4
SIFT (Scale Invariant Feature Transform)

SIFT é um algoritmo de visão computacional publicado por David Lowe, em 1999 (Lowe, 1999) e patenteado nos EUA pela *University of British Columbia.*

SIFT é composto por duas partes distintas: o detector e o descritor. O detector SIFT é baseado em cálculos de diferença de Gaussianas e o descritor SIFT utiliza histogramas de gradientes orientados para descrever a vizinhança local dos pontos de interesse. A descrição de ambas as partes a seguir baseia-se (Lowe, 2004).

4.1.Etapas do Algoritmo SIFT

O algoritmo SIFT é executado através de quatro etapas principais: detecção de extremos, localização de pontos-chave, definição da orientação e descrição dos pontos-chave. As duas primeiras descrevem a parte do detector e as duas seguintes descrevem a formação do descritor.

A seguir, será explicado o funcionamento de cada estágio do algoritmo.

4.1.1. Detecção de Extremos

A primeira etapa da técnica SIFT consiste em buscar pontos que sejam invariantes a mudanças de escala da imagem, possibilitando a detecção de pontos com a câmera próxima ou distante do objeto de interesse. Tal objetivo é alcançado procurando características estáveis em diferentes escalas, utilizando uma função chamada de espaço de escala, que neste caso é a função Gaussiana.

Uma imagem I (x,y) passa a ser definida por uma $L(x,y,\sigma)$, no espaço-escala. Esta função é produzida pela convolução de uma função gaussiana, $G(x,y,\sigma)$, com a imagem, $I(x,y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{4.1}$$

onde

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2/2\sigma^2)} \tag{4.2}$$

Perceba que este filtro é variável à escala através do parâmetro σ.

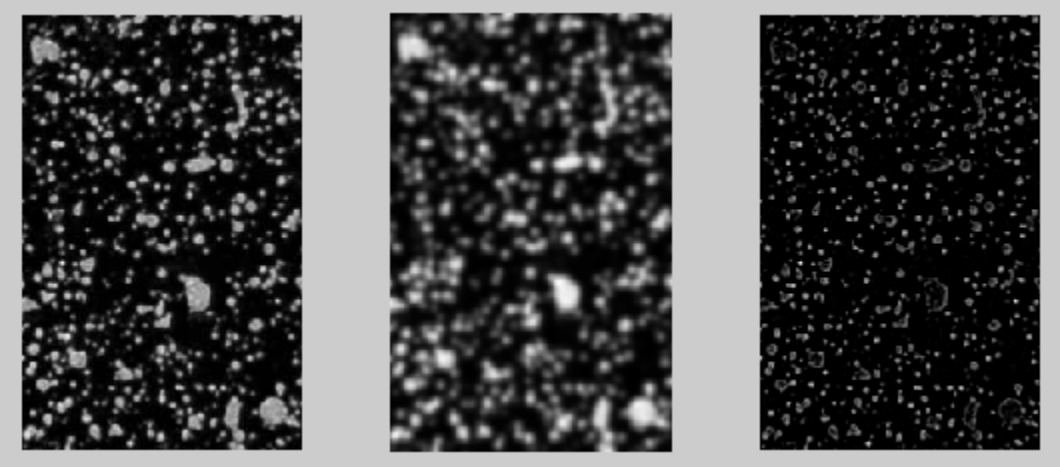

Figura 4.1 Exemplos de aplicação dos filtros. Imagem de padrão de granulados (esquerda), filtro Gaussiano (centro), e resultado da Diferença de Gaussianas (direita).

A eficiência da busca por pontos-chaves é aumentada com a utilização de uma função DoG ("*Difference of Gaussian*") formada pela diferença de imagens filtradas em escalas próximas, separadas por uma constante de escala *k*. A função DoG é definida por

$$DoG = G(x, y, k\sigma) - G(x, y, \sigma) \tag{4.3}$$

O resultado de efetuar a convolução de uma imagem com o filtro DoG é dado por

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{4.4}$$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{4.5}$$

Esta função DoG suaviza as imagens, e pode ser calculada pela simples subtração de imagens borradas por um filtro Gaussiano em escalas σ e $k\sigma$.

A utilização da função gaussiana tem o objetivo de obter mostras da imagem donde detalhes indesejados e ruídos são eliminados e características

fortes realçadas. Variando σ é possível encontrar tais características em diferentes escalas.

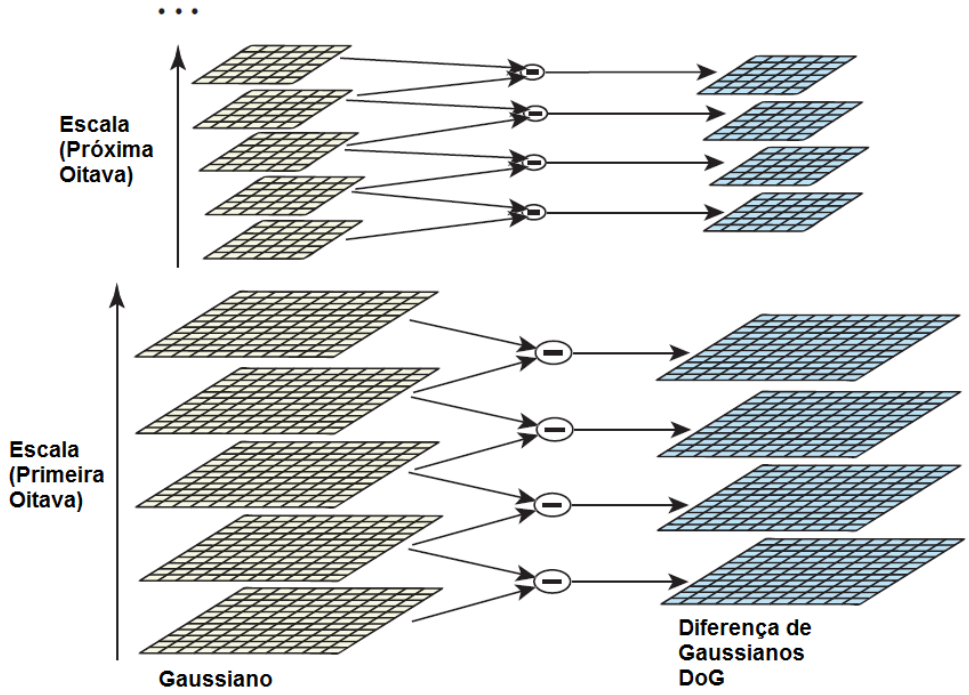

Figura 4.2 Representação do procedimento de obtenção das Diferenças de Gaussianas DoG para diversas oitavas de uma imagem. (Lowe 2004).

Um modo eficiente para a criação da Diferença de Gaussianas é esquematizado na Figura 4.2, cujos 4 passos são descritos a seguir.

1- A imagem inicial sofre convoluções incrementais com filtros Gaussianos para produzir imagens separadas por um fator de escala k no espaço de escala, representados na coluna esquerda.

2- Lowe considera que é necessário fazer a convolução da imagem até 2σ para ser possível a construção de descritores invariantes quanto à escala. Portanto, para se gerar em s intervalos, o fator de escala k é definido por $k = 2^{1/s}$, produzindo assim $s+3$ imagens na oitava de forma que a detecção de extremos cubra toda oitava.

3- Imagens em escalas adjacentes são subtraídas para produzir as imagens da Diferença do Gaussiano mostradas à direita (na Figura 4.2).

4- Uma vez processada a oitava, é reduzida a resolução da imagem (*downsample*) tomando-se cada segundo pixel da imagem no centro da oitava, gerando-se uma nova oitava (alteração da frequência de muestreo por um fator de dois), e voltando-se ao passo número 1.

A partir de então, será feita a detecção de extremos em cada intervalo de cada oitava. Um extremo é definido como qualquer valor no DoG maior do que todos os seus vizinhos no espaço-escala.

Os extremos são dados por valores de máximo ou mínimo locais para cada $D(x,y,\sigma)$, que podem ser obtidos comparando-se a intensidade de cada ponto com as intensidades de seus oito vizinhos na sua escala, com os nove pontos vizinhos na escala superior, e os nove vizinhos na escala inferior, representados na Figura 4.3. Na figura, o ponto marcado com "**X**" é comparado com seus vizinhos marcados como "**O**". As 3 imagens DoG apresentadas na Figura correspondem à diferença entre imagens adjacentes da pirâmide gaussiana.

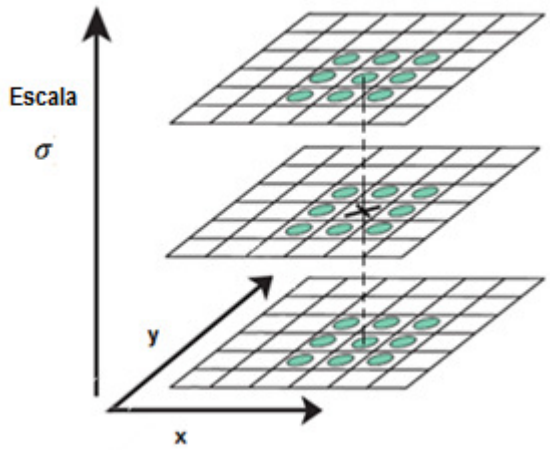

Figura 4.3 Detecção de extremos no espaço-escala (Lowe, 2004).

A próxima etapa é definir a localização dos pontos-chave e fazer o descarte de pontos instáveis.

4.1.2. Localização Precisa de Pontos Chaves

Todos os pontos detectados como extremos são candidatos a pontos-chave. Deseja-se agora calcular a localização exata destes pontos-chave.

O método consiste em ajustar uma função quadrática *3D* do ponto de amostragem local de modo a determinar uma localização interpolada do máximo.

Isto é feito utilizando uma expansão de Taylor da função Diferença de Gaussiano aplicada à imagem, *D(x,y,σ)*, deslocada de modo que a origem desta expansão esteja localizada no ponto de amostragem (Brown and Lowe, 2002):

$$D(\overline{x}) = D + \frac{\partial D^T}{\partial \overline{x}}\overline{x} + \frac{1}{2}\overline{x}^T \frac{\partial^2 D}{\partial x^2}\overline{x}... \tag{4.6}$$

$$\overline{x} = \left(x, y, \sigma\right)^T \tag{4.7}$$

Onde o valor de *D*, a sua primeira e segunda derivada são calculados no ponto de amostragem e $\overline{x}$, representa o deslocamento deste ponto.

A localização em *sub-pixels* do ponto de interesse é dada pelo extremo da função apresentada na equação (4.6). Esta localização, $\hat{x}$, é determinada ao se calcular a derivada de *D(* $\overline{x}$ *)* em relação a $\overline{x}$, e igualando o resultado a zero:

$$\frac{\partial D}{\partial \overline{x}} + \frac{\partial^2 D}{\partial \overline{x}^2}\hat{x} = 0 \tag{4.8}$$

Tem-se então a posição do extremo, dada por:

$$\hat{x} = -\frac{\partial^2 D^{T-1}}{\partial \overline{x}^2}\frac{\partial^2 D}{\partial \overline{x}} \tag{4.9}$$

O valor da função no extremo, $D(\overline{x})$, é útil para a rejeição de extremos instáveis com baixo contraste, que seriam sensíveis a ruído. Substituindo-se a equação (4.9) na equação (4.6) obtém-se:

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial \overline{x}}\hat{x} \tag{4.10}$$

E aconselhável segundo Lowe que se rejeitem valores de $\left|D(\hat{x})\right|$ inferiores a um determinado limiar. Em Brown e Lowe (2002), é aconselhado trabalhar com o valor 0.03 para esse limiar (assumindo-se que os tons de cinza dos *pixels* da imagem estejam normalizados em valores entre 0 e 1).

Além do procedimento apresentado para se descartar pontos, Lowe ainda aponta que a função DoG possui resposta "forte" ao longo de arestas, mesmo que a localização ao longo da borda seja mal determinada, i.e., pontos em arestas poderiam ser escolhidos como pontos de interesse, o que não é desejável. Mas estes pontos podem ser detectados e eliminados, como discutidos a seguir.

A eliminação de pontos chaves próximos de arestas é feita usando-se uma matriz Hessiana *2x2*, *H*, computada na localização e escala dos pontos-chaves na função *D*.

$$H(x, y) = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{4.11}$$

onde D_{xy} é a derivada de $D(x,y,\sigma)$ na localização e escala em relação a x e y; D_{xx} é a derivada segunda em relação a x; e D_{yy} é a derivada segunda em relação a y.

A Hessiana representa assim uma segunda derivada, permitindo mensurar as magnitudes das curvaturas de *D* a partir de seus autovalores.

As derivadas são estimadas através das diferenças entre pontos vizinhos à localização e escala definida, e pode ser aproximada por

$$D_{xx} = D(x+1, y, \sigma) - 2D(x, y, \sigma) + D(x-1, y, \sigma) \tag{4.12}$$

$$D_{yy} = D(x, y+1, \sigma) - 2D(x, y, \sigma) + D(x, y-1, \sigma) \tag{4.13}$$

$$D_{xy} = \begin{pmatrix} D(x-1, y+1, \sigma) - D(x+1, y+1, \sigma) \\ +D(x+1, y-1, \sigma) - D(x-1, y-1, \sigma) \end{pmatrix} / 4 \tag{4.14}$$

Determina-se α, o autovalor com maior magnitude, e β, o de menor. Pode-se, então, calcular a soma dos autovalores pelo traço de *H* e o produto pelo seu determinante:

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta \tag{4.15}$$

$$Det(H) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \tag{4.16}$$

Para o caso em que o determinante for negativo, as curvaturas possuem sinais diferentes, e o ponto é descartado, não sendo considerado um extremo. Sendo *r* a razão entre o autovalor de maior magnitude e o de menor, de modo que $\alpha = r\beta$, então

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \tag{4.17}$$

A equação (4.17) depende apenas da razão entre os autovalores, sendo independente de seus valores individuais. O valor de $(r+1)/r$ oferece uma medida

de quanto os autovalores são distintos, ou seja, é mínimo quando são idênticos e cresce com respeito ao valor de *r*. Assim, eliminam-se pontos (indesejáveis) próximos a extremidades descartando-se pontos abaixo de determinado limiar (*r*) :

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r} \tag{4.18}$$

A equação (4.18) é altamente eficiente de ser computada. Lowe propõe o uso de *r = 10*, assim eliminam-se pontos chaves que não são estáveis, apesar de estarem próximos de extremidades.

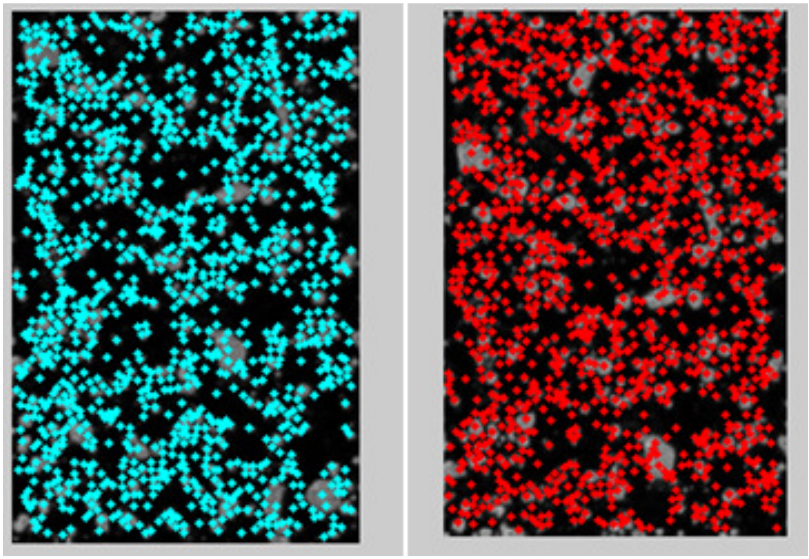

Figura 4.4 Pontos-chaves localizados em duas imagens, antes e após uma deformação trativa na direção vertical.

4.1.3. Atribuição da Orientação dos Descritores

A cada ponto-chave é atribuída uma orientação, que será utilizada mais adiante para se construir descritores invariantes quanto à rotação. Essa invariância é obtida através das características locais da imagem.

Calcula-se para cada amostragem da imagem na escala, $L(x,y,\sigma)$, a magnitude $m(x,y)$ e orientação $\theta(x,y)$ do gradiente usando as diferenças de *pixels*:

$$m(x,y) = \sqrt{\begin{pmatrix} (L(x+1,y) - L(x-1,y))^2 + \\ (L(x,y+1) - L(x,y-1))^2 \end{pmatrix}} \tag{4.19}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{(L(x,y+1) - L(x,y-1))}{(L(x+1,y) - L(x-1,y))}\right) \tag{4.20}$$

Monta-se um histograma das orientações para *pixels* em uma região vizinha ao redor do ponto-chave. O histograma possui 36 regiões, cobrindo todas as orientações possíveis (*0* a 2π), vide Figura 4.5 (Lowe, 2004).

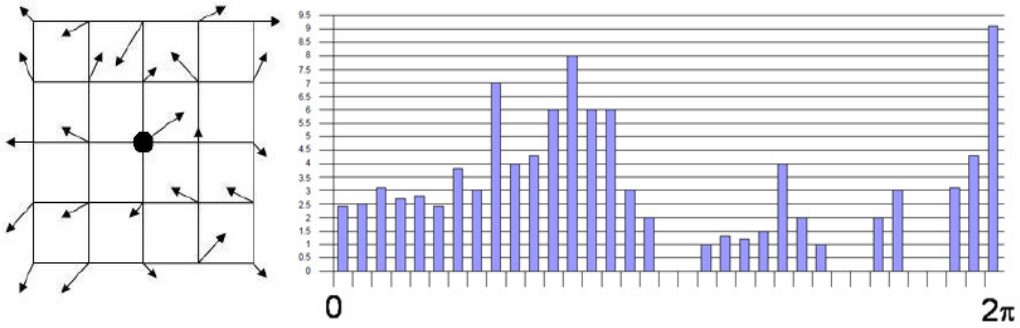

Figura 4.5 Histograma de orientações de um ponto-chave.

Cada ponto na vizinhança do ponto-chave é adicionado ao histograma com um valor de peso determinado. O primeiro peso é o valor da magnitude *m(x,y)* de cada ponto adicionado. O segundo peso é dado por uma janela Gaussiana circular com σ' igual a 1,5 vezes maior que a escala do ponto-chave. Esta janela é definida pela equação Gaussiana

$$g(\Delta x, \Delta y, \sigma') = \frac{1}{2\pi\sigma'^2} e^{-(\Delta x^2+\Delta y^2)/2\sigma'^2} \tag{4.21}$$

onde Δx e Δy são as distâncias entre cada ponto verificado e o ponto-chave.

O valor dos pesos calculados para cada ponto na vizinhança em *(x, y)* é atualizado na expressão:

$$h'_\theta = h_\theta + \alpha m(x, y).g(\Delta x, \Delta y, \sigma') \tag{4.22}$$

com

$$\alpha = \begin{cases} d/i, d < i \\ 0, d > i \end{cases}$$

onde h'_θ é a atualização de h_θ, e d é a distância absoluta em graus entre a orientação do ponto e o θ discretizado, e i é o intervalo em graus entre os θ's discretizados.

Picos no histograma de orientações correspondem a direções dominantes dos gradientes locais. Além do máximo são considerados também os picos que

correspondem a pelo menos 80% do valor deste máximo. Portanto, um mesmo ponto chave poderá ter mais de uma orientação associada.

O pico deste histograma é utilizado para definir a sua orientação. No caso de múltiplos picos de elevada amplitude, o ponto-chave receberá múltiplas orientações, tornando-se ainda mais estável para futura identificação. Ao final, uma parábola é usada para interpolar os três valores do histograma mais próximos ao pico, de forma a se obter uma melhor exatidão na sua posição.

A Figura 4.6 apresenta diversos pontos-chaves identificados em uma imagem de uma superfície metálica, cujas magnitudes e orientação são representadas por vetores.

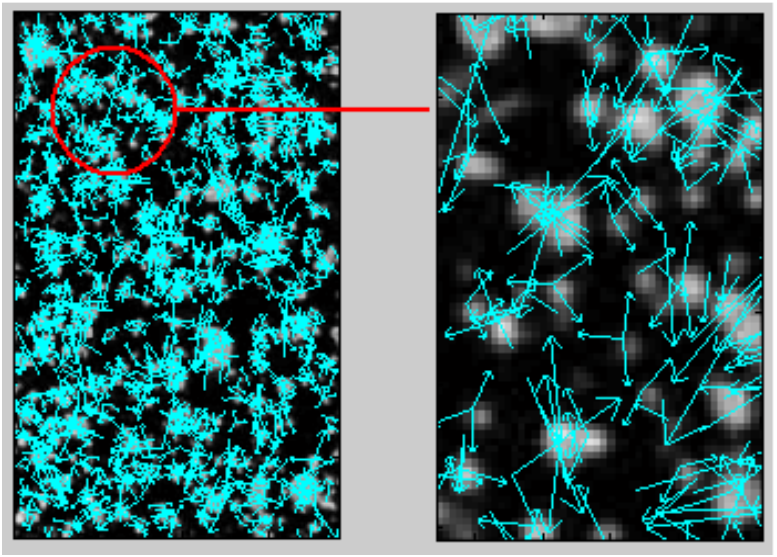

Figura 4.6 Atribuição de orientação e magnitude a cada ponto-chave.

Cada ponto-chave tem agora quatro dimensões:

- posição *x* e *y*;
- magnitude;
- orientação.

4.1.4.Construção do Descritor Local

Nesta seção, será atribuído a cada ponto-chave um descritor invariante a iluminação e ponto de vista 3D, tornando-os bem distinguíveis. É importante lembrar que os procedimentos a seguir serão feitos com os valores normalizados

em relação à orientação e magnitude de gradiente definidos na seção anterior para cada ponto-chave.

Para que os descritores tenham invariância à rotação, as orientações dos gradientes destes pontos são giradas de um ângulo correspondente à orientação do ponto-chave definida na seção anterior.

O descritor do ponto-chave é então criado computando-se as magnitudes e orientações dos gradientes que são amostradas ao redor da localização do ponto-chave. Este procedimento está mostrado na Figura 4.7, onde os gradientes são representados pelas pequenas setas em cada amostra da localização. São definidas *nxn* regiões de amostragem com *kxk pixels* cada ao redor da localização do ponto-chave.

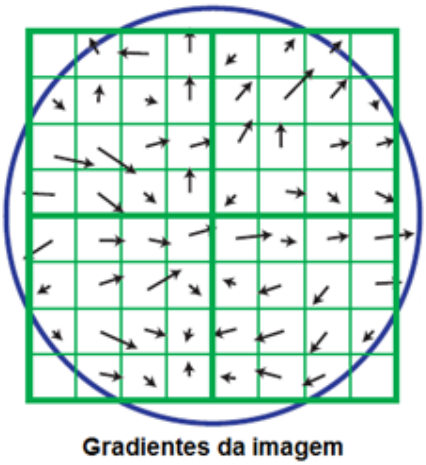

Figura 4.7 Mapa de gradientes para $n = 2$ regiões e $k = 4$ *pixels*. (Lowe, 2004).

Uma função Gaussiana é utilizada para dar peso à magnitude do gradiente em cada ponto na vizinhança do ponto-chave, com uma janela de suavização Gaussiana de escala σ igual à metade da largura da janela do descritor. Esse Gaussiano evita mudanças súbitas do descritor a pequenas mudanças na posição da janela, e também reduz a ênfase nos gradientes longe do centro do descritor, que são mais afetados por erros.

Uma vez efetuada a suavização dos gradientes, o descritor consiste de um vetor contendo os valores do histograma. No exemplo da Figura 4.8, o histograma tem 8 valores de orientação, cada um criado ao longo de uma janela de apoio de *4x4 pixels*. O vetor característico resultante tem 128 elementos com uma janela de apoio total de *16x16 pixels*.

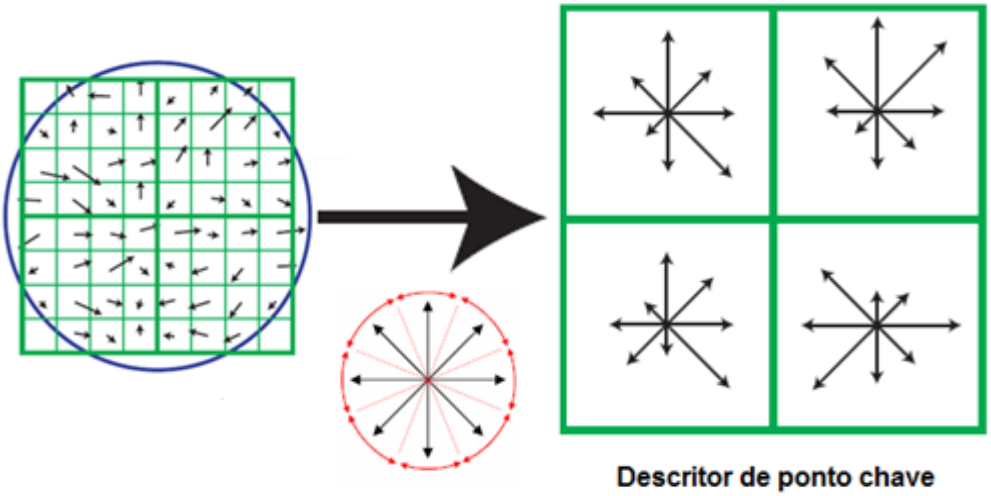

Figura 4.8 Construção do descritor para um ponto-chave de *2x2* com 48 elementos (Lowe, 2004).

No entanto, duas imagens de um mesmo objeto podem possuir variações de luminosidade que modifiquem sensivelmente os descritores obtidos. Assim, para que o descritor tenha invariância à iluminação, este é normalizado.

Os descritores são invariantes a mudanças homogêneas de brilho da imagem, uma vez que esta variação representa uma adição a todos os *pixels* da imagem de uma constante, e os descritores são calculados por diferenças de *pixels*. Quanto a mudanças homogêneas de contraste, representadas pela multiplicação de todos os *pixels* por uma constante, elas são corrigidas com a normalização dos descritores.

Variações não-lineares, causadas por saturação das câmeras ou por efeito de iluminação de superfícies tridimensionais em diferentes orientações, podem provocar elevada influência sobre as magnitudes dos descritores, mas com pouca influência na orientação. Reduz-se este efeito impondo um valor máximo às magnitudes. Após a normalização, todos os valores acima de um determinado limiar são ajustados para este limiar. Isto é feito para que direções com magnitude muito grande não dominem a representação do descritor. Lowe sugere usar um limiar 0,2. Isto significa que a correspondência para as grandes magnitudes dos gradientes não é tão importante se comparada com a distribuição das orientações.

Para cada imagem, são construídos diversos descritores, cada um referente a um ponto chave. Tem-se como resultado, portanto, um conjunto de descritores robustos que podem ser usados para fazer a correspondência da imagem em outra imagem, como será detalhado na próxima seção. Mais detalhes sobre a construção dos descritores SIFT são encontrados em Lowe (2004).

4.2. *Matching*: Encontrando os Pontos em Comum

A idéia de *matching* é extrair pontos-chave de duas imagens, e procurar os pontos correspondentes em cada uma, como exemplificado na Figura 4.9. A comparação de pontos é baseada na similitude dos descritores correspondentes. A obtenção de uma solução robusta para o problema da busca de pontos homólogos pode ser considerada como um elemento chave na automação das tarefas fotogramétricas (Schenk, 1999). Muitas das aplicações de Visão Computacional exigem a identificação de elementos repetitivos entre duas imagens.

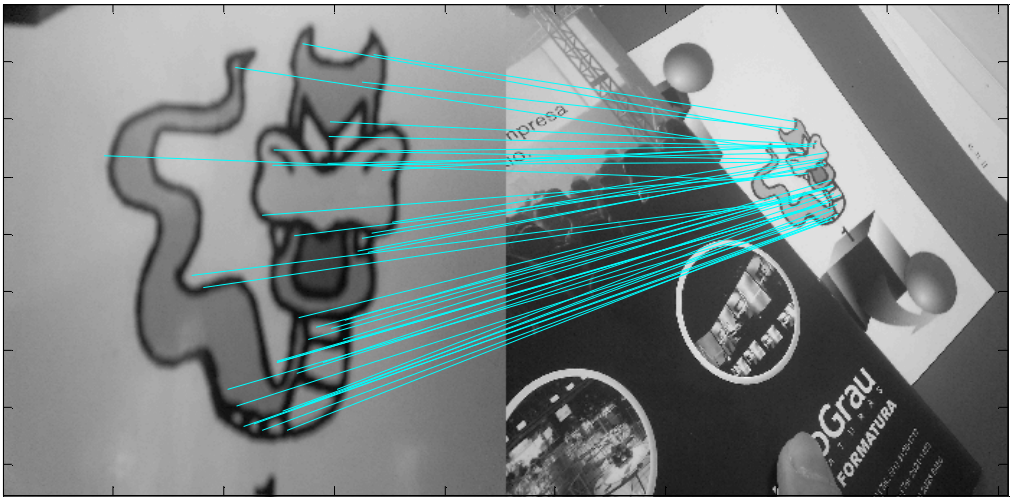

Figura 4.9 Processo de correspondência entre duas imagens através da técnica SIFT.

Quando se trabalha com SIFT, pontos de interesse são detectados pelo método e representados por descritores. Os descritores são vetores que podem ser comparados, por exemplo, utilizando-s a distância Euclidiana. Normalmente, os candidatos à melhor correspondência são pontos próximos, de maneira que o melhor candidato é o ponto que apresenta a menor distância Euclidiana.

Lowe utilizou uma modificação do algoritmo Árvore k-d chamado de método de *Best-Bin-First* (BBF) (Beis & Lowe, 1997), que pode identificar os

vizinhos mais próximos com elevada probabilidade, utilizando apenas uma quantidade limitada de esforço computacional.

O problema da correspondência, desta maneira, é reduzido à busca do vizinho mais próximo. No entanto, alguns pontos instáveis (*outliers*) 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 por um limiar (*threshold*) (Lowe, 2004). Lowe rejeitou todas as correspondências (*matches*) em que a relação de distância é superior a 0.8, o que elimina 90% das falsas correlações, porém apenas descartando menos de 5% das correspondências corretas. Portanto, as correspondências são assim eficientemente refinadas, e os falsos pares são descartados.

No próximo capítulo, a técnica SIFT é aplicada ao problema de medição de campos de deformação em componentes mecânicos.