Produzindo mapa de profundidade esparso com câmera monocular

Naiane Maria de Sousa[1], **Gabriel da Silva Vieira**[1]

[1]Laboratório de Visão Computacional – Núcleo de informática
Instituto Federal Goiano - Campus Urutaí
Urutaí – GO – Brasil

naimsousa1@gmail.com, gabriel.vieira@ifgoiano.edu.br

Abstract. *Depth maps aid in the three-dimensional reconstruction of a scene. In order to achieve an accurate identification of all observed points are needed several requirements, among them the camera calibration. In this process, intrinsic camera data are estimated in order to define transformation values of 3D space in a 2D image. This estimation is important because, along with other processes, such as stereo correspondence and triangulation, it contributes to the definition of collinearity between points in a scene. This paper shows a method to build a sparce depth map using a monocular camera.*

Resumo. *Mapas de profundidade auxiliam na reconstrução tridimensional de uma cena. Para se chegar a uma identificação precisa de todos os pontos observados, são necessários vários quesitos, dentre eles, processos bem estruturados como calibração de câmera. Nesse processo, dados intrínsecos da câmera são estimados a fim de definir valores de transformação do espaço 3D em uma imagem 2D. Essa estimativa é importante, pois juntamente com outros processos, como correspondência estéreo e triangulação, contribui para a definição de colinearidade entre pontos de uma cena. Esse artigo apresenta um método para construção de mapa de profundidade esparso com uso de câmera monocular.*

1. Introdução

Imagens estéreo são aquelas que permitem a sensação de profundidade. Essa impressão se dá a partir da análise de, pelo menos, duas imagens de uma cena obtidas de ângulos e locais diferentes simulando a visão humana [Laureano and Paiva 2013]. As técnicas de visão estéreo podem ser utilizadas, por exemplo, em aplicações de navegação robótica (possibilitando noções de profundidade à máquina) ou para mobilidade (auxiliando deficientes visuais na detecção de obstáculos) [Vieira et al. 2016].

Mapas de profundidade podem ser utilizados para melhorar a percepção de profundidade em uma cena capturada. Entretanto, um mapa de profundidade preciso é difícil de ser obtido [Jung 2013]. Na busca de maior precisão, há o estudo de calibração de câmeras e retificação de imagens, processos importantes para se estimar a colinearidade entre pontos.

Nesse artigo é apresentado um método de construção de mapas de profundidade esparso [1] com a utilização de uma câmera comum, monocular, em movimento. O artigo

[1]Mapa esparso é aquele construído levando em conta apenas a disparidade entre alguns pontos da imagem e não toda a informação, em contraposição ao mapa denso [Batista and Regis 2013].

está organizado da seguinte forma: na Seção 2 os processos que fundamentam esse estudo são discutidos individualmente; na Seção 3 os processos de reconstrução são reunidos, aplicados e os resultados são apresentados e discutidos; por fim, na Seção 4 as conclusões desse estudo são apontadas.

2. Material e Métodos

Os processos investigados nesse estudo iniciam com a calibração de câmera e finalizam com a representação tridimensional de pontos contidos nas imagens utilizadas. A seleção desses processos foi feita de acordo com os trabalhos de Hartley e Zisserman, [Hartley and Zisserman 2003], Yi Ma *et al.*, [Ma et al. 2012], Radke, [Radke 2013] e Furukawa, [Furukawa and Hernández 2015], e são discutidos a seguir.

2.1. Calibração de câmera

O processo de calibração consiste em estimar os parâmetros internos e externos de câmera, ou parâmetros intrínsecos e extrínsecos, além de coeficientes de distorção. Essas informações são utilizadas como meio para medir o tamanho de objetos, determinar a localização da câmera na cena, ou mesmo, corrigir distorções de lentes.

Se a câmera não está centrada no centro ótico, pode-se obter uma informação de translação adicional, o_x e o_y, se o *pixel* não está em escala unitária tem-se uma escala adicional nas direções dos eixos x e y por s_x e s_y e se os *pixels* não forem retangulares haverá um fator de inclinação (*skew*), s_θ.

$$\lambda \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} s_x & s_\theta & o_x \\ 0 & s_y & o_y \\ 0 & 0 & 1 \end{pmatrix}}_{\equiv K_s} \underbrace{\begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\equiv K_f} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{\equiv \Pi_o} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{1}$$

Depois da projeção de perspectiva Π_o (com comprimento focal 1), tem-se uma transformação adicional que depende dos parâmetros de câmera, que podem ser expressos pela matriz de parâmetros intrínsecos $K = K_s K_f$. Dessa forma, todos os parâmetros intrínsecos da câmera são agrupados na matriz:

$$K \equiv K_s K_f = \begin{pmatrix} f s_x & f s_\theta & o_x \\ 0 & f s_y & o_y \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

Como a transformação em coordenadas de imagem é uma função em relação a coordenadas de mundo X_o, tem-se:

$$\lambda x' = K \Pi_o X = K \Pi_o g X_o \equiv \Pi X_o, \tag{3}$$

onde g é a representação dos parâmetros extrínsecos (rotação e translação) em coordenadas homogêneas.

2.2. Caracterização de pontos

Nesse processo, modelos matemáticos são aplicados a fim de garantir *repetibilidade*, isto é, dada uma imagem diferente da mesma cena deve ser possível encontrar novamente o local correto da característica identificada inicialmente.

Um clássico detector de características foi publicado por Harris e Stephens, em 1988. O método proposto pelos autores é baseado no tensor de estrutura (*structure tensor*) que pode incluir um filtro gaussiano G de largura σ.

$$M(x) \equiv G_\sigma * \bigtriangledown I \bigtriangledown I^T = \int G_\sigma(x - x') \begin{pmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{pmatrix} (x')dx' \tag{4}$$

2.3. Correspondência estéreo

Medidas para correlacionar pontos são conhecidas na literatura como foto-consistência e são responsáveis por estimar a probabilidade de dois *pixels* (ou grupo de *pixels*) possuírem correspondência entre si. Um exemplo de medida de foto consistência é o NCC (*Normalized Cross Correlation*), usado, principalmente, quando há variação de iluminação e de materiais [Furukawa and Hernández 2015].

A correlação cruzada normalizada é definida como:

$$NCC(h) = \frac{\int_{w(x)} (I_1(x') - \bar{I}_1)(I_2(h(x')) - \bar{I}_2)}{\sqrt{\int_{w(x)} (I_1(x') - \bar{I}_1)^2 dx' \int_{w(x)} (I_2(x') - \bar{I}_2)^2 dx'}}, \tag{5}$$

onde $\bar{I}_1$ e $\bar{I}_2$ é a média de intensidade em uma janela de vizinhança $W_{(x)}$. Subtraindo essa intensidade média, a medida torna-se invariante a mudanças de intensidade aditivas $I \rightarrow I + \gamma$. Dividindo pela variação de intensidade de cada janela torna a medida invariante a mudanças multiplicativas $I \rightarrow \gamma I$.

2.4. Encapsulamento da geometria epipolar

A geometria epipolar relaciona pares de imagem. Matematicamente, as matrizes fundamental e essencial encapsulam a geometria epipolar e podem ser estimadas usando pontos correspondentes entre imagens.

A restrição epipolar providencia um relacionamento entre coordenadas de ponto 2D de um ponto 3D em cada uma das duas imagens, considerando também os parâmetros de transformação de câmera. A matriz

$$E = \hat{T}R \in \mathbb{R}^{3\times3} \tag{6}$$

é chamada de *matriz essencial*. Essa matriz codifica apenas os parâmetros extrínsecos da câmera e toma como base o conhecimento prévio dos parâmetros intrínsecos. Dessa forma, se os parâmetros intrínsecos da câmera são conhecidos pode-se trabalhar com coordenadas normalizadas $x_1 = K^{-1}x_1$.

A equação 7 mostra a relação entre matriz essencial e matriz fundamental, para $K_1 = K_2 = K$:

$$F = K^{-T} E K^{-1} \tag{7}$$

Com o uso dessas matrizes pode-se encontrar as linhas epipolares com as seguintes equações:

$$l_2 = Fx_1, l_1 = F^T x_2 \tag{8}$$

onde l_2 corresponde a linha epipolar na segunda imagem referente a um ponto na primeira imagem. Já l_1 representa a linha epipolar de um ponto da segunda imagem.

2.5. Triangulação

A triangulação é um processo que lida com o problema de encontrar a posição de um ponto no espaço dado a sua posição em duas ou mais imagens. Esse processo requer a intersecção de duas retas no espaço, que pode ser obtida usando as informações apresentadas pelos processos anteriores, como correspondência estéreo e parâmetros de câmera.

Como em cada imagem tem-se a transformação $x = \Pi X$, $x' = \Pi' X$, conforme equação 3, essas equações podem ser combinadas na forma $AX = 0$, que é uma equação linear em X.

Primeiramente, o fator de escala homogênea é eliminado pelo produto cruzado que gera três equações para cada ponto da imagem, dessas equações, duas são linearmente independentes. Por exemplo, para a primeira imagem, $x \times (\Pi X) = 0$ obtem-se:

$$\begin{aligned} x(p^{3T}X) - (p^{1T}X) &= 0 \\ y(p^{3T}X) - (p^{2T}X) &= 0 \\ x(p^{2T}X) - y(p^{1T}X) &= 0 \end{aligned} \tag{9}$$

onde p^{iT} são as linhas de Π. Uma equação na forma $AX = 0$ pode ser organizada como:

$$A = \begin{bmatrix} xp^{3T} - p^{1T} \\ yp^{3T} - p^{2T} \\ x'p'^{3T} - p'^{1T} \\ y'p'^{3T} - p'^{2T} \end{bmatrix} \tag{10}$$

Para resolver essa equação pode-se utilizar a transformação linear direta (DLT).

2.6. Reprojeção

A reprojeção consiste em projetar o ponto 3D estimado para coordenadas bidimensionais. A informação propiciada por esse processo ajuda a melhorar o processo de reconstrução tridimensional. Isto porque, ao reprojetar é fácil perceber se os métodos desenvolvidos direcionam a uma estimativa satisfatória. Essa reprojeção pode ser inclusive observada visualmente contribuindo com a qualidade esperada da reconstrução.

A matriz projetiva Π transforma um ponto 3D em coordenadas de imagem, conforme equação 3. A reprojeção é equacionada conforme 11:

$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \Pi X, x = \begin{pmatrix} \frac{\hat{x}}{\hat{z}} \\ \frac{\hat{y}}{\hat{z}} \end{pmatrix} \tag{11}$$

onde o sinal ˆ representa o vetor em coordenadas homogêneas.

3. Experimento e Resultados

Para se encontrar a distorção causada pela própria lente e os dados intrínsecos da câmera é necessário calibrá-la. Para a calibração da câmera foi usado um aplicativo de calibração, presente no software MATLAB, que faz a comparação entre pares de imagens obtidas de vários ângulos e a uma distância predominante de um *checker patern* (tabuleiro semelhante ao de xadrez) [Zhang 2000]. Esse tabuleiro caracteriza-se por cubos de $25\ mm$, projetados nesse estudo a uma distância de $30\ cm$ entre a câmera e o tabuleiro. A calibração resultou na seguinte matriz de dados intrínsecos (DI):

$$DI = \begin{bmatrix} 3882.36274605904 & 0 & 2398.61232837601 \\ 0 & 3896.39198050779 & 1736.67578288591 \\ 0 & 0 & 1 \end{bmatrix}$$

Para a aquisição das imagens foi produzido um *gantry* [2] (Figura 1) que possibilita melhor precisão do movimento da câmera.

Figura 1. *Gantry* planar construído durante o projeto.

Após a calibração, iniciou-se o processo de construção esparsa do mapa de profundidade, criando a cena a ser investigada (Figura 2).

A aquisição das duas imagens foi feita considerando uma translação horizontal da câmera, (modelo GEX550), de aproximadamente $2\ cm$ de uma imagem para a outra.

Os parâmetros físicos da câmera (matriz de dados intrínsecos) possibilitaram o conhecimento das características ópticas e geométricas do aparelho [Heikkila and Silven 1997]. Por meio desses parâmetros foram retiradas as distorções

[2]Equipamento análogo a uma ponte, com suporte para câmera possibilitando a aquisição de uma imagem estável e mais fiel quanto a metragem da translação, visto que a câmera se movimenta horizontalmente rente a uma régua para medição em milímetros e centímetros.

Figura 2. Cena criada com câmera monocular calibrada.

radial e tangencial causadas pela lente da câmera, conforme Figura 3 (a). A detecção de pontos na cena, (Figura 3 (b)), foi feita utilizando uma função baseada no algoritmo Harris-Stephens [Shi et al. 1994].

(a) Imagem com remoção de distorções.

(b) Pontos detectados na imagem.

Figura 3. Transformações e pontos encontrados na primeira imagem da cena.

A Figura 4 ilustra a correspondência entre pontos identificados nas duas imagens da cena. Isto é, pontos que possuiam similaridade foram correlacionados seguindo o método de foto consistência NCC.

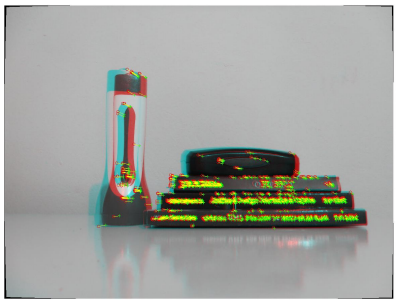
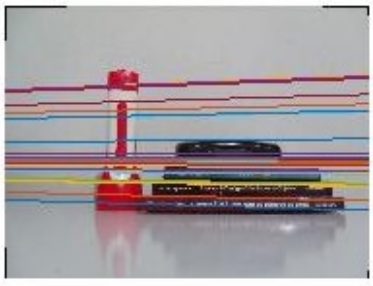


Figura 4. Correspondência estéreo entre as duas imagens da cena sobrepostas.

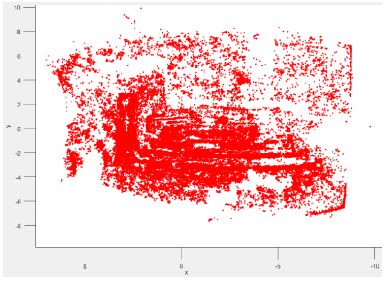
O resultado da colinearidade, produção e relacionamento de ponstos e suas respectivas linhas epipolares nas imagens da cena (Figura 5), foi obtido utilizando, novamente, os parâmetros intrínsecos da câmera. Logo, foi possível estimar a matriz essencial, que por sua vez possibilitou a normalização das imagens. Essa normalização permitiu a triangulação dos pontos, ou seja, sua computação tridimensional, formando uma nuvem de pontos com característica de profundidade sobre a cena, Figura 6 (a).

(a) Pontos na primeira imagem da cena.

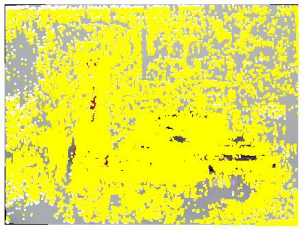


(b) Linhas epipolares na segunda imagem da cena

Figura 5. Pontos e linhas epipolares desenhadas a partir da correspondência das duas imagens que compõe a cena.



(a) Nuvem de pontos tridimensional.



(b) Reprojeção de pontos.

Figura 6. Nuvem de pontos criada durante a elaboração do mapa esparso de profundidade.

A reprojeção dos pontos, como ilustrado na Figura 6 (b), mostra que a estimativa 3D tem proximidade com os pontos 2D das imagens, embora outras abordagens devam ser empregadas para maior precisão. Nessa figura, os pontos em branco representam os pontos originais e os pontos em amarelo os pontos reprojetados.

A reconstrução dos pontos coloridos de acordo com os valores de cor da cena, aplicados sobre a nuvem de pontos, auxiliou na visualização do mapa esparso de profundidade construído, conforme apresentado na Figura 7.

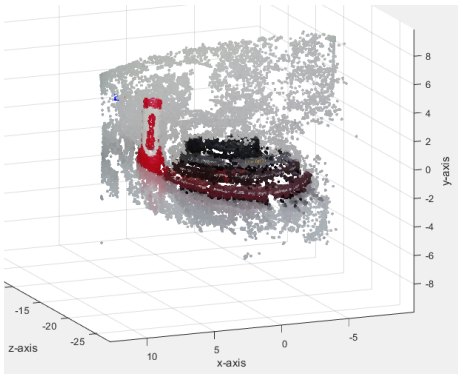


Figura 7. Mapa esparso de profundidade.

4. Conclusão

Os resultados alcançados no desenvolvimento do estudo são satisfatórios, visto que a estimativa de pontos 3D foi alcançada. Todavia, é um ponto de partida para que em outras pesquisas o método explore outras técnicas, objetivando uma maior precisão. Observa-se assim que dentro de cada processo outras abordagens podem ser empregadas a fim de obter resultados similares ou aperfeiçoados. Por exemplo, outros métodos de caracterização de pontos poderiam ser utilizados, outros métodos de triangulação poderiam ser aplicados, da mesma forma que outros modelos de calibração de câmera, correspondência estéreo e visualização.

Referências

Batista, N. A. R. and Regis, C. D. M. (2013). Obtenção da disparidade e dos mapas de profundidade em vídeos 3d. Number 23, pages 1160–1163. Revista Principia.

Furukawa, Y. and Hernández, C. (2015). Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148.

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

Heikkila, J. and Silven, O. (1997). A four-step camera calibration procedure with implicit image correction. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 1106–1112.

Jung, S. W. (2013). Enhancement of image and depth map using adaptive joint trilateral filter. In *Circuits and Systems for Video Technology*, pages 258–269. IEEE Transactions, 23(2) edition.

Laureano, G. T. and Paiva, M. V. (2013). Criação de mapas de disparidades empregando análise multi-resolução e agrupamento perceptual. In *IV Workshop de Visão Computacional*. IV Workshop de Visão Computacional.

Ma, Y., Soatto, S., Kosecka, J., and Sastry, S. S. (2012). *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media.

Radke, R. J. (2013). *Computer vision for visual effects*. Cambridge University Press.

Shi, J. et al. (1994). Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE.

Vieira, G., Soares, F., Parreira, R., Laureano, G., and Costa, R. (2016). Depth map production: approaches, challenges and applications. In *Proceedings of XII Workshop de Visão Computacional*, pages 323–328.

Zhang, Z. (2000). A flexible new technique for camera calibration. In *Pattern Analysis and Machine Intelligence*, volume 22, pages 1330–1334.