2
SIFT (Scale Invariant Feature Transform)

O SIFT é um algoritmo de visão computacional proposto e publicado pelo pesquisador David Lowe (Lowe, 1999), professor do departamento de Ciência da Computação na *University of British Columbia* no Canadá.

Neste trabalho foi utilizada a implementação do algoritmo SIFT disponibilizada pelo seu autor no site *http://www.cs.ubc.ca/~lowe/keypoints/*. Esta implementação consiste de um programa na linguagem C que localiza e extrai características relevantes de duas imagens selecionadas, fornecendo descritores invariantes a diversas transformações, como rotação, translação, escala, além de possuírem robustez a ruído e iluminação.

A seguir, são apresentadas de forma resumida as etapas que compõem este método e, na parte final do capítulo, é descrito como é feito o emparelhamento de pontos homólogos utilizando os descritores SIFT, processo chamado de *matching*.

Uma descrição completa do algoritmo pode ser encontrada em (Lowe, 2004), (Lowe, 2001) e (Lowe, 1999).

2.1.
Etapas do SIFT

A construção dos descritores SIFT é feita por meio de quatro etapas principais. As duas primeiras descrevem a parte do detector e as duas seguintes descrevem a formação do descritor.

2.1.1.
Detecção de Extremos

O primeiro passo consiste na construção de uma pirâmide de imagens, como detalhado a seguir. A partir de uma imagem de entrada, $I(x, y)$, novas versões da imagem são geradas pela aplicação sucessiva de um filtro de suavização Gaussiano, $G(x, y, \sigma_g)$, conforme mostra a equação abaixo:

$$F(x, y, \sigma_g) = G(x, y, \sigma_g) * I(x, y) \tag{2.1}$$

onde σ_g denota o desvio padrão do filtro Gaussiano e define a escala ou nível da imagem filtrada $F(x, y, \sigma_g)$.

O conjunto de imagens resultantes é chamado de oitava. Seguidamente, uma destas imagens tem seu tamanho reduzido à metade e serve de entrada para a próxima geração de oitavas. O processo pode ser repetido várias vezes até que se obtenha o número de oitavas desejado, formando-se assim uma pirâmide, como mostrada na Fig. 2.1.

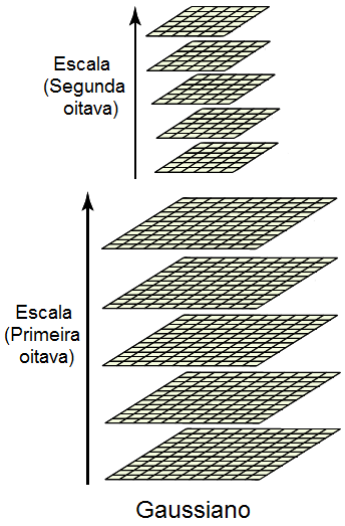

Figura 2.1 Pirâmide de Gaussianas.

A utilização da função gaussiana tem o objetivo de obter amostras da imagem, onde detalhes indesejados são suavizados ou eliminados e características fortes são realçadas. A variação de σ_g permite encontrar tais características em diferentes escalas.

Depois da geração da pirâmide de Gaussianas, em cada oitava é calculada a diferença entre imagens filtradas com escalas variadas por uma constante *k*. Esta operação é implementada usando uma função DoG (*Difference of Gaussian*) definida por:

$$D(x, y, \sigma_g) = F(x, y, k\sigma_g) - F(x, y, \sigma_g) \tag{2.2}$$

onde $k\sigma_g$ e σ_g representam escalas adjacentes da pirâmide Gaussiana.

O resultado da aplicação da Eq. (2.2) é a chamada pirâmide de diferença de Gaussianas, mostrada na Fig. 2.2. O seguinte passo é detectar valores de máximo

ou mínimo locais em cada nível da pirâmide, chamados de extremos, o que confere ao método invariância quanto à escala.

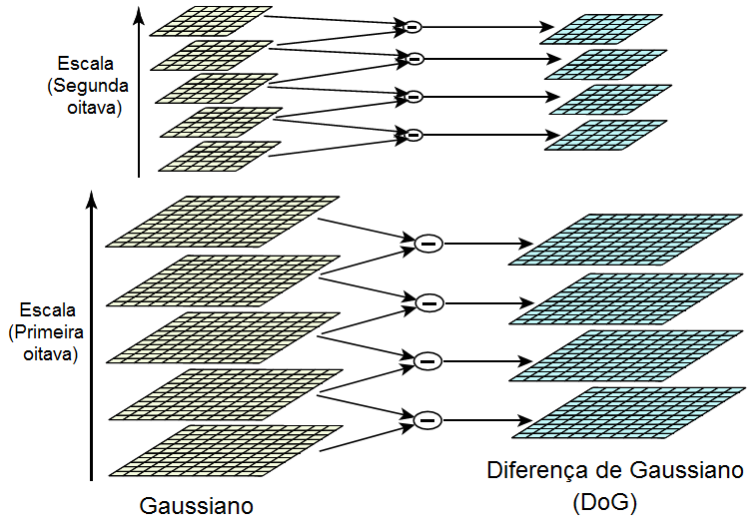

Figura 2.2 Processo de construção das imagens resultantes da Diferença de Gaussianas DoG.

Os extremos podem ser filtrados com ajuda da pirâmide de diferenças Gaussianas. Na Figura 2.3, por exemplo, a intensidade do *pixel* marcado como "X" é comparada com a dos seus vizinhos marcados como "O" na própria escala e nas escalas adjacentes. Desta forma, é feita a seleção dos pontos candidatos a serem pontos-chave.

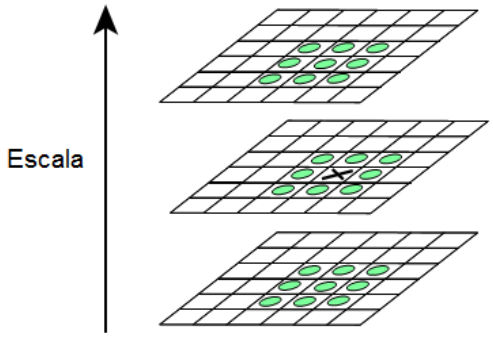

Figura 2.3 Detecção de extremos nas escalas adjacentes.

2.1.2. Localização Precisa dos Pontos-chave

Todos os pontos candidatos detectados na etapa anterior são ajustados para a localização, escala e razão das curvaturas principais. Isto permite retirar aqueles

que possuem baixo contraste, sensíveis a ruídos, ou aqueles mal localizados ao longo de uma aresta.

Para aumentara exatidão da localização em que foi detectado um extremo, uma superfície de segunda ordem é ajustada ao ponto de amostragem local de modo a determinar uma localização interpolada do máximo. Esta abordagem faz uso de uma expansão por séries de Taylor da função *D*, deslocada de modo que a origem seja o ponto de amostragem:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \tag{2.3}$$

onde *D* e suas derivadas são calculadas no ponto de amostragem e, $\mathbf{x} = (x, y, \sigma)^T$ é o deslocamento (*offset*) em relação a este ponto.

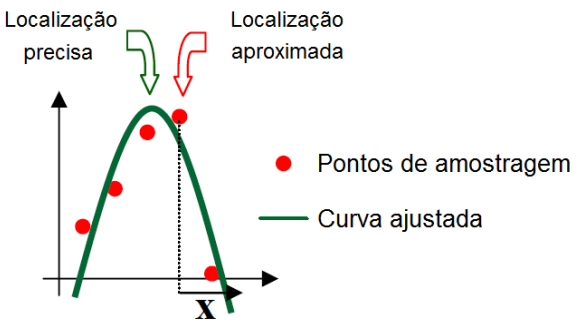

Figura 2.4 Localização precisa do ponto-chave.

A localização e a escala em que a função *D* alcança seu valor extremo passam a ser os novos valores de localização e escala do ponto sob análise. A localização do extremo, $\hat{\mathbf{x}}$, é estimada calculando-se a derivada da Eq. (2.3) em relação a **x** e igualando o resultado a zero, obtendo:

$$\hat{\mathbf{x}} = -\left(\frac{\partial^2 D}{\partial \mathbf{x}^2}\right)^{-1} \frac{\partial D}{\partial \mathbf{x}} \tag{2.4}$$

Como resultado, tem-se um sistema linear 3×3, solucionado por mínimos quadrados. Se $\hat{\mathbf{x}} > 0.5$, isto indica que o extremo está mais próximo de outro ponto. Neste caso, o ponto é realocado e uma nova interpolação é realizada.

Substituindo a Eq. (2.4) na Eq.(2.3), obtém-se o valor da função no extremo, $D(\hat{\mathbf{x}})$. Assim, tem-se

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D^T}{\partial \mathbf{x}}\hat{\mathbf{x}} \tag{2.5}$$

Este parâmetro é utilizado para rejeitar pontos instáveis de baixo contraste através de um limiar previamente definido. Mediante testes experimentais, Lowe sugere um limiar de 0.03, assumindo que os *pixels* estão normalizados entre valores de 0 e 1.

Todavia, fazer o refinamento por um limiar ainda não é suficiente, pois a função DoG possui altos valores de resposta ao longo de arestas, mesmo que a localização ao longo da borda seja mal determinada.

Para solucionar esse problema, calcula-se a razão entre curvaturas, partindo do pressuposto que: um pico mal definido na função DoG forma uma grande curvatura principal do lado oposto da borda e uma pequena curvatura na direção perpendicular.

Para isto, utiliza-se uma matriz Hessiana 2×2, H, na localização e escala do ponto-chave na função D.

$$H(x, y) = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{2.6}$$

onde D_{xx}, D_{yy} e D_{xy} são as derivadas parciais de segunda ordem da função D.

Como os autovalores de H são proporcionais às principais curvaturas de D, calcula-se a razão entre o traço (soma dos autovalores) e o determinante (produto dos autovalores) de H.

Considerando-se α o autovalor com maior magnitude e β o de menor magnitude, é possível calcular a soma e o produto destes autovalores:

$$\frac{Tr(H)^2}{Det(H)} = \frac{(D_{xx} + D_{yy})^2}{D_{xx}D_{yy} - (D_{xy})^2} = \frac{(\alpha + \beta)^2}{\alpha\beta} \tag{2.7}$$

Seguidamente, considera-se r como a razão entre o autovalor de maior magnitude e o de menor, de modo que $\alpha = r\beta$, tem-se

$$\frac{Tr(H)^2}{Det(H)} = \frac{(r+1)^2}{r} \tag{2.8}$$

Se o determinante é negativo, isto significa que as curvaturas possuem sinais diferentes e o ponto sob análise é descartado. Para definir quais pontos serão pontos-chave, aplica-se um limiar na razão de curvaturas, i.e.

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r_{\max}+1)^2}{r_{\max}} \tag{2.9}$$

Lowe propõe o uso de $r_{max} = 10$, assim eliminam-se pontos instáveis próximos a extremidades, descartando-se pontos abaixo de determinado limiar. Um exemplo de detecção e localização de pontos-chaves é mostrado na Fig. 2.5.

Figura 2.5 Exemplo de detecção de pontos-chaves. Imagem original (esquerda) e 9250 pontos localizados (direita).

2.1.3. Atribuição da Orientação dos Descritores

Nesta etapa, a cada ponto-chave é atribuída uma orientação correspondente à direção predominante do gradiente em torno do ponto-chave. Para tanto, seleciona-se a imagem filtrada, $F(x, y, \sigma_g)$, no nível da pirâmide Gaussiana com a escala mais próxima ao ponto-chave avaliado. A seguir, a magnitude $m(x, y)$ e a orientação $\theta(x, y)$ do gradiente na vizinhança em torno de cada ponto são calculados por:

$$m(x,y) = \sqrt{L_1^2 + L_2^2} \tag{2.10}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{L_2}{L_1}\right) \tag{2.11}$$

onde:

$$F_1 = F(x+1, y) - F(x-1, y) \tag{2.12}$$

$$F_2 = F(x, y+1) - F(x, y-1) \tag{2.13}$$

Para cada região, monta-se um histograma de orientações do gradiente divido em 10 intervalos, cobrindo todas as orientações possíveis de 0° a 360°, como mostrado na Fig. 2.6. Na montagem dos histogramas, a contribuição de cada ponto da vizinhança é ponderada por uma série de pesos. O primeiro, conforme uma função de distância normalizada entre a orientação dos *pixels* e a orientação do ponto-chave; o segundo, com base na magnitude do gradiente; e o terceiro, usando uma janela Gaussiana circular com desvio padrão 1,5 vezes maior do que a escala do ponto-chave. Com esses pesos, o histograma é atualizado.

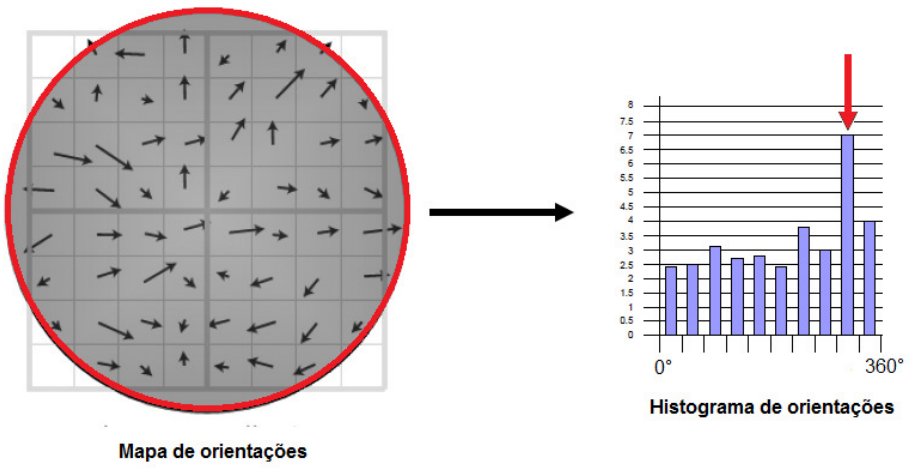

Figura 2.6 Determinação da orientação principal do ponto-chave.

Os picos no histograma de orientação representam as direções dominantes dos gradientes locais. Além do pico máximo, também são usados, para definir a orientação, os picos com valor acima de 80% em relação ao maior. No final, ainda se aplica um ajuste parabólico aos três valores mais próximos de cada pico, a fim de interpolar a posição com melhor exatidão.

A Fig. 2.7 mostra as orientações calculadas para cada ponto-chave localizado na imagem. Ao se atribuir uma orientação consistente para cada ponto-chave, podem-se representar os descritores em relação à orientação calculada, conseguindo-se assim invariância à rotação.

Finalmente, é possível construir os descritores para cada ponto-chave definido, que agora possui quatro dimensões: posição em *x*, posição em *y*, escala e orientação.

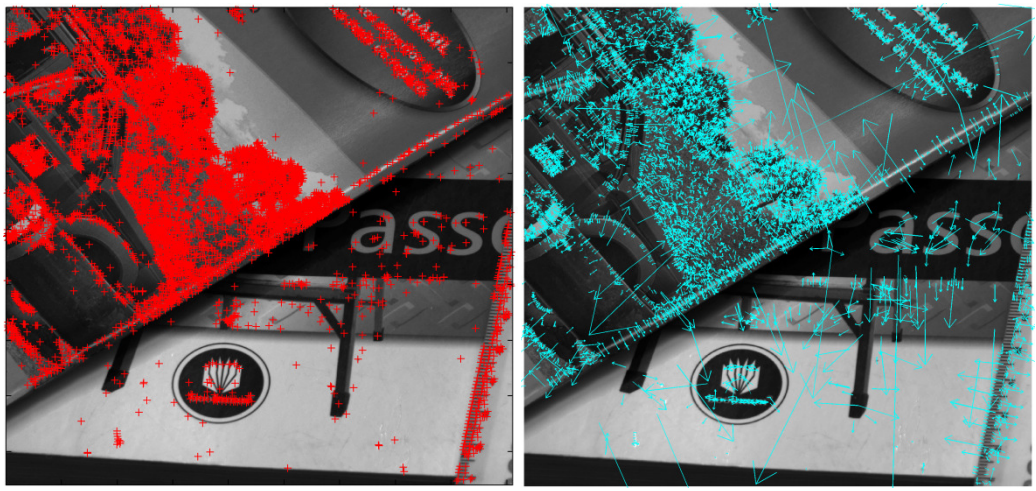

Figura 2.7 Exemplo da atribuição das orientações dos pontos-chave, localizados na imagem da esquerda, representadas por vetores na imagem da direita.

2.1.4. Construção do Descritor Local

Nesta etapa são calculados os descritores dos pontos-chave determinados na etapa anterior. Inicialmente, são calculados os gradientes em uma vizinhança ao redor de cada ponto-chave, os quais são ilustrados com pequenas setas no lado esquerdo da Fig. 2.8. Para isto, uma janela de suavização Gaussiana com desvio padrão igual à metade da janela do descritor é utilizada para dar peso à magnitude do gradiente de cada ponto vizinho. O Gaussiano evita mudanças súbitas do descritor a pequenas mudanças na posição da janela, e também reduz a ênfase nos gradientes mais afastados, que são os mais afetados por erros.

Depois de efetuada a suavização dos gradientes, a vizinhança em torno do ponto-chave é dividida em $n \times n$ regiões de $m \times m$ *pixels*. Para cada região, monta-se um histograma para 8 direções com base nas magnitudes dos pixels, a partir do qual o descritor é construído. Para o caso de um conjunto 4×4 de histogramas com 8 células de acumuladores, gera-se um vetor resultante de 128 elementos para cada ponto-chave, conforme mostrado na Fig. 2.8.

É importante notar que as orientações dos gradientes são medidas em relação à orientação do ponto-chave. Dessa forma, o uso de direções relativas ao invés de direções absolutas faz com que o descritor seja invariante quanto à rotação.

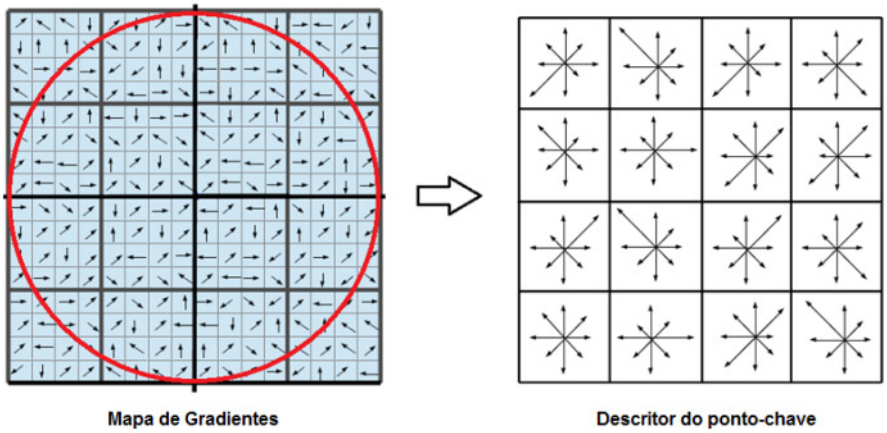

Figura 2.8 Construção do descritor SIFT.

Todavia, duas imagens de um mesmo objeto podem possuir variações de luminosidade que modifiquem sensivelmente os descritores obtidos. Para evitar isto, o vetor é normalizado e impõe-se um limiar máximo para que as direções com magnitudes muito acentuadas não prevaleçam na representação do descritor. Depois, o vetor é normalizado novamente. Este passo visa obter a invariância à iluminação.

Finalmente, para cada imagem são construídos diversos descritores, cada um referente a um ponto-chave. Tem-se como resultado, um conjunto de descritores robustos para identificar unicamente cada ponto-chave, que podem ser utilizados em processos de correspondência com outras imagens.

2.2. *Matching* ou Casamento de Pontos Correspondentes

O conceito de *matching* envolve procurar os pontos em comum em cada uma das imagens, baseado na semelhança dos descritores. Normalmente, os candidatos à melhor correspondência são descritores similares, de maneira que os melhores candidatos podem ser escolhidos através do algoritmo do vizinho mais próximo.

A tarefa de se encontrar pontos correspondentes entre as duas imagens pode ser definida como se segue. Dadas duas imagens I_1 e I_2, com descritores representados por $des1_i$ e $des2_j$, respectivamente, onde i e j são os índices dos descritores para cada imagem. Assim,

$$des1_i = (\varphi_1, \varphi_2, \varphi_3 \ldots \varphi_{128}) \tag{2.14}$$

$$des2_j = (\psi_1, \psi_2, \psi_3 \ldots \psi_{128}) \tag{2.15}$$

onde φ *e* ψ são as magnitudes de cada elemento do descritor.

O vizinho mais próximo do descritor $des1_i$ para um i dado é definido pelo descritor $des2_j$ que possua a menor distância Euclidiana em relação a $des1_i$. Ou seja, se deseja encontrar j que minimize a função:

$$\left|des1_i - des2_j\right| = \sqrt{\left((\varphi_1 - \psi_1)^2 + (\varphi_2 - \psi_2)^2 + \cdots + (\varphi_{128} - \psi_{128})^2\right)} \tag{2.16}$$

Isto é feito para todo i, de modo a serem encontrados todos os pares de descritores correspondentes. No entanto, alguns pontos instáveis são detectados ao longo do processo, levando a falsas correspondências. Para a eliminação desse problema, um método para comparar a menor distância com a segunda melhor distância é usado, selecionando somente correspondentes próximos de um limiar:

$$limiar = \frac{vizinho\,mais\,próximo}{segundo\,vizinho\,mais\,próximo} \tag{2.17}$$

Na Fig. 2.9 são apresentadas as funções de densidade de probabilidade típicas para correspondências estabelecidas com sucesso e falsas correspondências, em termos da relação entre as distâncias ou limiar.

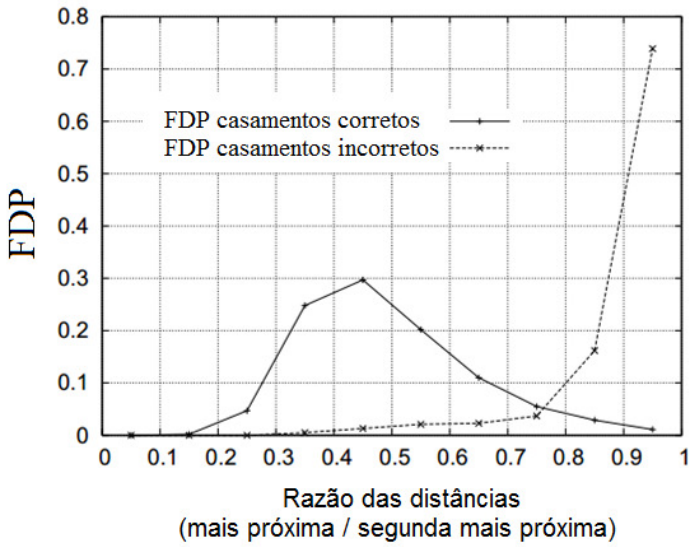

Figura 2.9 Função densidade de probabilidade típica para correspondência de pontos SIFT (Lowe, 2004).

Por exemplo, um limiar de 0,8 elimina em média 90% das falsas correspondências em diversas imagens estudadas, porém apenas descartando

menos de 5% das correspondências corretas. Portanto, as correspondências são assim eficientemente refinadas e os falsos pares são descartados.

Neste trabalho, com o objetivo de reduzir custo computacional, calcula-se o produto escalar entre os descritores ao invés das distâncias Euclidianas. Verifica-se que o cálculo do produto escalar é uma boa aproximação em relação às distâncias Euclidianas para pequenos ângulos.

Um exemplo do processamento da técnica SIFT e do algoritmo para emparelhamento de pontos correspondentes pode ser observado na Fig. 2.10.

Figura 2.10 Exemplo de *matching* entre duas imagens utilizando o algoritmo SIFT.

No capítulo seguinte, fundamentos de estereoscopia são apresentados para permitir a localização tridimensional dos pontos-chave localizados, identificando os deslocamentos nas três dimensões.