



UNIVERSIDADE DO ESTADO DO
RIO DE JANEIRO



INSTITUTO POLITÉCNICO
GRADUAÇÃO EM ENGENHARIA
DE COMPUTAÇÃO

Pedro Felipe Pena Barata

Aplicações práticas em técnicas de reconstrução 3D

Nova Friburgo

2017



UNIVERSIDADE DO ESTADO DO
RIO DE JANEIRO



INSTITUTO POLITÉCNICO
GRADUAÇÃO EM ENGENHARIA
DE COMPUTAÇÃO

Pedro Felipe Pena Barata

Aplicações práticas em técnicas de reconstrução 3D

Trabalho de Conclusão de Curso apresentado, como requisito parcial para obtenção do título de Graduado em Engenharia de Computação, ao Departamento de Modelagem Computacional do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Ricardo Fabbri

Nova Friburgo

2017

Pedro Felipe Pena Barata

Aplicações práticas em técnicas de reconstrução 3D

Trabalho de Conclusão de Curso apresentado, como requisito parcial para obtenção do título de Graduado em Engenharia de Computação, ao Departamento de Modelagem Computacional do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

Aprovada em 30 de 09 de 2017.

Banca Examinadora:

Prof. Dr. Ricardo Fabbri (Orientador)
Departamento de Modelagem Computacional – UERJ

Prof. Dr. Edirlei Soares
Departamento de Modelagem Computacional – UERJ

Prof. Dr. Roberto Pinheiro
Departamento de Modelagem Computacional – UERJ

Nova Friburgo
2017

AGRADECIMENTOS

RESUMO

BARATA, Pedro Felipe Pena. *Aplicações práticas em técnicas de reconstrução 3D.* 2017. 29 f. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) – Departamento de Modelagem Computacional, Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2017.

A partir dos anos 2000, a área de reconstrução 3D vem sido amplamente explorada. No início, sensores de alcance, tanto aéreos quanto terrestres, eram empregados em diferentes aplicações, devido à facilidade de manuseio e ao baixo custo. Porém, constantes melhorias na tecnologia, sobretudo, nos *hardwares* e *softwares* no âmbito da reconstrução, fizeram com que hoje, quase duas décadas depois, novas técnicas surgissem.

Muitos cientistas que utilizavam a fotogrametria converteram seus esforços na área dos sensores à laser. Pois além de executarem uma reconstrução mais rápida, possuem uma altíssima acurácia, compensando seu alto custo inicial. Isto dificultou e desacelerou o processo de descoberta de novos algoritmos e métodos na área da fotogrametria.

Hoje em dia, graças à esse avanço, a fotogrametria, aliada a novos algoritmos, como o *Structure of Motion (SfM)*, pontos em comum e de combinação de imagens, por exemplo, consegue competir com scanners à laser e sensores de alcance.

ABORDAR O SFM, SIFT E CMVS (POUCO) iii ?????

Com uma combinação de algoritmos, com o *SfM*, junto com o SIFT () e o CMVS (), é possível gerar uma reconstrução satisfatória apenas utilizando uma câmera de um *smartphone*.

O trabalho foi estruturado da seguinte maneira: previamente apresentam-se os objetivos do projeto, destacando suas funcionalidades e metas, a seguir divide-se em capítulos; O Capítulo 1, que introduz o funcionamento de cada algoritmo e técnica empregada, apresentando e debatendo, comparativamente pontos à favor e contra; O Capítulo 2 é dedicado à ferramenta gráfica utilizada para a obtenção dos resultados (VisualSfM). Finalmente, apresentamos os resultados e conclusões do trabalho, bem como sugestões para implementações e trabalhos futuros.

ABORDAR OS "TODO'S" DO HANGOUTS

Palavras-chave: Reconstrução densa. Nuvem de pontos. SfM. Triangulação.

ABSTRACT

BARATA, Pedro Felipe Pena. . 2017. **29** f. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) – Departamento de Modelagem Computacional, Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2017.

Keywords:

LISTA DE FIGURAS

Figura 1 - Kinects de primeira geração (a) consistindo de câmeras e projetores infra-vermelho (b) e de segunda geração, consistindo de tecnologia ToF (c). Ambos os kinects são largamente utilizados para escaneamento em tempo real, formando a base de scanners manuais (d), porém nem sempre são úteis para preservação detalhada de patrimônio. Um dos objetivos deste projeto é explorar os limites desta tecnologia.	11
Figura 2 - A reconstrução usando-se Kinect (de primeira ou segunda geração) usando software atual de super-resolução (c) fornece precisão similar a um sistema estéreo de média resolução, inferior um sistema a laser de alta qualidade (d) porém de baixo custo e muito mais versátil devido ao sistema de aquisição manual e a software amplamente utilizado e desenvolvido(????).	12
Figura 3 - Protótipo do scanner a laser de triangulação. O objeto a ser escaneado é uma réplica em tamanho real de um sarcófago egípcio (a). O scanner foi reconfigurado para escanear objetos maiores, pois a escultura possui 517 centímetros (b), o da cabeça também sofreu uma reconfiguração, este scanner gira em 90 graus, que faz o laser rotacionar, da posição horizontal para a vertical e também roda em torno da cabeça como um todo (c). Para a reconstrução, o primeiro passo foi alinhar cerca de 100 scans em diversas posicoes, após isso, utilizado um alinhamento automatico em pares dos scans, utilizando um algoritmo modificado de iteracoes de pontos próximos (ICP - <i>iterated-closest-points</i>). Após isso, faz-se um processo de relaxação global a fim de minimizar erros de alinhamento por toda a estátua. Depois de alinhados, usa-se o algoritmo de profundidade volumétrica de processamento de imagens (VRIP - <i>volumetric range image processing</i> - do Brian Curless) (d).	12
Figura 4 - Algumas esculturas do Jardim do Nêgo	13
Figura 5 - A reconstrução usando-se apenas imagens, sem controle de aquisição, como em um vídeo de um smartphone filmado em torno do objeto, fornece uma nuvem de pontos, que pode ser densificada (??????????), ou atribuída de curvas (?????????), de forma a preservar a resolução em áreas de alto conteúdo informativo. Tais representações estão sendo atualmente unificadas na pesquisa da área. Este projeto propõe explorar os limites da reconstrução 3D usando-se apenas imagens, no contexto de preservação de patrimônio.	13

Figura 6 - É aplicado um filtro gaussiano na imagem original (a), com $\sigma = 1$, tendo como resultado a imagem (b). Um outro filtro gaussiano é usado, porém, neste caso, o $\sigma = 2$ (c). Após isso, subtrai-se (b) de (c), obtendo o filtro DoG (d).	17
Figura 7 - Exemplo de funcionamento de detecção de espaço-escala extrema	18
Figura 8 - Exemplo do resultado obtido do histograma orientado	20
Figura 9 - Exemplo de um descritor de pontos-chaves, com uma matriz 2x2 e uma região 8x8	20
Figura 10 - Uma triangulação utilizando um ponto qualquer, X_j . Onde cada câmera C_1, C_2, C_3 possui um <i>feature</i> correspondente a cada uma delas, respec- tivamente, X_{1j}, X_{2j}, X_{3j} .	21
Figura 11 - Erro proveniente da reprojeção, onde os pontos x e x' estão mais próximos das medidas reais da imagem.	22

SUMÁRIO

	INTRODUÇÃO	10
0.1	Introdução e Justificativa	10
0.1.1	<u>O Jardim do Nêgo, Nova Friburgo</u>	13
0.2	Objetivos	14
0.3	Organização deste manuscrito	15
1	PONTOS DE INTERESSE	16
1.1	SIFT – Scale Invariant Feature Transform	16
1.1.1	<u>Detecção de espaço-escala extremos</u>	16
1.1.2	<u>Localização de pontos-chaves</u>	18
1.1.3	<u>Atribuição de orientação</u>	19
1.1.4	<u>Descriptor de pontos-chaves</u>	20
1.1.5	<u>Combinação de pontos-chaves</u>	21
1.2	Triangulação – Full pairwise image matching	21
2	RECONSTRUÇÃO DENSA	23
2.1	Introdução	23
2.2	HPMVS	23
2.2.1	<u>falar sobre o hpmvs...</u>	23
2.3	MVE	23
2.3.1	<u>Guia de reconstrução com o MVE</u>	24
3	KINECT	26
4	VISUALSFM	27
5	EXPERIMENTOS	28
	CONCLUSÃO	29

INTRODUÇÃO

0.1 Introdução e Justificativa

A reconstrução 3D de cenas gerais a partir de múltiplos pontos de vista usando-se câmeras convencionais, sem aquisição controlada, é um dos grandes objetivos de pesquisa em visão computacional, ambicioso até mesmo para os dias de hoje. Aplicações incluem a reconstrução de modelos 3D para uso em videogames (??), filmes (??), arqueologia, arquitetura, modelagem 3D urbana (*e.g.*, Google Streetview); técnicas de *match-moving* em cinematografia para fusão de conteúdo virtual e filmagem real (??), a organização de uma coleção de fotografias com relação a uma cena (*e.g.*, o sistema *Phototourism* (??) e a funcionalidade *Look Around* do Google Panoramio e Street View), manipulação robótica, e a metrologia a partir de câmeras na indústria automobilística e metal-mecânica.

Os desafios estão ligados às escolhas de grande escala de representações adequadas e de técnicas que possam modelar simultaneamente com materiais drásticamente diferentes (*e.g.*, não-Lambertianos), modelos geométricos (*e.g.*, variedades curvilíneas gerais, descontinuidades, texturas, deformações, em escalas diferentes), tipos de regiões (com ou sem textura), condições de iluminação variadas, sombras, fortes diferenças de perspectivas, desbalanceamento devido a excesso de detalhes em partes menos importantes, número arbitrário de objetos e câmeras não-calibradas.

Mesmo que um sistema completo esteja fora do alcance da tecnologia atual, um progresso significativo tem sido atingido nos últimos anos. Por um lado, uma tecnologia operacional tem evoluído, mais recentemente para sistemas de grande escala (????), a partir do desenvolvimento da detecção robusta de *features* (??), o *fitting* robusto e seleção de correspondências baseados em RANSAC, e o desenvolvimento de métodos de geometria projetiva para calibrar duas ou três imagens e progressivamente adicionar imagens e extrair estrutura 3D dessas *features* na forma de nuvens de pontos. Com o código fonte do sistema Bundler (????) liberado por Noah Snavely, e sua subsequente incorporação ao sistema VisualSfM (??), é possível utilizar este sistema para a reconstrução de patrimônio.

No paradigma usando-se apenas imagens convencionais – denominado **reconstrução estéreo multiocular passiva** – a posição das câmeras são estimadas a partir apenas de imagens, usando pontos de interesse, em seguida uma nuvem de pontos é reconstruída ???. As câmeras podem então ser utilizadas para obter modelos mais detalhados de reconstrução, como algoritmos de densificação (??) e interpolação (??) da nuvem de pontos, bem como demais algoritmos densos de visão estéreo multi-perspectiva/multi-ocular, como os do grupo de Michel Goesele (??), também com código disponível. Tais algoritmos, no entanto, têm problemas, em particular a reconstrução suaviza partes bem-

delineadas do objeto, e pode conter buracos em áreas homogêneas. Pode-se, portanto, utilizar a reconstrução 3D de curvas do pesquisador proponente (?????????) para auxiliar na reconstrução mais bem-delinada nesses casos problemáticos, bem como para ajudar no problema de escalabilidade quando a reconstrução 3D se torna muito grande. Um segundo paradigma, denominado **reconstrução estéreo multiocular ativa**, tem se tornado viável devido à indústria de videogames, e consiste na utilização de sistemas que alteram o funcionamento de câmeras convencionais, típicamente usando-se projetores infra-vermelho, laser ou câmeras ToF (time of flight), como no caso dos dispositivos Kinect, figura 1.

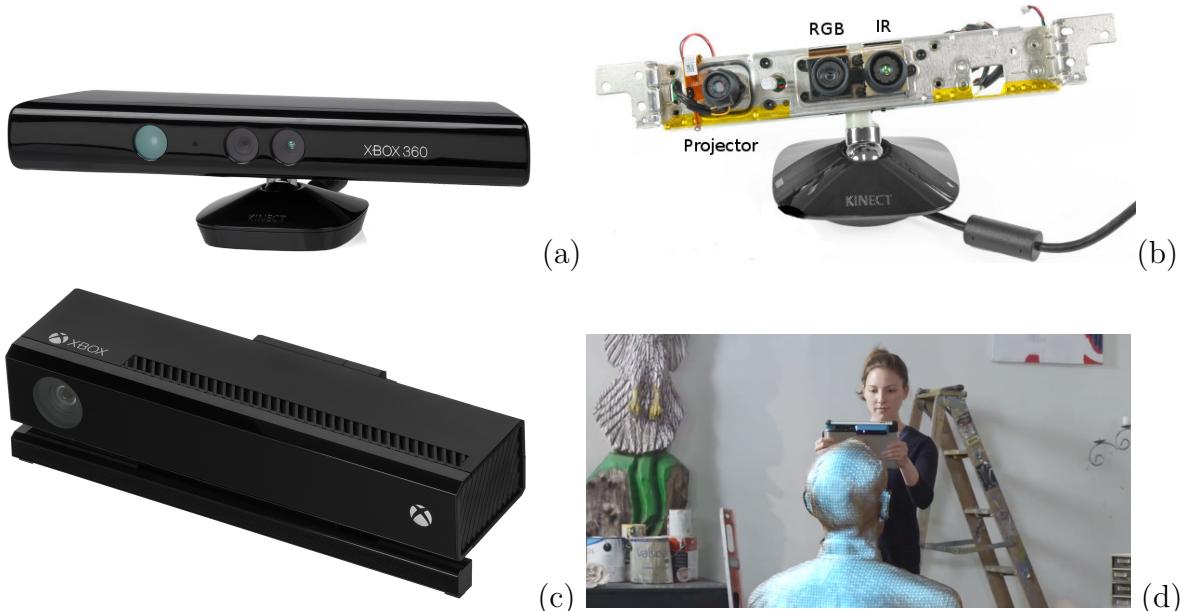


Figura 1 - Kinects de primeira geração (a) consistindo de câmeras e projetores infra-vermelho (b) e de segunda geração, consistindo de tecnologia ToF (c). Ambos os kinects são largamente utilizados para escaneamento em tempo real, formando a base de scanners manuais (d), porém nem sempre são úteis para preservação detalhada de patrimônio. Um dos objetivos deste projeto é explorar os limites desta tecnologia.

A preservação de patrimônio tem sido realizada tradicionalmente com scanners dedicados de alto custo, como no projeto David 3. O projeto teve início em 1992 e tem como objetivo a utilização de scanners a laser de profundidade (*rangefinder scanners*), aliado com algoritmos que combinam diferentes profundidades e cores da imagem, para realizar uma digitalização da parte externa e da superfície de forma acurada da estátua de David (porém, esse método pode ser utilizado em diferentes objetos no mundo real, como partes de máquinas, artefatos culturais e na indústria de video games, por exemplo). Para as partes mais detalhadas, foi utilizado um scanner de menor escala que faz uma pequena triangulação com laser de profundidade.

Seria de grande interesse explorar os dois paradigmas supracitados para avaliar as possibilidades disponíveis no estado da arte de reconstrução 3D para o escaneamento de

baixo custo para a preservação de Patrimônio. O que se pode atingir com apenas uma filmagem de esculturas realizada por um smartphone, sem calibração prévia e *in situ*, ou seja, sem ambiente controlado? Como esta reconstrução se compara nos dias de hoje com a reconstrução realizada por um scanner padrão baseado em Kinect?

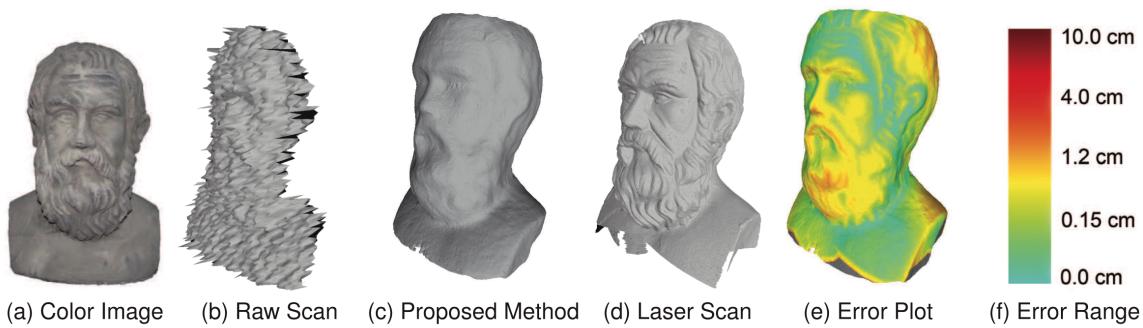


Figura 2 - A reconstrução usando-se Kinect (de primeira ou segunda geração) usando software atual de super-resolução (c) fornece precisão similar a um sistema estéreo de média resolução, inferior a um sistema a laser de alta qualidade (d) porém de baixo custo e muito mais versátil devido ao sistema de aquisição manual e a software amplamente utilizado e desenvolvido(????).

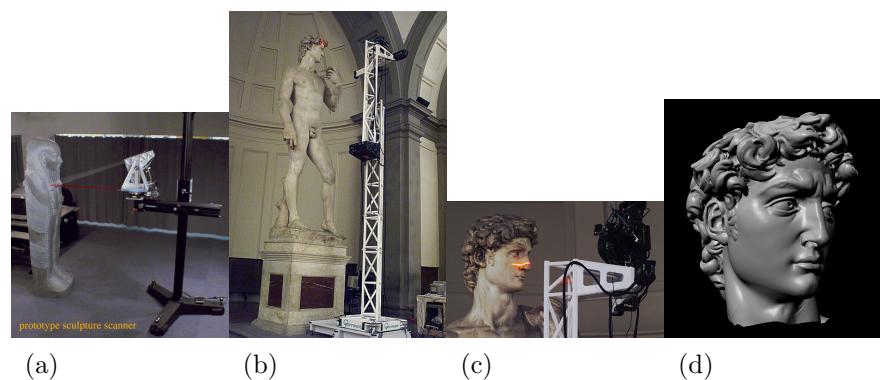


Figura 3 - Protótipo do scanner a laser de triangulação. O objeto a ser escaneado é uma réplica em tamanho real de um sarcófago egípcio (a). O scanner foi reconfigurado para escanear objetos maiores, pois a escultura possui 517 centímetros (b), o da cabeça também sofreu uma reconfiguração, este scanner gira em 90 graus, que faz o laser rotacionar, da posição horizontal para a vertical e também roda em torno da cabeça como um todo (c). Para a reconstrução, o primeiro passo foi alinhar cerca de 100 scans em diversas posicoes, após isso, utilizado um alinhamento automatico em pares dos scans, utilizando um algoritmo modificado de iteracoes de pontos próximos (ICP - *iterated-closest-points*). Após isso, faz-se um processo de relaxação global a fim de minimizar erros de alinhamento por toda a estátua. Depois de alinhados, usa-se o algoritmo de profundidade volumétrica de processamento de imagens (VRIP - *volumetric range image processing* - do Brian Curless) (d).

0.1.1 O Jardim do Nêgo, Nova Friburgo

No caso de Nova Friburgo, há a necessidade redobrada de preservação do patrimônio, em especial devido às chuvas e deslizamentos inerentes à região. O Jardim do Nêgo consiste em grandes esculturas em encostas, cobertas por um tapete de vegetação, as quais desfrutam de grande reconhecimento regional e internacional (??), figura 4.



Figura 4 - Algumas esculturas do Jardim do Nêgo

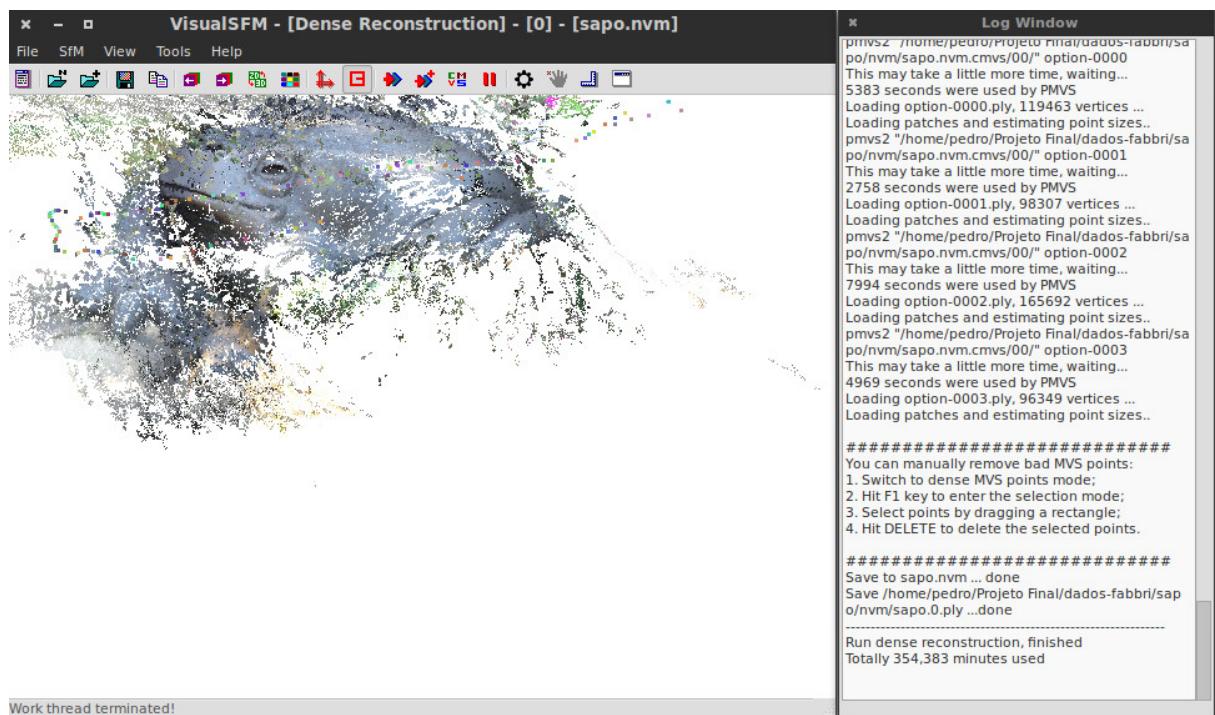


Figura 5 - A reconstrução usando-se apenas imagens, sem controle de aquisição, como em um vídeo de um smartphone filmado em torno do objeto, fornece uma nuvem de pontos, que pode ser densificada (??????????), ou atribuída de curvas (??????????), de forma a preservar a resolução em áreas de alto conteúdo informativo. Tais representações estão sendo atualmente unificadas na pesquisa da área. Este projeto propõe explorar os limites da reconstrução 3D usando-se apenas imagens, no contexto de preservação de patrimônio.

Idealizado e criado por Geraldo Simplicio (Nêgo), artista cearense que mora no

local a mais de 30 anos, ganhou notoriedade por suas esculturas de barro, com traços singulares e técnicas únicas. Hoje, trabalha para reconstruir o Jardim da tragédia de 2011 na região serrana, onde algumas estruturas foram destruídas. Portanto, com o consentimento do Nêgo, surgiu a motivação desta pesquisa: além de explorar métodos de reconstrução, também tem o objetivo de eternizar um patrimônio que é reconhecido no mundo todo.

A preservação das esculturas do Jardim do Nego se torna um desafio à pesquisa em reconstrução 3D, pois apresentam curvas bem delineadas, que são representadas de maneira suavizada e empobrecida por métodos convencionais. Algumas esculturas apresentam pouca textura, quase sem nenhum padrão de textura/musgo. Seria de grande interesse avaliar o potencial de técnicas atuais de reconstrução 3D geral sem controle de aquisição, as quais têm seu código fonte disponível na internet.

0.2 Objetivos

O presente projeto pretende fazer com que o aluno ganhe experiência com técnicas modernas de reconstrução 3D fotogramétrica, no contexto de uma aplicação bem-definida de preservação de patrimônio. A entrada do sistema deverá ser um conjunto de vídeos realizados por câmeras de baixo custo, ou um conjunto de escaneamentos realizados por scanners à mão de baixo custo baseados em Kinect.

O objetivo concreto do aluno será explorar as tecnologias supracitadas para desenvolver um esquema de escaneamento usando software aberto, câmeras e scanners de baixo custo, representando o estado da arte em reconstrução 3D sem restrições de aquisição. Perguntas fundamentais a serem respondidas são: que nível de detalhe, facilidade e precisão se pode obter usando-se apenas imagens e software aberto? É possível utilizar scanners de baixo custo baseados em Kinect com melhorias significativas em termos de qualidade, conveniência ou tempo de processamento? Quais são as restrições desses sistemas? Seria útil na prática uma reconstrução de curvas para auxiliar na reconstrução de nuvem de pontos e de superfícies densas? Onde o estado da arte deve ser avançado de forma a permitir uma solução mais conveniente e completa para a preservação de patrimônio?

O principal objetivo em termos de pesquisa científica será comparar as diferentes abordagens do estado da arte disponíveis para reconstrução 3D e explicitar suas limitações práticas. O aluno deverá, com o entendimento das abordagens, desenvolver um esquema de aquisição de esculturas que permita ampliar os detalhes ou ajudar a resolver os problemas dos métodos. Com a experiência obtida, o aluno estará pronto para desenvolver pesquisa futura na área de reconstrução 3D, com conhecimento de causa para avaliar direções de pesquisa de efetivo e alto impacto na prática.

0.3 Organização deste manuscrito

O trabalho foi estruturado da seguinte maneira: previamente introduzimos os métodos baseados em pontos de interesse no Capítulo 1, destacando suas funcionalidades. No Capítulo 2 discutimos e aprofundamos o funcionamento de cada algoritmo de reconstrução densa empregados, apresentando e debatendo, comparativamente, pontos à favor e contra; O Capítulo 3 é apresentada a técnica de reconstrução utilizada nos *Kinects*, da Microsoft; Com isso, temos o Capítulo 4, que é dedicado à ferramenta gráfica utilizada para a obtenção dos resultados (VisualSfM) dos algoritmos de reconstrução densa utilizados. Finalmente, apresentamos os resultados e conclusões do trabalho, bem como sugestões para implementações e trabalhos futuros.

1 PONTOS DE INTERESSE

A reconstrução 3D pelo método da estrutura do movimento, ou *Structure from Motion (SfM)*. Tem como base a utilização de pontos de interesse (*features*), que são pontos ou áreas em comum entre as imagens usadas na reconstrução. Para encontrar estes pontos, diferentes algoritmos são empregados.

1.1 SIFT – *Scale Invariant Feature Transform*

Primeiramente, utiliza-se o SIFT (algoritmo de detecção de pontos de interesse, invariante à escala e à transformações, como rotação, translação e iluminação da imagem, por exemplo). O algoritmo pode ser dividido em cinco etapas, das quais:

- Detecção de espaço-escala extremos – *Scale-space Extrema Detection*
- Localização de pontos-chaves – *Keypoint Localization*
- Atribuição de orientação – *Orientation Assignment*
- Descritor de pontos-chaves – *Keypoint Descriptor*
- Combinação de pontos-chaves – *Keypoint Matching*

1.1.1 Detecção de espaço-escala extremos

Em casos com cantos pequenos, a detecção funciona bem. Porém, raramente utilizaremos a mesma janela para detectar pontos-chaves em imagens com diferentes escalas, pois utilizamos imagens grandes e, consequentemente, cantos grandes. Para isso, precisamos de janelas grandes também.

Para resolver este problema, o filtro de escala-espacó é usado: o Laplaciano de Gaussiano (*Laplacian of Gaussian* – LoG). O LoG atua como um detector de partículas em diferentes tamanhos σ . (Onde σ é o parâmetro de escala). Por exemplo, o núcleo gaussiano com σ baixo, tem como resposta um alto valor para um canto pequeno. Enquanto um núcleo gaussiano com alto σ , se encaixa bem para um canto maior. Com esta lógica, podemos encontrar um máximo local através da escala e o espaço, o que nos fornece uma lista de $(x, y\sigma)$, o que significa que existe um ponto-chave em potencial, com o par (x, y) na escala σ .

Porém, como o LoG é um pouco custoso, computacionalmente. O SIFT utiliza um algoritmo aproximado do LoG, o DoG (Diferença de Gaussianos – *Difference of Gaussians*). O DoG é a diferença de um filtro Gaussiano de uma imagem, com dois valores diferentes de escala σ .

Uma aplicação prática do filtro DoG é a sequência de imagens a seguir:



Figura 6 - É aplicado um filtro gaussiano na imagem original (a), com $\sigma = 1$, tendo como resultado a imagem (b). Um outro filtro gaussiano é usado, porém, neste caso, o $\sigma = 2$ (c). Após isso, subtrai-se (b) de (c), obtendo o filtro DoG (d).

Uma vez que o DoG é aplicado, as imagens são utilizadas com o espaço e escala extremos. Por exemplo, na imagem 7, um pixel é comparado com seus 8 vizinhos, assim como comparado com os 9 pixels na próxima escala e os 9 pixels na escala anterior. Se esse pixel é um local extremo, ele é um ponto-chave em potencial. Isto é, este ponto-chave é melhor representado nesta escala.

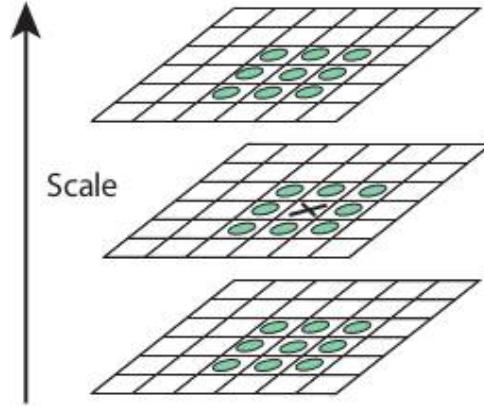


Figura 7 - Exemplo de funcionamento de detecção de espaço-escala extrema

No caso, as funções ficariam da seguinte forma:

$$g_\sigma(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{x^T x}{\sigma^2}} \quad (1)$$

$$I_\sigma = g_\sigma * I, \sigma >= 0 \quad (2)$$

$$\nabla^2 g_\sigma(x) \quad (3)$$

$$DoG_\sigma(o, s) = I_\sigma(o, s + 1) - I_\sigma(o, s) \quad (4)$$

Onde 1 é a função padrão do operador gaussiano (núcleo), a equação 3 é o operador LoG, 4 é o operador DoG e ∇^2 é o operador Laplaciano.

1.1.2 Localização de pontos-chaves

A localização dos pontos extremos pode cair em um extremo local e não global. Logo, após a utilização do DoG e com os pontos-chaves em potencial localizados, eles precisam ser refinados para melhorar o resultado. Para isso, são utilizadas Séries de Taylor na escala e no espaço, e, se a intensidade nesse extremo é menor que o valor limite, este é rejeitado. Os *frames* do SIFT (pontos-chaves) são extraídos baseados nos

extremos locais (picos) a partir do DoG. Numericamente, extremos locais são elementos que possuem um menor (ou maior) valor em uma vizinhança em um espaço 3x3x3 (em escala e espaço). Depois de extraídos, estes pontos são interpolados quadraticamente (este passo é muito importante, especialmente nas escalas de menor resolução, para ter uma localização precisa do ponto-chave na resolução completa). Finalmente, eles são filtrados para eliminar respostas de baixo contraste ou respostas próximas as bordas.

Picos que são pequenos, na maior parte das vezes são gerados a partir de ruídos e necessitam ser descartados também. Isso é feito com uma comparação de valor absoluto do DoG no pico com o valor do pico limite e é descartado caso este valor é menor que o limite.

Para eliminar respostas em bordas, normalmente os picos mais rasos ou horizontais, são gerados por bordas e não possuem características estáveis, portanto estes picos precisam ser removidos. Para isso, dado um pico (x, y, σ) , o algoritmo avalia a matriz Hessiana (x,y) do DoG na escala σ . Então é computado um valor para esta equação (5):

$$v = \frac{(T_r D(x, y, \sigma))^2}{\text{Det } D(x, y, \sigma)} \quad (5)$$

Onde, T_r é o traço, ou seja, $T_r(H) = D_{xx} + D_{yy}$ e a matriz D é do tipo

$$D = \begin{bmatrix} \frac{\partial^2 \text{DoG}}{\partial x^2} & \frac{\partial^2 \text{DoG}}{\partial x \partial y} \\ \frac{\partial^2 \text{DoG}}{\partial x \partial y} & \frac{\partial^2 \text{DoG}}{\partial y^2} \end{bmatrix}$$

No caso, v possui um valor mínimo (igual a 4) quando os autovalores da Jacobiana são iguais (pico curvado) e aumentam à medida que um dos autovalores aumenta e os outros permanecem baixos. Os picos são retidos se $v < \frac{(t_e+1)(t_e+1)}{t_e}$, onde t_e é o limite da borda.

1.1.3 Atribuição de orientação

Agora, uma orientação é atribuída a cada ponto-chave para obter a invariância à rotação da imagem. Uma vizinhança é obtida, dependente da escala, do gradiente da magnitude 6 e da direção 7 (usando diferenças finitas), ao redor da localização do ponto-chave.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (6)$$

$$\theta = \tan^{-1} \left(\frac{(L(x, y + 1) - L(x, y - 1))}{(L(x + 1, y) - L(x - 1, y))} \right) \quad (7)$$

Então, um histograma de 36 orientações (*bins*) cobrindo 360 graus é criado. Onde ele é ponderado pelo gradiente da magnitude e por uma janela Gaussiana circular onde σ vale 1.5 em relação à escala do ponto-chave 8.

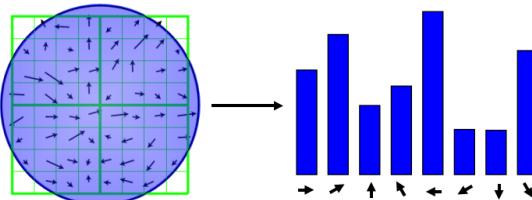


Figura 8 - Exemplo do resultado obtido do histograma orientado

O ponto mais alto do histograma é obtido e qualquer pico acima de 80% é considerado no cálculo da orientação. Pontos-chaves são criados com a mesma localização e escala, mas em diferentes direções, o que contribui para a estabilidade da correspondência.

1.1.4 Descriptor de pontos-chaves

Com os pontos-chaves criados a partir do histograma orientado, cria-se agora o descriptor de pontos-chaves.

Uma vizinhança 16x16 ao redor do ponto-chave é escolhida e esta mesma vizinhança é dividida em 16 sub-blocos 4x4. Para cada bloco, um histograma orientado com 8 *bin* é criado. Logo, temos 128 valores válidos de *bin*. Esses valores são representados em forma de vetor para expressar o descriptor de pontos-chaves 9.

Além disso, são tomadas algumas medidas para deixar o descriptor mais robusto, como, por exemplo, invariante à luminosidade, rotação, etc.

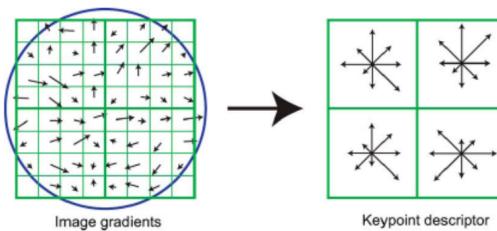


Figura 9 - Exemplo de um descriptor de pontos-chaves, com uma matriz 2x2 e uma região 8x8

1.1.5 Combinação de pontos-chaves

Pontos-chaves entre duas imagens são combinados a partir da identificação da vizinhança mais próxima. Mas, em alguns casos, a segunda combinação mais próxima pode ser parecida com a primeira. Isso se dá por ruídos presentes nas imagens ou algo assim.

Nesse caso, a razão da distância mais próxima para a segunda distância mais próxima é utilizada. Se essa razão for maior que 0.8, essa combinação é descartada. Esse método elimina cerca de 90% de combinações falsas, enquanto descarta apenas cerca de 5% de combinações corretas.

1.2 Triangulação – *Full pairwise image matching*

Com os pontos de interesse (*features*) extraídos, podemos agora fazer a triangulação entre os pontos das imagens.

A triangulação nada mais é que uma estimativa de um ponto em 3 dimensões, dado pelo menos duas câmeras conhecidas, onde, cada câmera com a projeção do *feature* correspondente àquele ponto 3D 10.

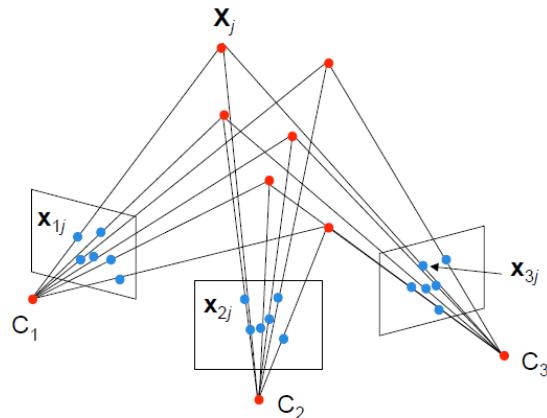


Figura 10 - Uma triangulação utilizando um ponto qualquer, X_j . Onde cada câmera C_1, C_2, C_3 possui um *feature* correspondente a cada uma delas, respectivamente, X_{1j}, X_{2j}, X_{3j} .

Infelizmente, não é tão simples assim. Existem muitos fatores que contribuem para aumentar a dificuldade da triangulação: ruídos, posição das câmeras, o feixe das projeções não se encontram no mesmo ponto 3D, não se tem informação da projeções nas câmeras, etc. Entretanto, existem diversos algoritmos para resolução de cada um dos problemas enfrentados.

Como o foco deste manuscrito está no *software* de reconstrução VisualSfM, atentamos apenas ao algoritmo utilizado pelo mesmo. Neste caso, ele utiliza um algoritmo

chamado *Bundle Adjustment*.

Este algoritmo é utilizado para minimizar o erro geométrico proveniente da reprojeção da triangulação 11. Dentro dele, existe a técnica de resolução de problemas não-lineares chamada Levenberg-Marquardt.

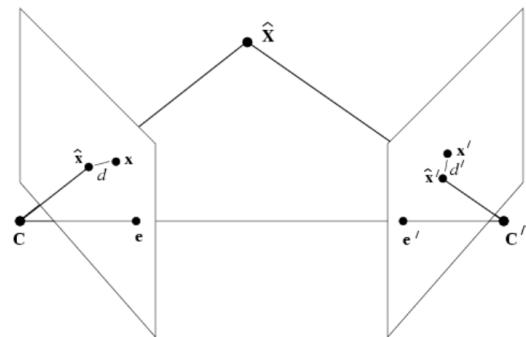


Figura 11 - Erro proveniente da reprojeção, onde os pontos x e x' estão mais próximos das medidas reais da imagem.

2 RECONSTRUÇÃO DENSA

2.1 Introdução

2.2 HPMVS

2.2.1 falar sobre o hpmvs...

2.3 MVE

Um dos algoritmos utilizados para a técnica de reconstrução densa é o MVE – *Multi-View Environment*, feito por Simon Fuhrmann, Fabian Langguth e Michael Goesele. Este algoritmo utiliza fotos e produz uma malha triangular superficial como resultado. Diferente das reconstruções baseadas nas geometria das imagens, o MVE é focado na reconstrução multi-escala, um quesito importante na reconstrução de esculturas e acervo cultural. Portanto, com esta técnica é possível reconstruir grandes volumes de dados, contendo regiões detalhadas em alta resolução, em comparação com o resto da cena. O sistema ainda possui uma interface gráfica para o uma reconstrução baseada no SfM, amigável ao usuário (UMVE), onde permite a visualização e inspeção das imagens, mapas de profundidade e renderizar cenas e malhas 3D.

Sua base de operação é basicamente:

1. Estrutura da formação – *Structure-from-Motion* (SfM)

- Reconstrói os parâmetros da câmera (posição e orientação) e seus dados de calibração (distância focal e distorção radial), encontrando correspondências esparsas mas estáveis entre as imagens. (Já foi abordado em outra seção deste manuscrito).

2. Múltiplas visões estéreo – *Multi-View Stereo* (MVS)

- Utiliza a posição estimada das câmeras, encontrando as correspondências visuais nas imagens. Estas correspondências são trianguladas, produzindo a informação 3D, e, consequentemente a reconstrução 3D densa.

3. Reconstrução de superfícies – Surface Reconstruction

- Tem como entrada uma densa nuvem de pontos, ou mapas de profundidade individuais. Produz uma malha superficial globalmente consistente.

Como não existem muitas opções para algoritmos de SfM, o MVE permite a utilização de *softwares* externos como o *Bundler* ou o próprio *VisualSfM*.

Uma vez com o passo do SfM feito, partimos para o MVS. Com os parâmetros de câmera conhecidos, a reconstrução densa geométrica é feita. Existem diversos algoritmos para a reconstrução densa, o MVE no caso, utiliza um algoritmo próprio, feito por um de seus criadores, Michael Goesele (*Multi-View Stereo for Community Photo Collections approach*), que reconstrói um mapa de profundidade para cada foto.

Embora abordagens baseadas em mapeamentos de profundidade produzirem uma grande quantidade de redundância, (isso se dá por causa das inúmeras fotos que são sobrepostas e possuírem partes similares da mesma cena), este algoritmo é altamente escalável para grandes cenas, pois apenas um pequeno conjunto de fotos vizinhas é necessário para a reconstrução. Outra vantagem da utilização dos mapas de profundidade como representação intermediária é que a geometria é parametrizada em seu domínio natural, e os dados por foto (como a cor, por exemplo) estão diretamente acessíveis nas imagens.

A redundância excessiva nos mapas de profundidade pode ser pesada. Não com relação ao armazenamento, mas na questão do processamento computacional exigido nos mapas de profundidade. Porém, esta abordagem foi capaz de produzir uma geometria detalhada e superar o ruído nos mapas de profundidades individuais.

2.3.1 Guia de reconstrução com o MVE

Tirando fotos: Um bom conjunto de dados é gerado se algumas regras simples forem seguidas:

- Para que o algoritmo do MVS consiga fazer uma triangulação com qualquer posição 3D, o conjunto de dados terá que ter, no mínimo, cinco fotos.
- As fotos devem ser tiradas com uma boa quantidade de sobreposição. A menos que o conjunto de dados se torne muito grande, uma grande quantidade de fotos não prejudicará a qualidade. Mas terá uma compensação do sistema, no que diz respeito à qualidade e desempenho.
- Para a triangulação funcionar, é necessário que tenha o efeito de paralaxe ?? (Aparente mudança na posição do objeto). Ou seja, é interessante que o conjunto de imagens seja duplicado.
- A câmera deverá ser reposicionada, de preferência.

Criando uma cena: Uma visualização contém dados por exibição (como imagens, mapas de profundidade ou outros dados). Uma cena é uma coleção de visualizações, que

constitui um conjunto de dados. Uma nova cena pode ser criada utilizando a interface gráfica UMVE, ou por linha de comando (*makescene*).

Tecnicamente, a cena é criada como um diretório no sistema de arquivos (com o nome do conjunto de dados). Este, por sua vez, contém outro diretório (*views*), com todas as visualizações guardadas com uma extensão de arquivos em .MVE.

Criar uma nova cena, criará apenas o diretório (*views*) vazio. A importação de fotos criará arquivos .MVE para cada foto. Esse processo importará meta-dados provenientes das imagens (*tags EXIF*), que é necessário para estimar a distância focal para cada foto. Caso estes meta-dados não estejam disponíveis, uma distância focal padrão é assumida pelo sistema, porém se essa distância adotada for uma péssima suposição, com relação ao conjunto de dados utilizado, pode vir a acontecer erros no SfM.

Reconstrução SfM: Pode ser configurada e iniciada usando a interface gráfica UMVE ou por linha de comando (*sfmrecon*). A interface guia através da detecção de *features*, combinação emparelhada (*pairwise matching*) e uso incremental do SfM. Que, por sua vez, a reconstrução SfM começa a partir de um par inicial, e adiciona, de forma incremental, mais vistas à reconstrução.

Múltiplas visões estéreo – *Multi-View Stereo* (MVS): Usando as imagens junto com os parâmetros obtidos das câmeras, é possível reconstruir a geometria densa utilizando o MVS. Isso pode ser feito utilizando a interface gráfica (UMVE) ou por linha de comando (*dmrecon*).

O parâmetro mais importante é o nível de resolução em que os mapas de profundidade são reconstruídos: Caso seja nível 0 (ou L0), a reconstrução é feita usando o tamanho original das imagens. Se for nível 1 (ou L1), a reconstrução corresponde a metade do tamanho (um quarto dos números de pixels), e assim por diante.

Com a resolução das câmeras atuais, uma reconstrução L0 é raramente usada, pois geram mapas de profundidade mais dispersos com um custo computacional elevado, o que acarreta em dificuldades para encontrar as correspondências densas das imagens. Geralmente utiliza-se o L2, pois o processo é mais rápido, gerando mapas de profundidades completos, já que utiliza imagens menores.

Reconstrução de superfícies – Surface Reconstruction: Utiliza-se a linha de comando *scene2pet*, que combina todos os mapas de profundidade em uma única e grande nuvem de pontos. Nesta fase, um valor de escala é atribuído a cada ponto, que indica o tamanho atual da região da superfície na qual o ponto foi mensurado. Esta informação adicional permite o uso de várias propriedades benéficas usando a abordagem de reconstrução da superfície FSSR. A seguir, as ferramentas FSSR calculam uma representação volumétrica de escala múltipla a partir dos pontos (na qual não precisa de nenhum ajuste de parâmetros explícitos) e uma malha final é extraída. Esta malha pode parecer desordenada devido à regiões não confiáveis e à componentes isolados, oriundos de medidas imprecisas. Logo, a malha é limpa, retirando pequenos componentes isolados e regiões

não confiáveis da superfície.

Experiência com o MVE: A utilização do *software* é bem intuitiva, seja por linha de comando ou pela interface gráfica (neste modo, fica mais fácil visualizar cada etapa da reconstrução). Amplamente configurável, podendo escolher a vizinhança, escala, manter o mapa de profundidade, ver os dados *EXIF* de cada imagem, por exemplo.

Porém, para a aplicação proposta neste projeto, não é muito interessante, visto que ele utiliza a informação das câmeras e das imagens e como as imagens usadas são, tecnicamente, vídeos cortados em determinados *frames*, não é possível obter a informação das câmeras, logo o *software* não tem tanta aplicabilidade neste caso, a menos que sejam tiradas fotos sequenciais de alguma escultura ou objeto que se deseja gerar a reconstrução densa, pois dessa forma, as informações necessárias das câmeras estarão armazenadas.

3 KINECT

4 VISUALSFM

5 EXPERIMENTOS

CONCLUSÃO