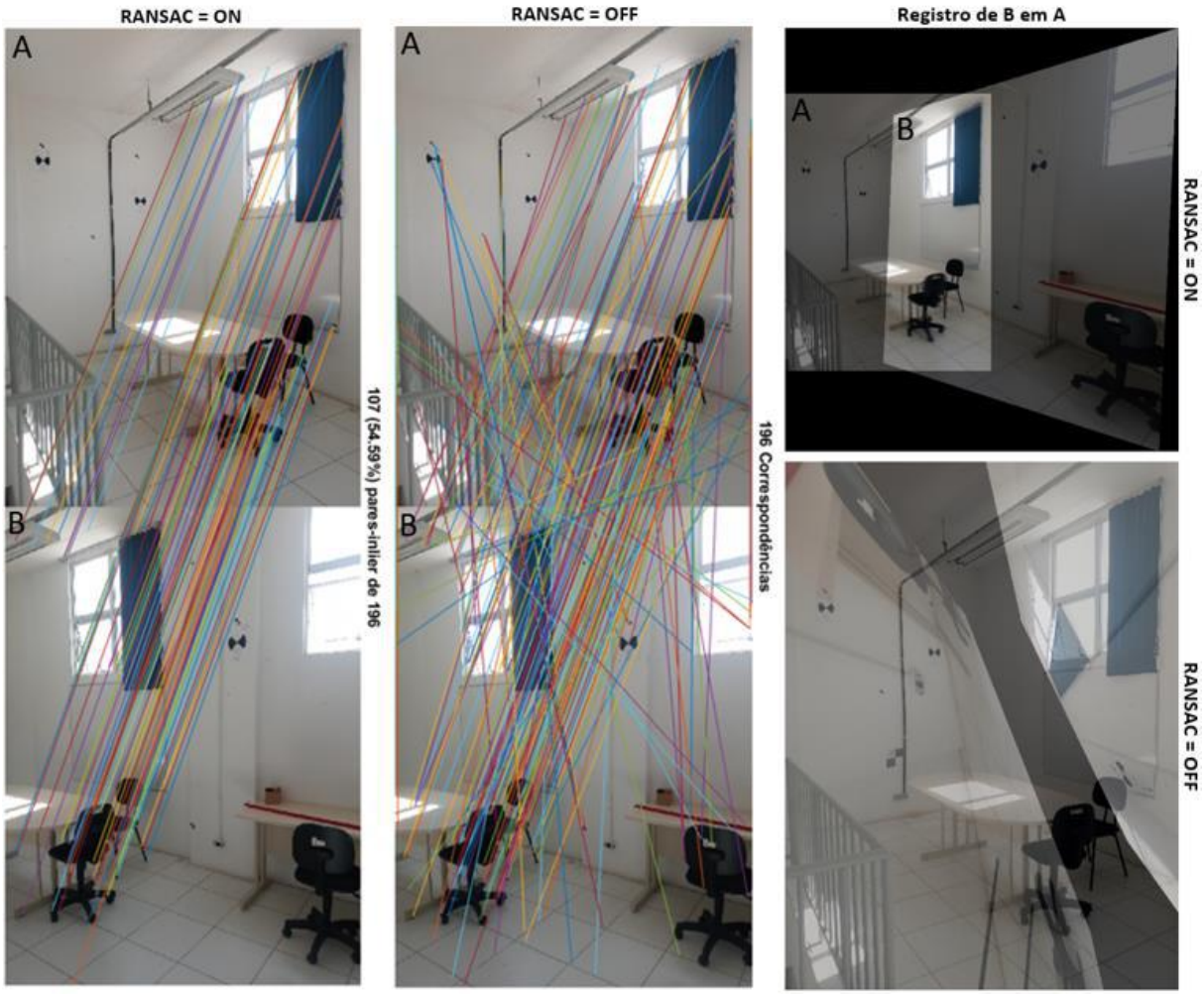
3. Resultados e Discussão

Em cada par se analisou o encontro de correspondências e a qualidade do mosaico resultante destas. O tempo total de execução da detecção de correspondências com o SIFT, classificação do RANSAC e cálculo da Homografia final, também é apresentado para cada par. Não se leva em consideração o tempo de importação, amostragem e desenho do par de imagens sobrepostas.

Registro dos pares 1 a 4 - imagens da câmera do LG-K51s

Estes são os pares mais desafiadores de serem registrados. Os 4 pares de imagens obtidos com câmera do LG-K51s podem ser observados nas Figuras 9, 10, 11 e 12. São apresentados os registros e as correspondências encontradas em cada par antes e depois do RANSAC. Faz-se uma breve discussão de cada resultados.

Figura 9 - Correspondências e mosaico do Par_1. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 1,9 segundos. Total de iterações do RANSAC: 31.

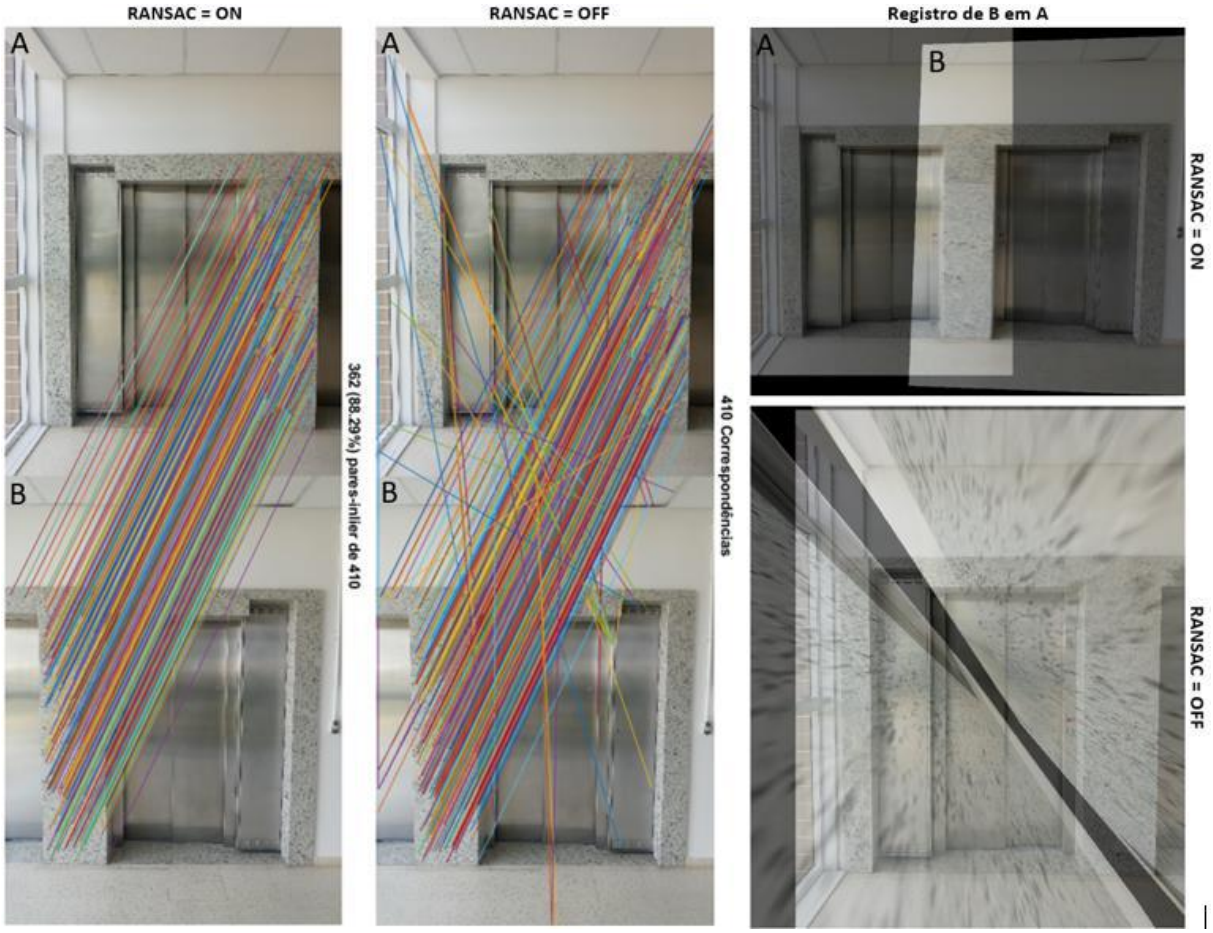


Fonte: Autores.

Na Figura 9 se vê que no par_1 há pouca sobreposição, a imagem A sobrepõe-se em menos de 30 % da imagem B e a

imagem B sobrepõe-se em menos de 40 % da imagem A. A rotação entre o par faz com que a homografia altere o comprimento da imagem B ao projetá-la no plano da imagem A durante a construção do mosaico. A quantidade de pares identificados pelo SIFT é bem reduzida, apenas 196 pares, todavia a homografia calculada com o RANSAC utiliza apenas 55 % destes, resultando em um mosaico adequado e de fácil interpretação, embora não seja perfeito. Por outro lado, a homografia calculada com todas as correspondências não é aceitável sob nenhum ponto de vista.

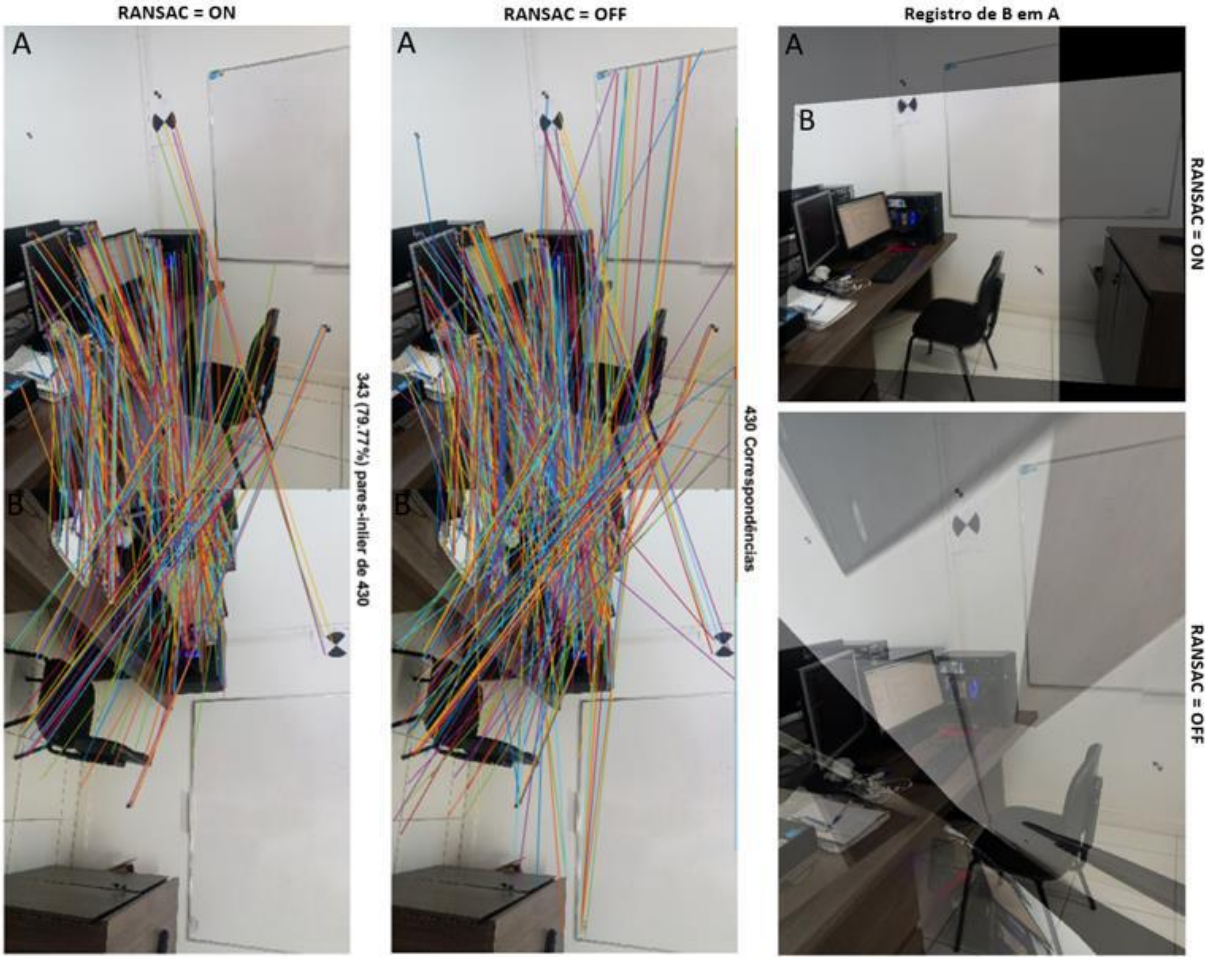
Figura 10 - Correspondências e mosaico do Par_2. A (imagem referência). B (imagem pesquisa). Tempo de execução (mediana): 2,5 segundos. Total de iterações do RANSAC: 6.



Fonte: Autores.

Na Figura 10 tem-se o Par_2, desta vez a baixa sobreposição é culpa da elevada translação entre as imagens, menor ainda do que as imagens do Par_1. Como esperado, o tempo de execução do algoritmo foi maior, embora a quantidade de inliers tenha sido maior. Isto não se deve ao RANSAC, que convergiu em penas 11 iterações devido à alta proporção de inliers, mas sim devido a etapa de extração e encontro de correspondências com o SIFT, responsável por quase todo o tempo de execução do modelo. Se observa que a homografia entregue foi suficiente para realizar o registro adequado do par de imagens, e caso se desconsidere a etapa do RANSAC o resultado será desastroso. Isto mostra que apenas 11.7 % de outliers nas correspondências é suficiente para corromper totalmente o cálculo da homografia.

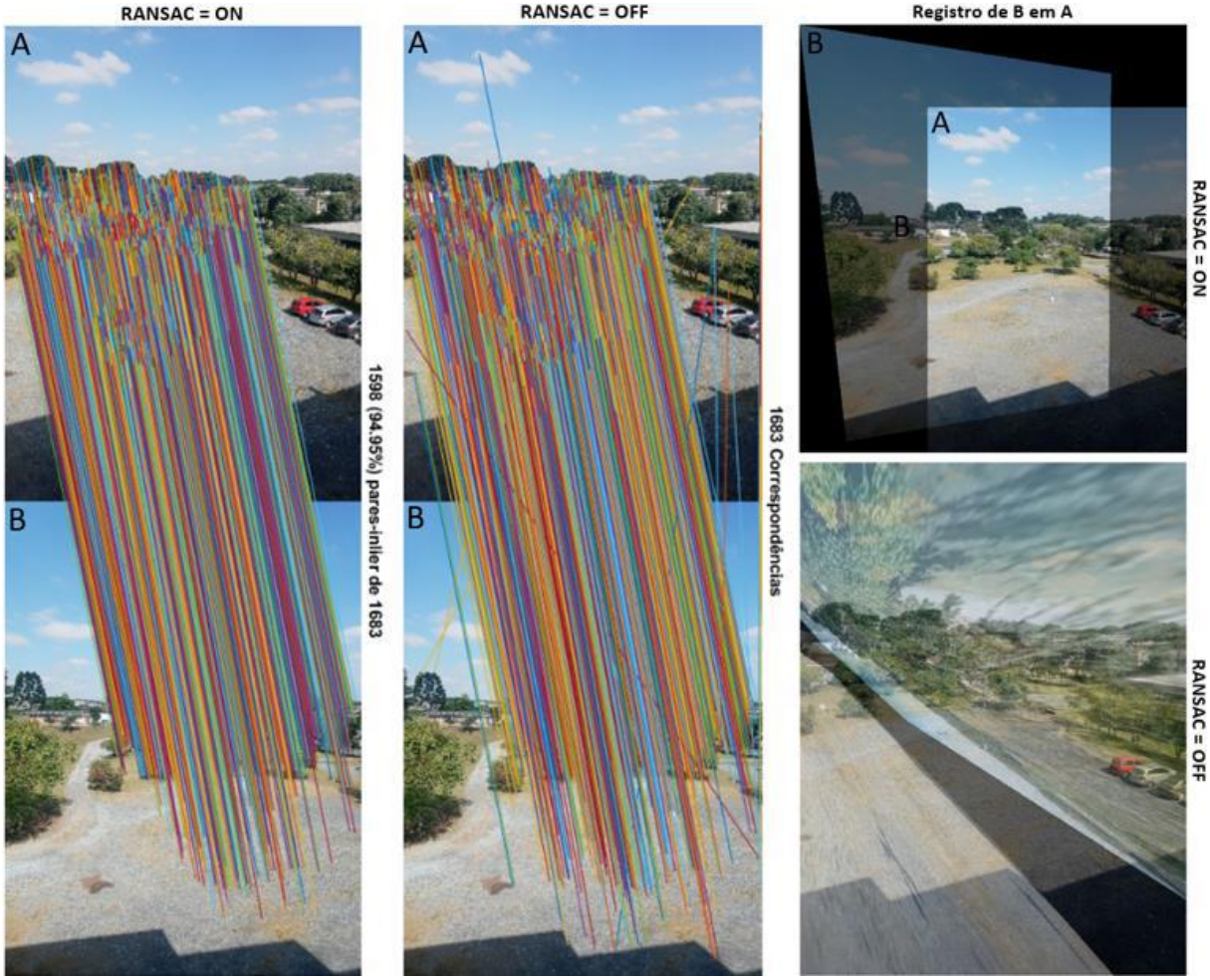
Figura 11 - Correspondências e mosaico do Par_3. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 2,0 segundos. Total de iterações do RANSAC: 15.



Fonte: Autores.

Na Figura 11 tem-se o Par_3, a principal característica deste par é o fato de a câmera estar rotacionada 90º em torno do seu eixo, onde a imagem A está na vertical e a imagem B está na horizontal. A sobreposição mútua entre as imagens é maior que 50 %, mas a taxa de inliers foi menor do que no par anterior. O tempo de execução foi curto, isto mostra que rotações não são um problema para o registro de imagens, desde que haja sobreposição adequada no par. O resultado do registro é semelhante aos anteriores, há um leve borrão na parte sobreposta e a homografia calculado sem o RANSAC é desastrosa.

Figura 12 - Correspondências e mosaico do Par_4. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 5,7 segundos. Total de iterações do RANSAC: 5.



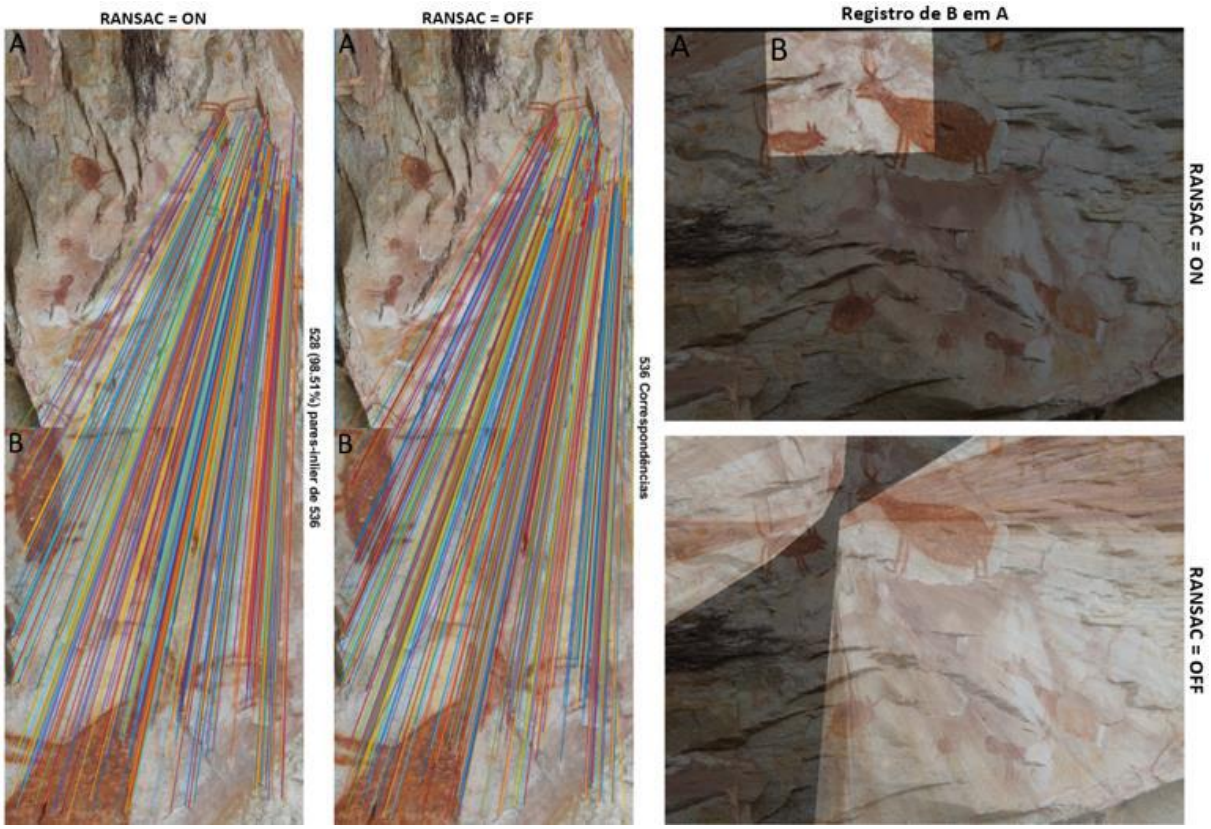
Fonte: Autores.

Na Figura 12 tem-se o Par_4, neste par a taxa de inliers entre as correspondências foi elevada ($\omega \cong 0{,}95$). Isto fez com que o praticamente não ocorressem borrões na área sobreposta. Mais uma vez é evidente que uma abordagem que utilize todas as correspondências dadas pelo SIFT não é aceitável, pois apenas 5% de outliers nos pares foi suficiente para degradar o mosaico construído com a homografia encontrada sem o modelo RANSAC. Devido a alta taxa de inliers o RANSAC praticamente não necessitou de iterações para convergir. O tempo de execução foi maior devido ao número de correspondência do SIFT. Em relação ao par com menos correspondências (Par_1) a quantidade de correspondências encontradas no Par_4 foi 8,6 vezes maior, mas o tempo de execução foi apenas 3 vezes maior.

Registro dos pares 5 a 8 - imagens da câmera profissional (Nikon-P600)

O resultado do modelo nos pares de imagens do sítio arqueológico da Pedra Pintada, obtidos com câmera profissional Nikon-P600, podem ser observados nas Figuras 13, 14, 15 e 16. São apresentados os registros e as correspondências encontradas em cada par antes e depois do RANSAC. As imagens com as correspondências estão na vertical para que a organização seja mantida igual aos casos anteriores.

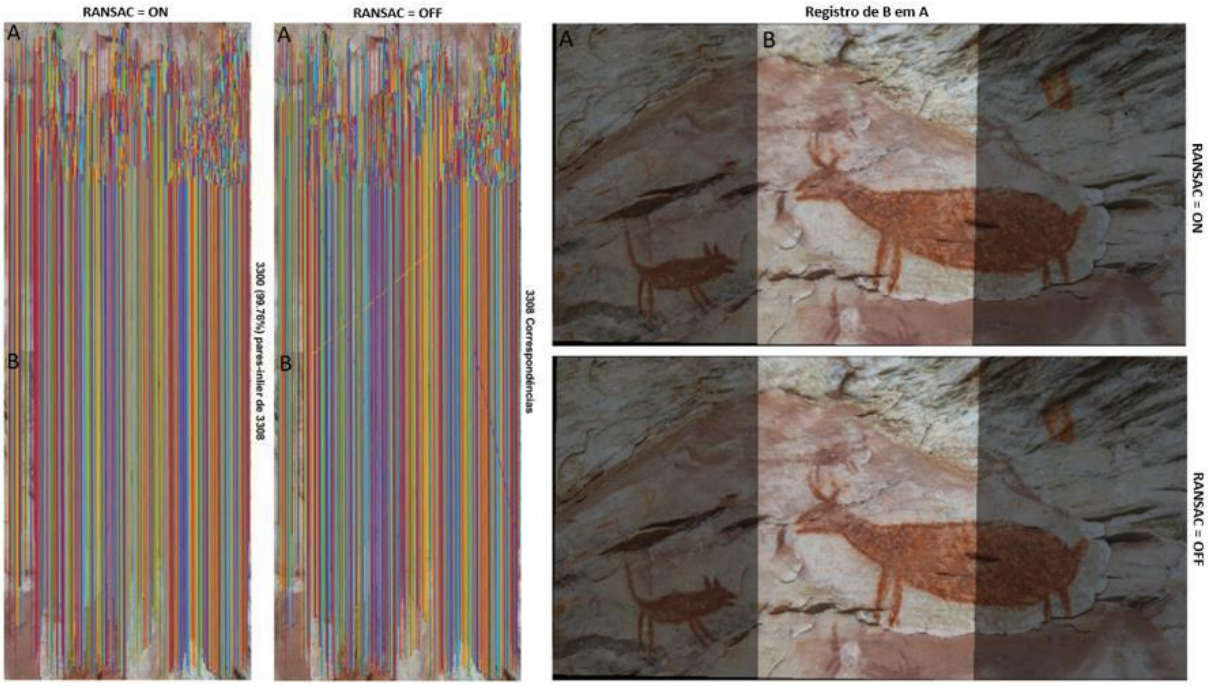
Figura 13 - Correspondências e mosaico do Par_5. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 2,6 segundos. Total de iterações do RANSAC: 3.



Fonte: Autores.

Na Figura 13 tem-se o Par_5, neste par há uma elevada mudança de escala entre as fotografias, a imagem A está distante e a imagem B está bem próxima da Pedra Pintada. A taxa de inliers é de quase 99%, o que praticamente não exige interações do RANSAC, todavia, caso esta etapa seja desconsiderada, a homografia ainda falha completamente. Apenas 1,5 % de *outliers* é suficiente para corromper o resultado do registro. O tempo de execução foi similar ao Par_2 e ao Par_3, assim como a quantidade de pares correspondentes encontrados.

Figura 14 - Correspondências e mosaico do Par_6. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 5,3 segundos. Total de iterações do RANSAC: 2.



Fonte: Autores.

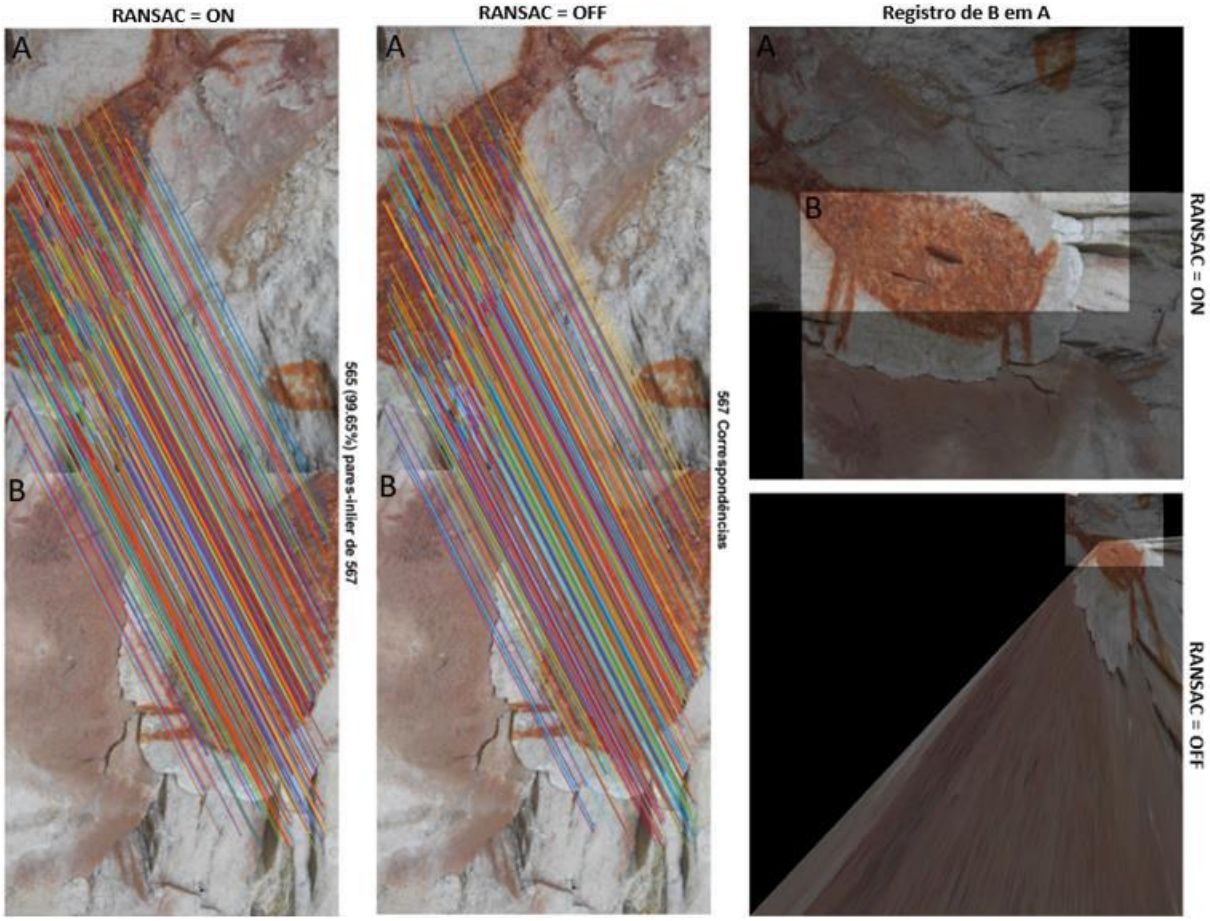
Na Figura 14 tem-se o Par_6, a sobreposição média deste par é de 50%. Novamente a quantidade de inliers foi bastante alta, o que exigiu apenas duas iterações do RANSAC. Pela primeira vez não ocorreram diferenças grosseiras entre os registros com e sem o RANSAC. Neste par, os *outliers* respondem por apenas 0,25 % das correspondências, o que é uma consequência direta da qualidade da imagem. O tempo de execução foi longo, pois mais 3308 correspondências foram encontradas, sendo apenas 8 delas ruins.

Figura 15 - Correspondências e mosaico do Par_7. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 3,6 segundos. Total de iterações do RANSAC: 5.

Fonte: Autores.

Na Figura 15 tem-se o Par_7, este par tem menos de 20 % de sobreposição entre as imagens. Se nota que a quantidade de iterações do RANSAC não ultrapassa uma dezena mesmo neste par onde a corrupção das correspondências é superior a 40 %. O tempo de execução foi similar ao do Par_5, a quantidade de correspondências encontradas também, isto ocorre porque a quantidade de sobreposição entre as imagens é similar. Existe uma correlação entre o tempo de execução de extração de feições, a quantidade de sobreposição, e a quantidade de correspondências encontradas. Esta última, conforme se observou nas imagens menos detalhadas, varia muito com a textura dos objetos.

Figura 16 - Correspondências e mosaico do Par_8. A (imagem de referência). B (imagem de pesquisa). Tempo de execução (mediana): 2,9 segundos. Total de iterações do RANSAC: 2.



Fonte: Autores.

Na Figura 16 tem-se o Par_8, este par tem sobreposição aproximada de 30 %. Neste par apenas duas correspondências foram detectadas como *outliers*, o que corresponde a apenas 0,35 % do total de pares, contudo, foi o suficiente para deteriorar totalmente a homografia calculada, resultando em um mosaico absurdo. Fica claro que, quando se trada do resultado do registro, não importa a proporção de *outliers* nos dados, mas sim a magnitude destes em relação ao restante do conjunto. Diante deste último resultado é evidente a importância de modelos que se baseiem medidas robustas, tais como a mediana, MAD, índice interquartílico, média truncada, entre outras, que desconsideram os extremos durante a análise dos dados. O tempo de execução foi inferior ao par anterior, mesmo com mais correspondências.

4. Considerações Finais

Este trabalho apresentou um modelo para registro de pares de imagens que opera por meio do encontro de correspondências calculadas com o SIFT, filtra as correspondências com o modelo RANSAC adaptado ao descritor, e aplica a transformação projetiva de Homografia estimada. Os resultados revelam que a integração do SIFT e do RANSAC com o modelo de Homografia, na geração de mosaicos, continua a ser um framework de grande confiabilidade, capaz de eliminar falsas correspondências e entregar mosaicos de qualidade com fotografias profissionais ou não. Para imagens obtidas com sensores comuns, como o do smartphone de baixo custo LG-K51s, o resultado ideal somente pode ser garantido se houver cuidado na tomada das imagens. Deve-se garantir que não haja grande mudança de posição entre as capturas, de forma que o

par de imagens se sobreponha em pelo menos 60 % da área de ambas.

Considerando os mosaicos gerados com todos os pares, são apontadas duas conclusões principais: (1) a geometria da tomada de fotos é um critério essencial para garantir a geração de mosaicos geometricamente consistentes; (2) poucas correspondências entre as imagens permitem a geração de mosaicos consideravelmente melhores que mosaicos gerados com muitas correspondências que estejam contaminadas com erros.

Há a possibilidade de acelerar a convergência do RANSAC com checagens da geometria dos pares correspondentes, as geometrias degeneradas, como pontos próximos da colinearidade, podem ser removidas previamente nas iterações de amostragem, todavia essa alternativa não traria benefícios na velocidade do framework como um todo, pois o encontro das correspondências pelo descritor é o que majoritariamente consome mais tmepo.

Recomenda-se para trabalhos futuros testar outros descritores de imagens alternativos ao SIFT, tais como ORB (Rublee et al., 2011), SUFR (Bay et al., 2008) e BRISK (Leutenegger et al., 2011), em particular, recomenda-se observar a sensibilidade do descritor ao limiar de erro ε do RANSAC e o tempo de encontro de correspondências. Outra recomendação promissora de pesquisa futura é na direção ao registro de nuvens de pontos 3D, pois o encontro de correspondências entre imagens pode ajudar no encontro de correspondências entre nuvens de pontos 3D, desde que sejam conhecidos os parâmetros de transformação entre a câmera e o laser scaner que levanta os pontos da nuvem (Weinmann, 2016).

Uma das vantagens disso é que o encontro de correspondências entre imagens é muito mais rápido e robusto do que o encontro de correspondências entre nuvens de pontos, pois imagens são um dado estruturado com espaço de busca significativamente menor. Atualmente, o tempo gasto no registro entre nuvens de pontos 3D massivas é um dos maiores gargalos no processamento deste tipo de dado, tal etapa é feita de forma quase sempre manual, e quando é feita de automática, ainda leva horas para ser concluída (Theiler et al., 2015); (Dong et al., 2018).

Agradecimentos

Os autores agradecem à Universidade Federal do Paraná (UFPR) pelos recursos disponibilizados e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo incentivo à pesquisa com bolsas de estudo.

Referências

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Belo, F. A. W. (2006). Desenvolvimento de Algoritmos de Exploração e Mapeamento Visual para Robôs Móveis de Baixo Custo.

Berveglieri, A. (2014). *Localização automática de pontos de controle em imagens aéreas baseada em cenas terrestres verticais*.

Bouguet, J.-Y. Camera calibration toolbox for Matlab (2008). Computational vision at the California institute of technology. 2008. <http://www.vision.caltech.edu/bouguetj/calib_doc/>.

Brown, M., & Lowe, D. G. (2002). Invariant features from interest point groups. *BMVC*, *4*, 398–410.

Bruce, A., & Bruce, P. (2019). *Estatística Prática para Cientistas de Dados*. Alta Books.

Chum, O., & Matas, J. (2008). Optimal randomized RANSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(8), 1472–1482.

Dong, Z., Yang, B., Liang, F., Huang, R., & Scherer, S. (2018). Hierarchical registration of unordered TLS point clouds based on binary shape context descriptor. *ISPRS Journal of Photogrammetry and Remote Sensing*, *144*, 61–79.

dos Santos, M. C., & Rocha, A. (2012). Revisao de Conceitos em Projeçao, Homografia, Calibraçao de Câmera, Geometria Epipolar, Mapas de Profundidade e Varredura de Planos. *Unicamp, Campinas, Tech. Rep*.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, *65*(20), 20TR01.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

Kumar, R., Anandan, P., Irani, M., Bergen, J., & Hanna, K. (1995). Representation of scenes from collections of images. *Proceedings IEEE Workshop on Representation of Visual Scenes (In Conjunction with ICCV'95)*, 10–17.

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. *2011 International Conference on Computer Vision*, 2548–2555.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Mills, A., & Dudek, G. (2009). Image stitching with dynamic elements. *Image and Vision Computing*, *27*(10), 1593–1602.

Paul, S., & Pati, U. C. (2021). A comprehensive review on remote sensing image registration. *International Journal of Remote Sensing*, *42*(14), 5396-5432.

Pons, J.-P., Keriven, R., & Faugeras, O. (2007). Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, *72*(2), 179–193.

Raguram, R., Frahm, J.-M., & Pollefeys, M. (2008). A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. *European Conference on Computer Vision*, 500–513.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision*, 2564–2571.

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Theiler, P. W., Wegner, J. D., & Schindler, K. (2015). Globally consistent registration of terrestrial laser scans via graph optimization. *ISPRS Journal of Photogrammetry and Remote Sensing*, *109*, 126–138.

Torr, P. H. S., & Zisserman, A. (2000). MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, *78*(1), 138–156.

Wang, J., & Watada, J. (2015). Panoramic image mosaic based on SURF algorithm using OpenCV. *2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP) Proceedings*, 1–6.

Weinmann, M. (2016). *Reconstruction and analysis of 3D scenes* (Vol. 1). Springer.

Yang, Z.-L., & Guo, B.-L. (2008). Image mosaic based on SIFT. *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 1422–1425.

Zuliani, M. (2009). RANSAC for Dummies. *Vision Research Lab, University of California, Santa Barbara*.