2
Visão Computacional

Este capítulo descreve tópicos relacionados à visão computacional subjacentes ao presente trabalho.

A primeira seção apresenta como determinar parâmetros intrínsecos e extrínsecos da câmera e consequentemente a transformada entre a posição da câmera e do objeto, a partir de pares de pontos no espaço e na imagem.

A segunda seção apresenta uma técnica muito conhecida na área de visão computacional chamada SIFT. Este algoritmo é capaz de gerar pares de pontos correspondentes entre duas imagens de diferentes vistas de um mesmo objeto, em diferentes condições de iluminação, rotação ou escala das imagens.

2.1.
Processamento da imagem

Um sistema de visão estereoscópica baseia-se na relação entre as coordenadas de pontos no espaço tridimensional e as coordenadas das projeções destes pontos na imagem. Normalmente estas equações referem-se ao sistema de coordenadas da câmera, sendo as projeções dos pontos na imagem dadas pelas coordenadas linha e coluna em número de *pixels*.

No modelo conhecido como câmera de orifício, apresentado na Figura 2, denomina-se distância focal f a distância entre o plano da imagem e o centro O de coordenadas da câmera. A linha que passa por O e é perpendicular ao plano da imagem é conhecida como eixo ótico. E o ponto onde ocorre à interseção é chamado de ponto principal ou centro da imagem o (Trucco & Verri, 1998).

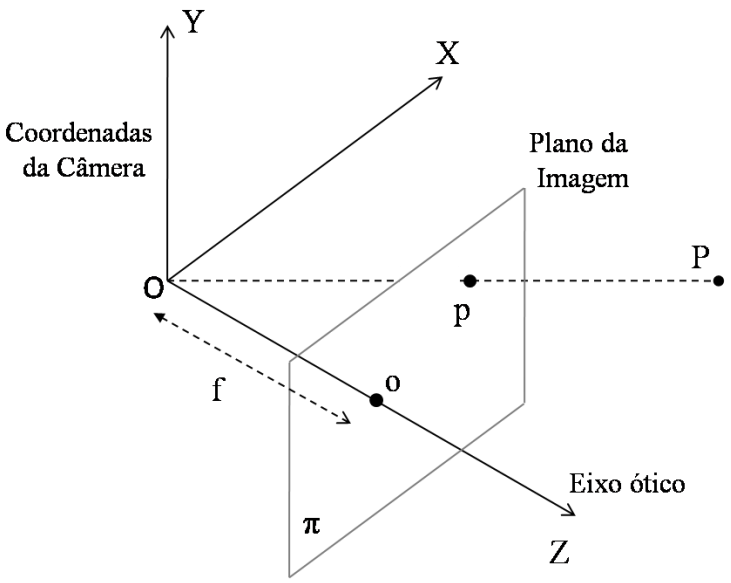

Figura 2 - Modelo de Projeção da Câmera

As equações básicas do sistema de coordenadas da câmera são:

$$x = f\frac{X}{Z} \tag{1}$$

$$y = f\frac{Y}{Z} \tag{2}$$

sendo (x, y) as coordenadas do ponto na imagem, e (X, Y, Z) as coordenadas do ponto no espaço, conhecidas na literatura de visão computacional como coordenadas no mundo.

A relação entre as coordenadas de um ponto no mundo e as coordenadas de sua projeção na imagem, envolve dois grupos de parâmetros chamados parâmetros extrínsecos e intrínsecos.

Os parâmetros extrínsecos definem a localização e orientação de um ponto no sistema de coordenadas da câmera em relação ao mundo. Já os parâmetros intrínsecos são necessários para relacionar o ponto nas coordenadas do *pixel* com o sistema de coordenadas da câmera.

2.1.1. Parâmetros extrínsecos

Os parâmetros extrínsecos definem a transformação geométrica que relaciona unicamente o sistema de coordenadas da câmera e o sistema de coordenadas do mundo. São eles:

- um vetor de translação 3D, T, que descreve a posição relativa entre as origens dos dois sistemas,
- uma matriz ortogonal 3x3 R, que expressa a rotação relativa entre os eixos de um e outro sistema de coordenadas.

Dessa forma, a relação entre o sistema de coordenadas do ponto P no mundo P^m e na câmera P^c é:

$$P^c = R\,P^m + T \tag{3}$$

na qual

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \tag{4}$$

e

$$T = \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} \tag{5}$$

como é ilustrado na Figura 3.

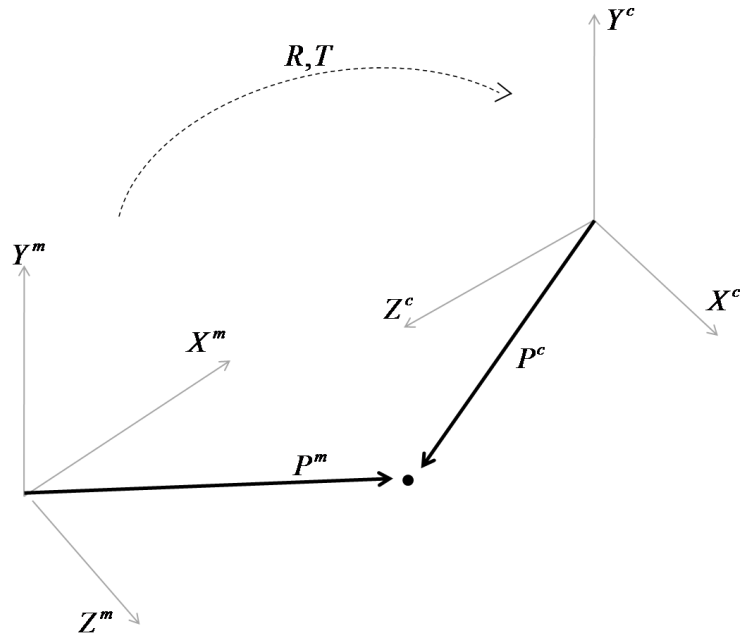

Figura 3 - Transformação do sistema de coordenadas do mundo para sistema de coordenadas da câmera

2.1.2.
Parâmetros intrínsecos

Os parâmetros intrínsecos caracterizam as influências óticas, geométricas e digitais da câmera na imagem. São eles:

- a distância focal;
- as dimensões horizontal e vertical dos pixels na matriz de sensores;
- as coordenadas do vetor que representa deslocamento da origem do sistema de coordenadas da imagem relativamente ao ponto em que o eixo ótico atinge o plano de projeção;
- coeficientes relativos à distorção geométrica causada pela ótica.

Ao determinar a relação de um ponto no sistema de coordenadas da imagem em pixel $\left(x^{im}, y^{im}\right)$ e do mesmo ponto no sistema de coordenadas da câmera $\left(x^{c}, y^{c}\right)$, tem-se que:

$$x^{c} = -\left(x^{im} - o_x\right)s_x \tag{6}$$

$$y^{c} = -\left(y^{im} - o_y\right)s_y \tag{7}$$

onde $\left(o_x, o_y\right)$ são as coordenadas em *pixels* do centro da imagem e $\left(s_x, s_y\right)$ o tamanho efetivo do pixel (em milímetros) nas direções horizontal e vertical.

Na maioria dos casos, as influências óticas podem ser modeladas simplesmente como distorções radiais, da seguinte forma:

$$x^{c} = x_{dis}\left(1 + k_1 r^2 + k_2 r^4\right) \tag{8}$$

$$y^{c} = y_{dis}\left(1 + k_1 r^2 + k_2 r^4\right) \tag{9}$$

sendo $\left(x_{dis}, y_{dis}\right)$ as coordenadas distorcidas dos pontos e $r^2 = x_{dis}^2 + y_{dis}^2$ (Trucco & Verri, 1998). Pode-se observar nas equações (8) e (9) que a distorção no centro da imagem é sempre nula, e cresce conforme o afastamento do centro. Estas distorções podem ser significativas dependendo da câmera utilizada, no entanto são desprezadas no presente trabalho.

2.1.3.
Modelo da câmera

Uma vez determinados os valores dos parâmetros intrínsecos e extrínsecos, pode-se relacionar diretamente o sistema de coordenadas da imagem com o do mundo, sem que seja necessário explicitar o sistema de coordenadas da câmera. Desta forma, substituindo as equações (4) a (7) em (3), tem-se que:

$$x^{im} - o_x = -f_x \frac{r_{11}X^m + r_{12}Y^m + r_{13}Z^m + T_x}{r_{31}X^m + r_{32}Y^m + r_{33}Z^m + T_z} \tag{10}$$

$$y^{im} - o_y = -f_y \frac{r_{21}X^m + r_{22}Y^m + r_{23}Z^m + T_y}{r_{31}X^m + r_{32}Y^m + r_{33}Z^m + T_z} \tag{11}$$

onde $f_x = f / s_x$ e $f_y = f / s_y$.

Observando que as equações (10) e (11) possuem o mesmo denominador, pode-se assumir para cada par de pontos $\left(\left(X_i^m, Y_i^m, Z_i^m\right), \left(x_i^{im}, y_i^{im}\right)\right)$ a equação

$$\begin{aligned} &\left(x_i^{im} - o_x\right) f_y \left(r_{21}X_i^m + r_{22}Y_i^m + r_{23}Z_i^m + T_y\right) = \\ &\left(y_i^{im} - o_y\right) f_x \left(r_{11}X_i^m + r_{12}Y_i^m + r_{13}Z_i^m + T_x\right) \end{aligned} \tag{12}$$

Generalizando, podem-se considerar as coordenadas transladadas $\left(x_i^{im} - o_x, y_i^{im} - o_y\right) = \left(x_i^{im}, y_i^{im}\right)$, considerando que a origem das coordenadas da imagem está no ponto (0,0). Considerando também $\alpha = fx / fy$, a equação (12) pode ser representada como uma equação linear de 8 coeficientes desconhecidos:

$$\begin{aligned} &x_i^{im} X_i^m v_1 + x_i^{im} Y_i^m v_2 + x_i^{im} Z_i^m v_3 + x_i^{im} v_4 - \\ &y_i^{im} X_i^m v_5 + y_i^{im} Y_i^m v_6 + y_i^{im} Z_i^m v_7 + y_i^{im} v_8 = 0 \end{aligned} \tag{13}$$

onde

$$\begin{aligned} v_1 &= r_{21} & v_5 &= \alpha\, r_{11} \\ v_2 &= r_{22} & v_6 &= \alpha\, r_{12} \\ v_3 &= r_{23} & v_7 &= \alpha\, r_{13} \\ v_4 &= T_y & v_8 &= \alpha\, T_x. \end{aligned}$$

2.1.4.
Calibração da Câmera

O processo de determinar os valores dos parâmetros do modelo da câmera é chamado de calibração da câmera. Escrevendo a equação (13) para N pontos cujas coordenadas num referencial do mundo $\left[X_i^m, Y_i^m, Z_i^m\right]$ e suas projeções $\left[x_i^{im}, y_i^{im}\right]$

na imagem são conhecidas, tem-se um sistema homogêneo de N equações lineares como abaixo:

$$Av = 0 \tag{14}$$

onde $v = [v_1,...,v_8]^T$ e a matriz A de dimensão Nx8 é dada por:

$$A = \begin{bmatrix} x_1^{im}X_1^m & x_1^{im}Y_1^m & x_1^{im}Z_1^m & x_1^{im} & -y_1^{im}X_1^m & -y_1^{im}Y_1^m & -y_1^{im}Z_1^m & -y_1^{im} \\ x_2^{im}X_2^m & x_2^{im}Y_2^m & x_2^{im}Z_2^m & x_2^{im} & -y_2^{im}X_2^m & -y_2^{im}Y_2^m & -y_2^{im}Z_2^m & -y_2^{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^{im}X_N^m & x_N^{im}Y_N^m & x_N^{im}Z_N^m & x_N^{im} & -y_N^{im}X_N^m & -y_N^{im}Y_N^m & -y_N^{im}Z_N^m & -y_N^{im} \end{bmatrix} \tag{15}$$

A solução do sistema linear homogêneo da equação (14) é dada pelo auto-vetor correspondente ao menor autovalor de A^TA (Forsyth e Ponce, 2003). Considerando que todos os parâmetros da câmera possuem um fator de escala γ, tem-se que $v = \bar{v}$, e que:

$$\bar{v} = \gamma(r_{21}, r_{22}, r_{23}, T_y, \alpha r_{11}, \alpha r_{12}, \alpha r_{13}, \alpha T_x) \tag{16}$$

Pelas propriedades da matriz de rotação, sabe-se que $r_{21}^2 + r_{22}^2 + r_{23}^2 = 1$, e a partir disso pode-se concluir que:

$$\sqrt{\bar{v}_1^2 + \bar{v}_2^2 + \bar{v}_3^2} = \sqrt{\gamma^2\left(r_{21}^2 + r_{22}^2 + r_{23}^2\right)} = |\gamma| \tag{17}$$

De forma similar, assumindo que $r_{11}^2 + r_{12}^2 + r_{13}^2 = 1$ e que $\alpha > 0$, tem-se que:

$$\sqrt{\bar{v}_5^2 + \bar{v}_6^2 + \bar{v}_7^2} = \sqrt{\gamma^2\alpha^2\left(r_{11}^2 + r_{12}^2 + r_{13}^2\right)} = \alpha|\gamma| \tag{18}$$

Nesta etapa, já foram determinadas as duas primeiras linhas da matriz de rotação, as componentes x e y do vetor de translação, bem como o fator de escala γ e o fator de forma α. Sabendo que a terceira linha da matriz de rotação pode ser determinada pelo produto vetorial entre as duas primeiras linhas, resta apenas determinar a componente z do vetor de translação e a distância focal na direção horizontal f_x.

Para determinar estes últimos parâmetros, retorna-se às equações (10) ou (11) e obtém-se a solução do sistema minimizando o erro quadrático de:

$$A\begin{pmatrix} T_z \\ f_x \end{pmatrix} = b \tag{19}$$

.

Utilizando os mesmos N pares de pontos, tem-se

$$A = \begin{bmatrix} x_1^{im} & \left(r_{11}X_1^m + r_{12}Y_1^m + r_{13}Z_1^m + T_x\right) \\ x_2^{im} & \left(r_{11}X_2^m + r_{12}Y_2^m + r_{13}Z_2^m + T_x\right) \\ \vdots & \vdots \\ x_N^{im} & \left(r_{11}X_N^m + r_{12}Y_N^m + r_{13}Z_N^m + T_x\right) \end{bmatrix} \quad (20)$$

e

$$b = \begin{pmatrix} -x_1^{im}\left(r_{31}X_1^m + r_{32}Y_1^m + r_{33}Z_1^m\right) \\ -x_2^{im}\left(r_{31}X_2^m + r_{32}Y_2^m + r_{33}Z_2^m\right) \\ \vdots \\ -x_N^{im}\left(r_{31}X_N^m + r_{32}Y_N^m + r_{33}Z_N^m\right) \end{pmatrix} \quad (21)$$

Desta forma, determinam-se todos os parâmetros intrínsecos e extrínsecos da câmera: matriz de rotação R, vetor de translação T, distância focal nas direções horizontal e vertical f_x e f_y, e fator de forma α.

2.2. SIFT (*Scale Invariant Feature Transform*)

O método SIFT (Lowe, 2004) tem como principal objetivo a extração de feições invariantes, chamados pontos chave, em uma imagem, podendo ser utilizado para estabelecer correspondências entre diferentes vistas de um objeto ou de uma cena. Estas feições são invariantes quanto à escala e à rotação da imagem, e são robustas contra distorção, mudanças no ponto de vista do ponto, ruídos na imagem, e variações na iluminação da cena.

Um exemplo do resultado do método SIFT aplicado a uma imagem pode ser visto na Figura 4, onde em azul estão indicados os pontos-chaves determinados pelo algoritmo, bem como a "orientação" do ponto de acordo com o algoritmo.

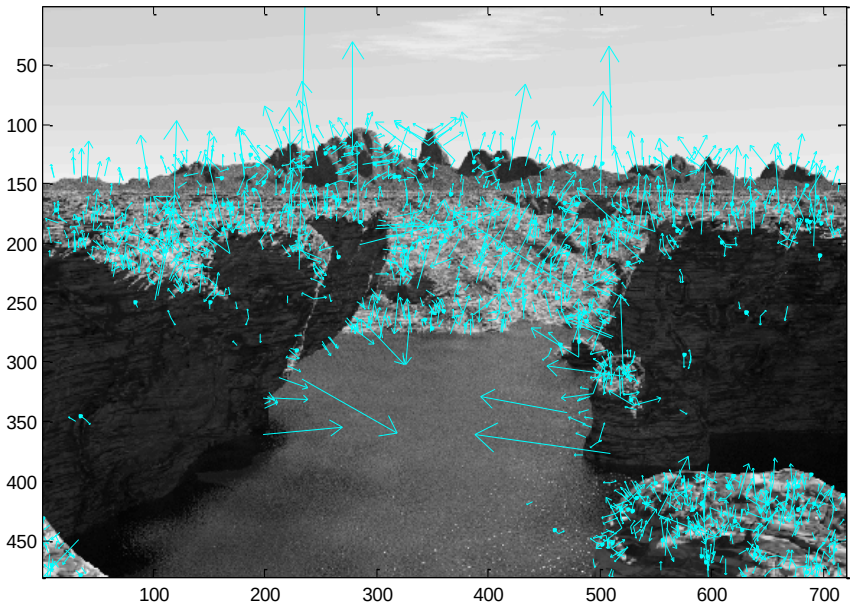

Figura 4 - Resultado do método SIFT aplicado a uma imagem

Apresenta-se a seguir uma descrição sucinta do algoritmo SIFT que consiste de 5 passos sequenciais:

a. Detecção dos pontos chave;
b. Eliminação dos pontos chave "fracos";
c. Determinação da orientação dos pontos chave;
d. Cálculo dos descritores dos pontos chave;
e. Pareamento.

2.2.1. Detecção dos pontos chave

O método tem como entrada uma imagem inicial $I(x, y)$ a partir da qual são criadas diversas imagens em diferentes escalas. A cada uma das imagens produzidas é aplicado um filtro gaussiano $G(x, y, \sigma)$ de suavização com diferentes valores de desvio padrão σ, criando-se assim imagens em vários níveis ou escalas, ou seja

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{22}$$

onde $*$ representa a operação de convolução em x e y, $L(x, y, \sigma)$ representa a imagem suavizada pela gaussiana, e

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2} \quad (23)$$

Alcançando um certo nível ocorre uma sub amostragem da imagem que reduz suas dimensões à metade. Cada conjunto de imagens de mesmas dimensões formam uma oitava.

Em seguida imagens de níveis adjacentes são subtraídas para produzir as DOG's (Diferença-de-Gaussianas):

$$D(x,y,\sigma)=(G(x,y,k\sigma)-G(x,y,\sigma))*I(x,y) \quad (24)$$

$$=L(x,y,k\sigma)-L(x,y,\sigma) \quad (25)$$

onde k é o fator de suavização aplicado às gaussianas.

Na assim chamada pirâmide DOG (apresentadas na Figura 5) procuram-se os pontos extremos (máximos ou mínimos) alongo das 3 dimensões (x,y,σ). Cada ponto da pirâmide DOG é comparado com seus 8 vizinhos na mesma escala, além dos 9 pontos vizinhos nas escalas acima e abaixo, conforme mostra a Figura 6. O ponto só é selecionado como ponto chave se seu valor na pirâmide DOG for maior (ou menor) do que o de todos os seus vizinhos.

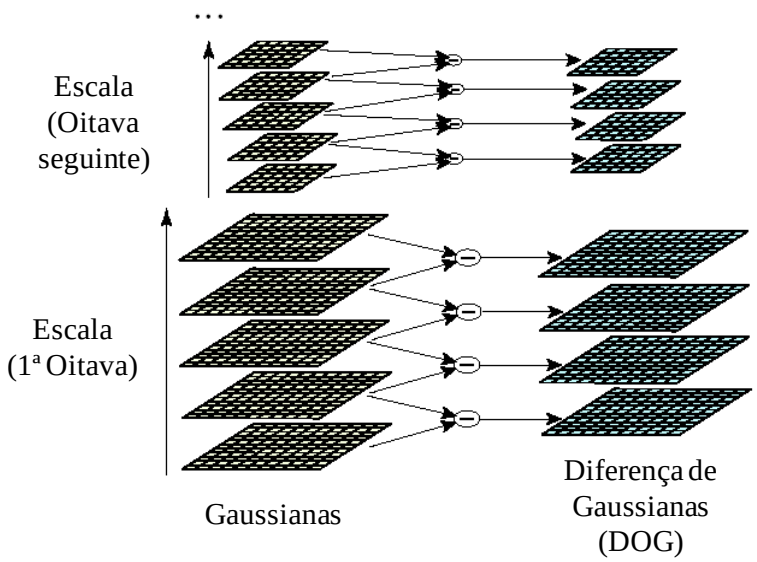

Figura 5 - Gaussianas aplicadas a cada oitava, e a partir de suas subtrações surgem as DOG's (Diferença-de-Gaussianas)

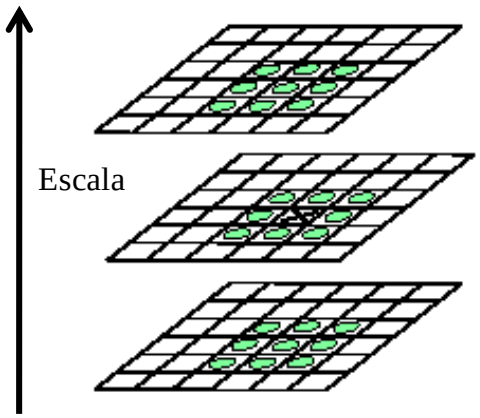

Figura 6 - Detecção dos máximos e mínimos nas diferenças entre gaussianas (DOG's)

2.2.2. Eliminação dos pontos chave "fracos"

Nesta etapa é feita uma filtragem dos pontos chave considerados "fracos", ou seja, que estão sobre arestas e/ou cuja vizinhança tem baixo contraste. A condição do contraste baseia-se no valor na pirâmide DOG sobre o ponto. A condição quanto a localizar-se sobre arestas é verificada utilizando um algoritmo que tem a mesma base teórica dos algoritmos de detecção de cantos já mencionados (Harris e Stephens, 1988).

2.2.3. Determinação da orientação dos pontos chave

Para cada ponto resultante da etapa anterior, calculam-se a magnitude e a orientação do gradiente em cada posição pertencente à vizinhança em torno do ponto, aplicando-se as equações abaixo:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \qquad (26)$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \qquad (27)$$

Monta-se a partir destes valores um histograma de orientações, contendo 36 faixas, cobrindo os 360°. Cada ponto é acrescido ao histograma amortecido por uma gaussiana circular. Picos no histograma de orientação representam a direção dominante do gradiente local. O maior pico de orientação no histograma é encontrado juntamente com quaisquer outros picos 80% do pico mais alto, e são

utilizados para criar *keypoints* com sua orientação. Mais de uma orientação pode ser associada a um *keypoint*. Sendo assim, cada *keypoint* possui 4 dimensões: localização em x, localização em y, escala e orientação.

2.2.4. Cálculo dos descritores dos pontos chave

A Figura 7 ilustra o cálculo do descritor do *keypoint*. Primeiramente a magnitude e orientação do gradiente são amostradas ao redor da localização do *keypoint*. A fim de garantir invariância à orientação, as coordenadas do descritor e a orientação do gradiente são rotacionadas de acordo com a orientação do *keypoint*. Os gradientes são representados no lado esquerdo da Figura 7, por setas indicando assim sua orientação.

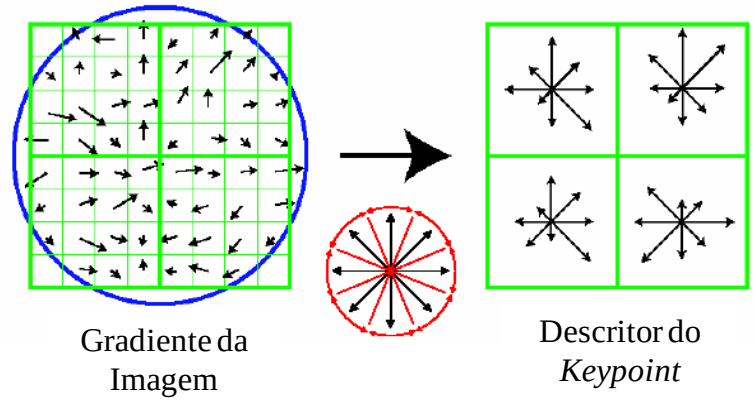

Figura 7 - Orientação dos pontos da vizinhança (esquerda), e descritor do *keypoint* (direita)

O descritor do *keypoint* pode ser observado no lado direito da Figura 7. Cada quadrante do descritor contém a soma dos gradientes onde as setas, em 8 diferentes direções, contém o comprimento equivalente a magnitude do histograma.

A fim de evitar influências nas fronteiras do descritor, aplica-se um peso maior aos pontos centrais. Este peso será igual a $1-d$, sendo d a distância entre o ponto e o centro (horizontal e vertical) da vizinhança.

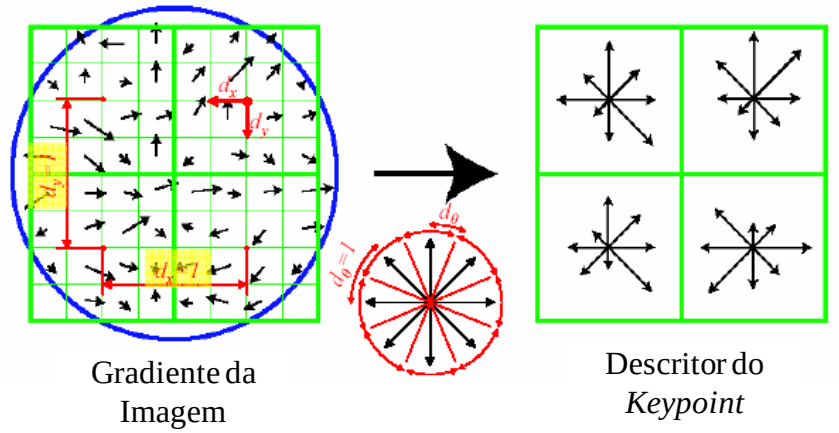

Figura 8 - Descritor dos pontos

Para garantir a invariância à iluminação, o vetor descritor do ponto deve ser normalizado. Suas componentes são limitadas em 0,2, e então o vetor é normalizado novamente.

Assim é criado o descritor de cada *keypoint*. Em seu artigo Lowe apresenta um vetor 2x2 descritor partindo de 8x8 pontos de vizinhança. No entanto, os experimentos realizados tanto em seu artigo, quanto neste trabalho, utilizaram um vetor descritor de tamanho 4x4, admitindo assim uma vizinhança ao redor do ponto de 16x16.

2.2.5. Pareamento

Seguindo todas as etapas mencionadas acima, o método é capaz de determinar os pontos-chaves a partir de uma imagem de entrada, e seus descritores.

Para iniciar o processo de *matching* (determinação dos pares de pontos correspondentes), um par de imagens deve ser submetido ao processo do SIFT, obtendo assim o descritor de cada imagem.

A medida de correlação entre dois pontos é dada pela distância euclidiana entre seus descritores. A fim de determinar o correspondente de um ponto da primeira imagem na segunda, cada ponto do primeiro descritor é comparado com todos os pontos do segundo descritor, e o que apresentar a menor distância será considerado. Para evitar correlações falsas, se a diferença entre os dois primeiros pontos com menor distância for maior que 80%, os pontos são descartados.

Na Figura 9 pode-se observar um exemplo de resultado do método de *matching*.

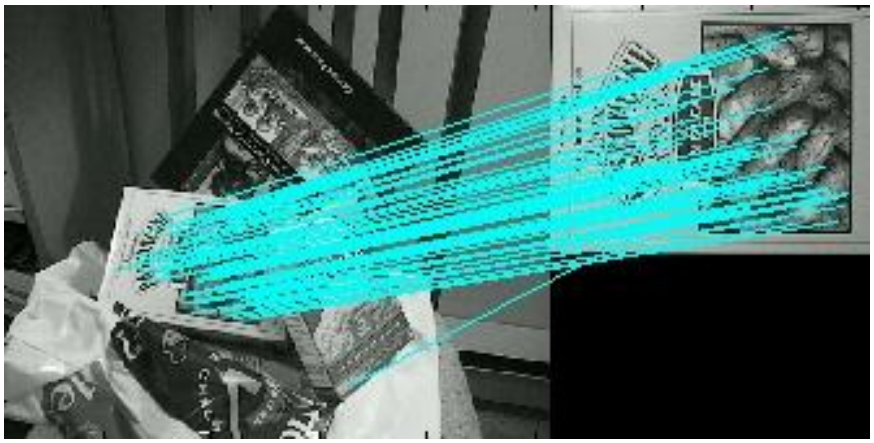

Figura 9 - Resultado do algoritmo de detecção de pontos correspondentes (Lowe, 2004)