# MODELAGEM 3D URBANA A PARTIR DE IMAGENS DO STREETVIEW

**M. L. Rossi, M. L. Lopes, e G. A. Carrijo**

## RESUMO

Na constante evolução do planejamento urbano contemporâneo, a tecnologia, principalmente as ferramentas computacionais, desempenham um papel cada vez mais fundamental e indispensável no apoio a tomada de decisões. Neste contexto, os mapeamentos e informações geográficas apresentam uma vasta gama de mapas 2D, assim como a ampla possibilidade de obtê-los através de sensoriamento. Entretanto ferramentas que disponibilizam informações 3D de áreas urbanas e que permitam uma melhoria na interpretação dos dados e leitura real da informação, estão em surgimento e crescimento. Por outro lado, muitas cidades já contam com dados visuais de fotogrametria terrestre, como vistas panorâmicas de 360 graus de vários pontos da cidade que podem ser utilizadas como base para modelagens 3D. Assim esse trabalho procura aplicar técnicas de scanneamento 3D de áreas urbanas utilizando essas imagens, como alternativa ao emprego lazer scanners de alto custo para disponibilizam informações 3D das áreas urbanas.

## 1 INTRODUÇÃO

Cada vez mais novas técnicas e ferramentas de planejamento urbano têm dado atenção em questões como projeção do crescimento das cidades, planejamento territorial, administrativo e tributário. Comumente em tempos anteriores as decisões do planejamento urbano se davam da forma que o gestor do município visualizava individualmente como sendo o mais interessante para a cidade, levando a resultados ineficientes (Laurini, 1982A). Atualmente, diversos trabalhos têm sido realizados para transformar a tarefa de planejamento urbano algo mais científico (Laurini, 1982B). Na tentativa de melhorar o processo de planejamento urbano os gestores têm exigido mecanismos mais inteligentes no gerenciamento das cidades e, dessa forma, tem-se aumentado a utilização de Sistema de Informação Geográfica - SIGs no planejamento territorial.

Compondo a base das ferramentas SIG existem uma vasta gama de mapeamentos 2D e/ou a possibilidade de obtê-los através de sensoriamento, como as imagens aéreas ou de satélites (Laurini, 2001). Porém, Laurini (2001) apresenta a necessidade de utilização de modelos 3D nas ferramentas SIG e, para isso, tem-se recorrido, atualmente, ao emprego de lazer scanners, por sistemas de LIDAR (da sigla inglesa Light Detection And Ranging) para escanear e obter modelos 3D das cidades. No entanto os atuais equipamentos de LIDAR possuem um custo elevado para a aquisição e, também, para a atualização. Por

outro lado, várias cidades já apresentam fotografias de suas vias e fachadas de edificação disponíveis através de serviços de fotogrametria terrestre, similares ao Google Street View®, ou mesmo o próprio serviço da empresa Google, que podem ser utilizadas no lugar do escaneamento como forma de obter um modelo 3D.

Através de técnicas de processamento digital de imagens, principalmente as de visão computacional, é possível gerar modelos 3D de objetos observados por uma câmera. Juntamente ao processamento das imagens pode-se aplicar técnicas de robótica móvel, que são capazes de mapear os caminhos percorridos por um robô, e criar mapas de elevada precisão sem o auxílio de GPS (da sigla inglesa Global Positioning System ou Sistema de Posicionamento Global em português).

Assim, com a união de técnicas de processamento de imagem, para gerar modelos 3D, e das técnicas de mapeamento da robótica móvel, para o mapeamento, é possível criar modelos 3D de cidades utilizando ferramentas simples, como fotogrametria terrestre que apresenta as fachadas de edificação ou através da utilização de câmeras que possuem custo mais acessível comparado aos atuais scanners à laser (LIDAR) em troca de um maior custo computacional.

Dessa forma este trabalho busca apresentar uma metodologia computacional que permita gerar esses modelos 3D para as ferramentas SIG através de várias imagens em 2D, em especial as de fotogrametria terrestre no nível das vias.

## 2 SCANEAMENTO 3D UTILIZANDO FOTOGRAFIAS

Para que o processo de scaneamento 3D utilizando fotografias possa ser compreendido é necessário verificar o fenômeno físico por trás do processo da fotografia. A Figura 1 apresenta o processo físico por trás da fotografia. Pela Figura 1 pode-se verificar que cada um dos pontos do objeto [$P_i$] possuem três coordenadas [$X_i$, $Y_i$, $Z_i$]. Porém, os pontos da imagem [$p_i$] possuem apenas duas coordenadas [$u_i$, $v_i$]. Dessa forma, pode-se dizer que o processo de fotografia perde informações da profundidade de uma cena.

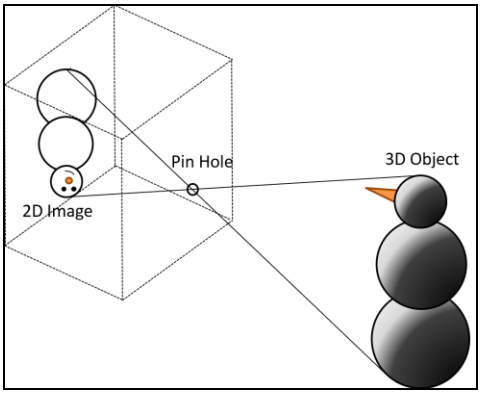


**Fig. 1 Perda da profundidade no processo da fotografia**

De acordo com Forsyth, e Ponce (2002) o modelo matemático para uma câmera de orifício (modelo mais simples de câmera) é definido de acordo com a Equação (1) e simplificado como (2).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & s & x_0 & 0 \\ 0 & f & y_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

$$[p_i] = [C] \,.\, [P_i] \quad (2)$$

Para as Equações (1) e (2) as coordenadas dos pontos da cena $[P_i]$ (valores de X, Y e Z) e coordenadas dos pontos na imagem $[p_i]$ (valores de u e v) são escritas no sistema de coordenadas homogênea (Semple e Kneebone, 1998; Mohr, 1993). A matriz [C] é denominada de matriz da câmera (ou matriz de calibração da câmera), pois ela carrega as características da câmera como: f que corresponde à distância focal; s corresponde à distorção do pixel; e $x_0$ e $y_0$ correspondem à posição do centro focal da cena.

Devido ao seu formato a matriz [C] é irreversível e, assim, pode-se afirmar que é impossível recuperar a profundidade da cena conhecendo-se apenas uma imagem e a matriz da câmera. Portanto, para obter a profundidade utiliza-se mais de uma imagem e, então, se realiza a triangulação dos pontos (Hartrley, 1997).

O processo de triangulação consiste em relacionar o mesmo ponto da cena em duas imagens diferentes, das quais já se conhecem as posições das câmeras, e, dessa forma, recuperar a informação sobre a coordenada perdida durante a aquisição das imagens, ou seja, a profundidade do ponto na cena. A Figura 2 apresenta de forma simplificada o processo de triangulação.

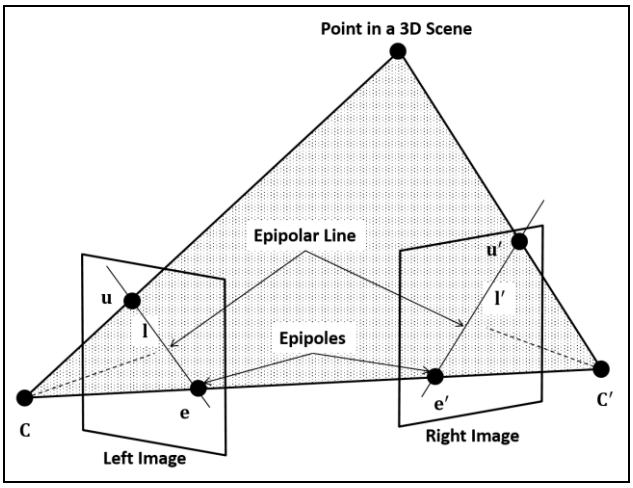


**Fig. 2 Processo de triangulação**

A triangulação é bem-sucedida quando o mesmo ponto de uma cena for identificado em diferentes imagens e, também, se as posições das câmeras forem conhecidas.

Para correlacionar os pontos de forma adequada os mesmos não devem ser escolhidos ao acaso, pois isso pode levar a correlações inválidas. Os pontos a serem escolhidos devem possuir certas características que os permitam serem identificados em imagens distintas. Esses pontos são, então, conhecidos como Pontos Característicos. Existem vários algoritmos que permitem identificar pontos característicos nas imagens, como os

algoritmos: FAST (Rosten e Drummond, 2006), Harris (Harris e Stephens, 1988), SURF (Bay *et al*, 1988), BRISK (Leutenegger *et al*, 2011) e o MSER (Matas, 2002).

Conhecido os vários pontos característicos em duas imagens distintas basta correlacioná-los, ou seja, encontrar os pares correspondentes. Esse processo é feito de forma que o conjunto apresente o menor erro possível. Técnicas como a dos mínimos quadrados podem levar a erros como apresentados na Figura 3.

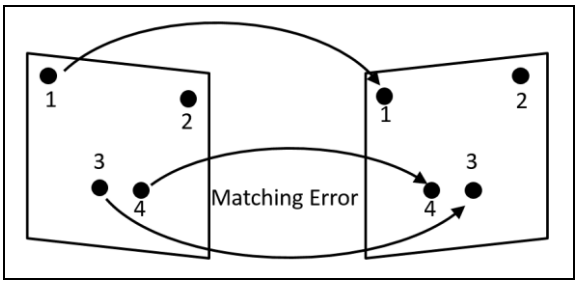


**Fig. 3 Erro no processo de correlação**

Na tentativa de contornar esses erros, algoritmos mais robustos capazes de superar essa dificuldade foram desenvolvidos. A Mediana e o M-estimadores são estimadores robustos bastante utilizados em análises de estatísticas. Entretanto, para o condicionamento robusto dos modelos paramétricos utilizados em visão computacional o algoritmo RANSAC (da sigla inglesa RANdom SAmple Consensus ou Consenso de Amostras escolhidas aleatoriamente em português) tornou-se o método padrão (Fischler e Bolles, 1981).

Obtido o correlacionamento (ou casamento) dos pontos característicos das imagens, a matriz de características das câmeras e as posições das câmeras é possível fazer a triangulação desses pontos e, assim, encontrar a distância de cada ponto da cena até as câmeras. Obtido a distância dos pontos em relação as câmeras e conhecido as posições das câmeras torna-se possível o posicionamento dos pontos no espaço criando, assim, uma nuvem de pontos no espaço que representa o objeto observado. Essa nuvem que utiliza apenas os pontos casados é denominada de 3D Esparso.

A Figura 4 apresenta o resultado da metodologia descrita utilizando uma bancada de laboratório que possui uma câmera que se move ao redor de um objeto. Neste caso o ângulo de giro da câmera em cada foto é conhecido e, dessa forma, o movimento da câmera pode ser deduzido utilizando equações da robótica.

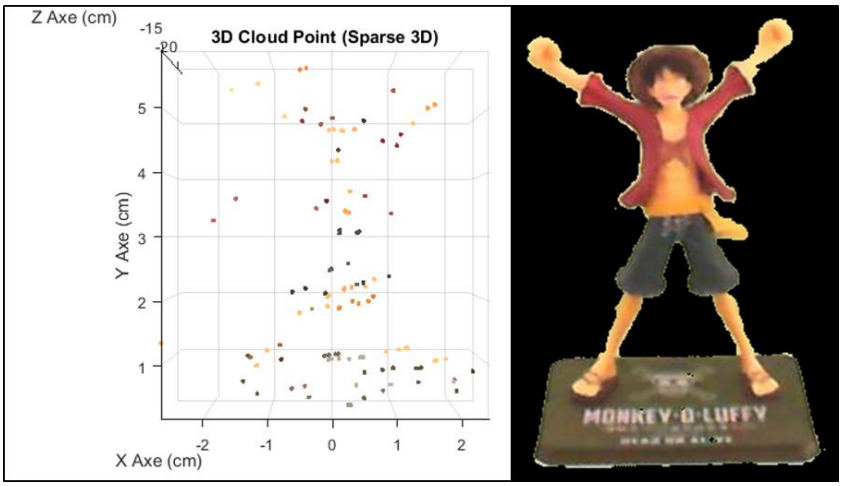


**Fig. 4 3D Esparso e o objeto utilizado**

Obtido a nuvem do 3D Esparso pode-se buscar pontos visíveis em duas, ou mais, fotografias que se localizam entre os pontos do 3D Esparso, aumentando a quantidade de pontos de forma a obter uma densa nuvem de pontos em 3D, denominado 3D Denso (Furukawa *et al*, 2008; Hartley e Zisserman, 2004). Após o 3D Denso é possível ligar os pontos com uma malha gerando, assim, um volume 3D. A Figura 5 apresenta o resultado do trabalho de Fitzgibbon e Zisserman (1998), no qual é gerado um modelo 3D de um objeto utilizando apenas fotografias do objeto. Neste trabalho Fitzgibbon e Zisserman (1998) implementam, além da metodologia do 3D Esparso, o 3D Denso e fecham os pontos com uma malha.

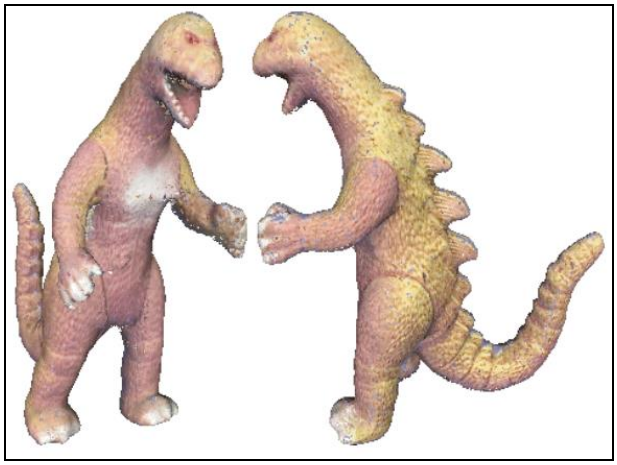

**Fig. 5 3D Denso e sua texturização**

## 3 PROCEDIMENTO PARA O RECONHECIMENTO DAS POSIÇÕES DE CÂMERA PARA O MODELAMENTO 3D DE ÁREAS URBANAS

Todos os resultados dos trabalhos apresentados na seção anterior foram obtidos ao utilizarem câmeras que se movem de acordo com modelos matemáticos. Entretanto, não há uma equação que possa descrever a movimentação de uma câmera movendo-se pelas cidades. Dessa forma, surge a necessidade de encontrar as posições de câmeras que se movem pela cidade, informação imprescindível para o processo de triangulação.

Para obtenção das posições das câmeras apropria-se da técnica da robótica móvel conhecida como SLAM (acrônimo de Simultaneous Localization And Mapping ou Mapeamento e Localização Simultânea). Essa é uma técnica recente que tem sido desenvolvida para uso em robôs móveis, permitindo que eles possam se localizar e, também, descrever o seu movimento através da visão computacional.

## 3.2 Entendendo o SLAM

O SLAM pode ser definido como uma tarefa onde um robô móvel avalia a sua posição ao analisar o ambiente. Essa análise pode ser através de uma imagem obtida por uma câmera e, assim, o robô móvel pode ser considerado uma câmera móvel. A função deste robô (ou câmera) é de mapear um ambiente, sem a utilização de qualquer informação previa deste, e, simultaneamente, localizar-se neste ambiente somente a partir do mapa gerado, das ações de controle recebidas e das medições realizadas através dos seus próprios sensores (Durrant-Whyte e Bailey, 2006). A realização desta tarefa é feita de forma incremental, aumentando a certeza da posição do robô e também ao refinamento na representação do ambiente pelo mapa gerado (Silveira, 2015).

Considerando um robô se deslocando em um ambiente, ao circular por este ambiente são adquiridas informações do ambiente por onde ele passa através de sua câmera. Estas informações são denominadas Landmarks e auxiliam o robô identificar a sua posição e fazer o mapeamento. A Figura 6 apresenta o deslocamento de um robô obtendo os Landmarks e a sua tentativa de se localizar na região.

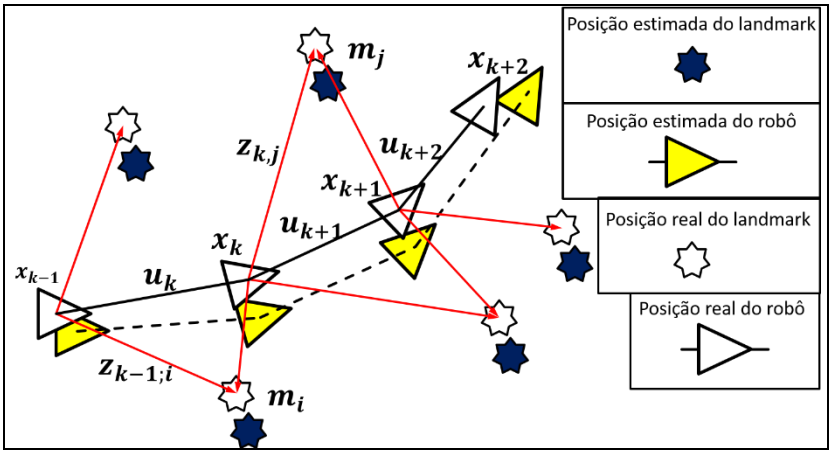


**Fig. 6 Robô mapeando uma área utilizando técnicas de SLAM**

Em cada instante de tempo k, apresentado na Figura 6, o robô móvel está em uma posição diferente e visualiza diferentes landmarks e pretende seguir uma ação.

Para facilitar a análise do SLAM Durrant-Whyte e Bailey (2006) recomendam as seguintes variáveis:

- $x_k$: Vetor de estado de posição. Define a localização do robô móvel no espaço;
- $u_k$: Vetor de ações de controle a serem aplicadas em um tempo k-1 para levar o veículo à posição $x_k$ em um tempo k;
- $m_i$: Vetor representando a localização da i-ésima landmark. Aqui é assumido que a posição das landmarks são invariantes no tempo;

- $z_{ik}$: Vetor que representa a posição da i-ésima landmark, medido a partir do robô pelos sensores do robô;
- $X_{0:k} = \{x_0, x_1, ..., x_k\} = \{X_{0:k-1}, x_k\}$: Conjunto que representa o histórico das localizações do robô móvel;
- $U_{0:k} = \{u_0, u_1, ..., u_k\} = \{U_{0:k-1}, u_k\}$: Conjunto que representa o histórico das ações de controle;
- $m = \{m_0, m_1, ..., m_k\}$: Conjunto que representa todas as posições de landmark;
- $Z_{0:k} = \{z_0, z_1, ..., z_k\} = \{Z_{0:k-1}, z_k\}$: Conjunto que representa as observações das landmarks.

A Figura 6 apresenta essas variáveis sugeridas em Durrant-Whyte e Bailey (2006) para problemas que envolvam o SLAM. De forma a reduzir as incógnitas apresentadas no problema do SLAM, deve-se conhecer ao menos uma das posições que, normalmente, é a posição $x_0$ como sendo o vetor posição O = {0;0;0}, também definido como a origem do sistema de coordenadas. A partir dessa hipótese os outros vetores de posição são calculados a partir de O e o movimento feito pela câmera.

Através da observação do ambiente, o robô corrige (ou revisa) as informações do seu movimento, utilizando m (tipicamente estático) como o mapa real do ambiente. Cada informação desse mapa possui uma posição real, representada a partir do vetor $m_i$. Quando o robô se movimenta ele mede, analisando as imagens, as distâncias até $m_i$ e essas medições de distâncias são representadas pelo conjunto $Z_{0:k}$ (Silveira, 2015).

Por fim, a abordagem do SLAM baseia-se na probabilidade estatística do robô estar em uma determinada coordenada após realizar medições de sua posição às landmarks visualizadas por ele. Matematicamente o problema do SLAM é representado pela distribuição de probabilidade indicada pela Equação (3) calculada a cada instante de tempo k (Durrant-Whyte e Bailey, 2006).

$$P(x_k, m | Z_{0:k}, U_{0:k}, x_0) \quad (3)$$

A Equação (3) descreve a densidade de probabilidade a posteriori conjunta sobre as localizações das landmarks e a posição da câmera no instante de tempo k. E para isso são consideradas todas as observações registradas e ações de controle recebidas até o tempo k, inclusive as do tempo k, e a posição inicial. Em outras palavras a Equação (3) traduz o problema de SLAM como sendo as chances que há da câmera estar posicionada no ponto $x_k$ observando as informações de m, dado que é conhecido as observações $Z_{0:k}$, os históricos de movimentações $U_{0:k}$, e a posição inicial $x_0$.

Uma solução recursiva é normalmente desejável, e pode ser realizada calculando-se a junção a posteriori a partir do teorema de Bayes (Bayes e Price, 1763; Allen, 1999). Neste caso deve-se iniciar com uma estimativa de $P(x_{k-1}, m | Z_{0:k-1}, U_{0:k-1})$, para o tempo k-1, o que também requer os modelos de observação e de transição de estados (Durrant-Whyte e Bailey, 2006; Thrun *et al*, 1999). Ou seja, cada vez que se calcula uma posição da câmera todas as posições anteriores devem ser revisadas e atualizadas.

Dessa forma, sempre que for encontrado um ponto já conhecido como, por exemplo, rodear uma quadra ou reencontrar um cruzamento, todas as informações de posição são recalculadas e aperfeiçoadas. A Figura 7 apresenta o resultado obtido por Torii *et al* (2009) ao aplicar as técnicas de SLAM com imagens do Google Street View®. Os pontos

vermelhos na Figura 7 representam as posições das câmeras encontradas apenas ao analisar as imagens.


**Fig. 7 Robô mapeando uma área utilizando técnicas de SLAM**

## 4 RESULTADOS

As técnicas demonstradas acima permitem combinar o scanneamento 3D através de fotografias com o mapeamento e o reconhecimento de posição do SLAM, tendo como material de análises fotogrametria terrestre tiradas no nível das vias como as do Google Street View® de forma a obter bons modelos 3D de cidades.

A Figura 8 apresenta o trabalho do *Center for Machine Perception* da *Czech Technical University* no qual fora feito a combinação dessas técnicas juntamente com fotografias tiradas no nível das vias.

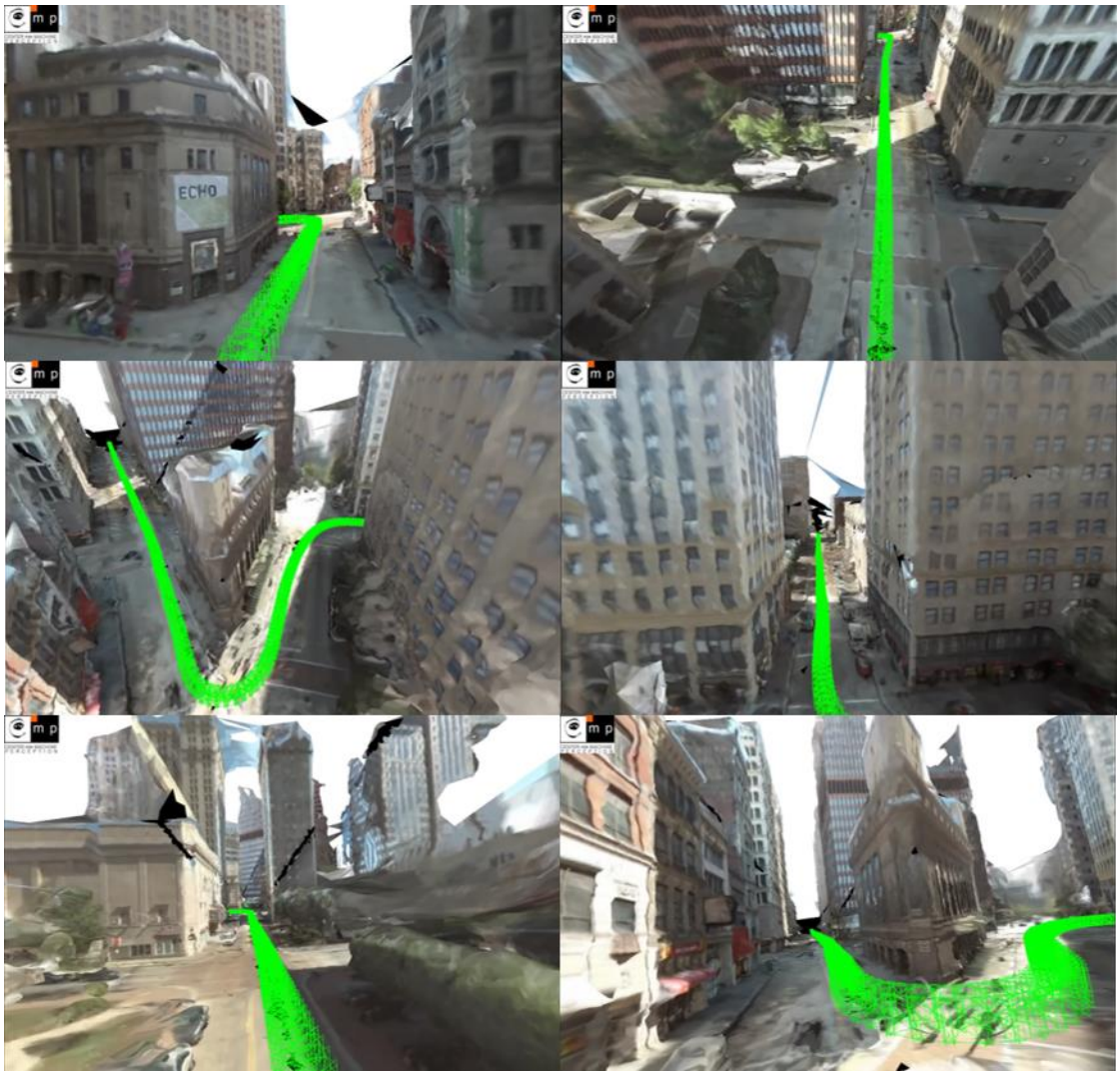


**Fig. 8 Scanneamento 3D através de fotografias combinado com SLAM**

Pela Figura 8 pode-se verificar que a combinação da técnica de scennaemento 3D através de fotografias juntamente com a técnica de SLAM é uma ferramenta poderosa para o modelamento 3D de áreas urbanas.

## 5 CONCLUSÃO

Ao observar os resultados apresentados na Figura 8 verifica-se que é possível obter modeles 3D de áreas urbanas, sem a necessidade de utilização de equipamentos como LIDAR, através de técnicas computacionais adequadas de processamento de imagem e de robótica permitindo, assim, obter modeles 3D de áreas urbanas de forma satisfatória utilizando apenas uma câmera fotográfica ou filmadora.

Além do mais, várias cidades contam com um acervo disponível de imagens de suas vias e fachadas de edificações, o que permite uma rápida implementação desses recursos para a modelagem 3D dessas cidades, obtendo mais um recurso a ser adicionado em ferramentas SIG para a geste planejamento urbano.

## 6 AGRADECIMENTOS

Os autores agradecem a Universidade Federal de Uberlândia e a Universidade Federal de Pelotas pela infraestrutura laboratorial. Eles também agradecem ao *Center for Machine Perception* da *Czech Technical University* pelos serviços prestados em seu site e, também, ao CNPQ, FAPEMIG e CAPES pelo apoio para a publicação desse artigo.

## 7 REFERÊNCIAS

Allen, R. (1999) **David Hartley on Human Nature: A Pragmatic Engagement with Contemporary Perspectives**. 1 ed. Nova York: State University of New York Press.

Bay, H., Tuytelaars, T. e Gool, L. V. (2008) SURF: Speeded Up Robust Features. **Computer Vision and Image Understanding**, 110(3), 346–359.

Bayes, M.; Price, M. (1963) An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. **Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S**. 370–418p. v.53.

Durrant-Whyte, H. F. e BAILEY, T. (2006) Simultaneous localization and mapping: part I. **IEEE Robotics & Automation Magazine**, 13(2), 99-110.

Fischler, M. A. e 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.

Fitzgibbon, A. W. e Zisserman, A. (1998) Automatic Camera Recovery for Closed or Open Image Sequences, **Proceedings 5TH EUROPEAN CONFERENCE ON COMPUTER VISION**, London, England, 2-6 Junho 1998.

Forsyth, D. e Ponce, J. (2002). **Computer Vision: A Modern Approach**. 1 ed. Nova York: Pearson.

Furukawa, Y. e Ponce, J. (2008) Accurate, Dense and Robust Multi-View Stereopsis, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 32(8), 1362-1376

Harris, C. e Stephens, M. A. (1988) Combined Corner and Edge Detector, **4th Alvey Vision Conference**, University of Manchester, Manchester, 31 Agosto-2 setembro 1988.

Hartley, R. I. e Zisserman, A. (2004) **Multiple View Geometry in Computer Vision**. 1 ed Cambridge: Cambridge University Press

Hartrley, R. I. (1997) In defense of the eight-point algorithm, **IEEE Transactions on Pattern Analysis anda Machine Intelligence**, 19(6), 580-593.

Laurini, R. (1982A) French Local Planning Practice, in: M. Batty e B. Hutchinson (eds.), **Systems Analysis in Urban Policy-Making and Planning**, New York: Plenum Press.

Laurini, R. (1982B) Nouveaux outils informatiques pour l'élaboration conjointe des plans d'urbanisme, **Proceedings 9th European Symposium on Urban Data Management Symposium (UDMS)**, Valencia, Spain, 26-29 Outubro 1982.

Laurini, R. (2001) **Information Systems for Urban Planning: a Hypermedia Cooperative Approach**. 1 ed Boca Raton: CRC Press.

Leutenegger, S., Chli, M. e Siegwart, R. (2011) BRISK: Binary Robust Invariant Scalable Keypoints, **Inernational Conference on Computer Vision (ICCV)**, Barcelona, Spain, 6-13 Novembro 2011.

Matas, J., Chum, O., Urban, M. e Pajdla, T. (2002) Robust Wide Baseline Stero from Maximally Stable External Regions, **The British Machine Vision Conference**, Cardiff, Reino Unido 2-5 Setembro 2002.

Mohr, R. (1993) Projective Geometry and Computer Vision, in C. H. Chen, L. F. Pau, Wang e P. S. Patrick (eds.), **Handbook of Pattern Recongnition & Computer Vision**, River Edge.

Rosten, E. e Drummond, T. (2006) Machine Learning for High-Speed Corner Detection, **9th European Conference on Computer Vision**, Graz, Áustria, 7-13 Maio 2006.

Semple, J. G. e Kneebone, G. T. (1998) **Algebraic Projective Geometry**, 1 ed Clarendon Press.

Silveira, L. (2015) **Sistema Bioinspirado para Mapeamento e Localização de Robôs Móveis em Ambientes Subaquáticos**, Universidade Federal do Rio Grande, Rio Grande.

Thrun, S., Fox, D. e Burgard, W. (2005) **Probabilistic robotics**, 2 ed The MIT Press.

Torii, A., Havlena, M. e Pajdla, T (2009) From Google Street View to 3D city models, **Proceedings 12th International Conference on Computer Vision Workshops (ICCV Workshops)**, Kyoto, Japão, 27 de setembro-4 de outubro 2009.