RONALDO DOS SANTOS ALBURNIO

MÉTODOS PASSIVOS DE RECONSTRUÇÃO 3D VOLTADOS À PRESERVAÇÃO DIGITAL DE ACERVOS CULTURAIS

CURITIBA

2012

RONALDO DOS SANTOS ALBURNIO

MÉTODOS PASSIVOS DE RECONSTRUÇÃO 3D VOLTADOS À PRESERVAÇÃO DIGITAL DE ACERVOS CULTURAIS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre. Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, Universidade Federal do Paraná.
Orientador: Prof. Dr. Luciano Silva
Co-Orientadora: Prof. Dra. Olga R. P. Bellon

CURITIBA

2012

Ministério da Educação
Universidade Federal do Paraná
Programa de Pós-Graduação em Informática

PARECER

Nós, abaixo assinados, membros da Banca Examinadora da defesa de Dissertação de Mestrado em Informática, do aluno Ronaldo dos Santos Alburnio, avaliamos o trabalho intitulado, *"Métodos passivos de reconstrução 3d voltados à preservação digital de acervos culturais"*, cuja defesa foi realizada no dia 24 de fevereiro de 2012, às 14:00 horas, no Departamento de Informática do Setor de Ciências Exatas da Universidade Federal do Paraná. Após a avaliação, decidimos pela aprovação do candidato.

Curitiba, 24 de fevereiro de 2012.

Prof. Dr. Luciano Silva
DINF/UFPR – Orientador

Prof. Dr. Paulo Urnau Gotardo
OSU/USA – Membro Externo

Profa. Dra. Olga Regina Pereira Bellon
DINF/UFPR – Membro Interno

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

AGRADECIMENTOS

Agradeço à CAPES que financiou este mestrado.

Aos professores Luciano Silva e Olga R. P. Bellon pela orientação e ideias ao longo desses anos.

Ao grupo IMAGO pelo ambiente propício à pesquisa e repleto de pessoas competentes e dispostas à ajudar.

Aos meus amigos pelo companheirismo e paciência ao me permitir divagar sobre problemas específicos enfrentados em cada etapa do trabalho, e especialmente a Leonardo Gomes e Leandro H. S. Yorinori pela ajuda na execução dos testes.

À minha família pelo suporte e amor incondicional.

CONTEÚDO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

A maioria dos métodos de aquisição que garantem a melhor reconstrução 3D (tridimensional) utilizam-se de *scanners* de profundidade, também conhecidos como *range scanners*. Tais métodos são chamados de métodos óticos ativos, nos quais a informação da superfície é obtida a partir da projeção controlada da iluminação incidente no objeto. Estas, porém, são soluções caras, principalmente devido ao alto preço do hardware envolvido no processo. A partir daí, surge a necessidade do desenvolvimento dos métodos conhecidos como métodos óticos passivos, que obtém a informação 3D a partir da análise de imagens de intensidade e resultam na profundidade absoluta da cena ou nas orientações de superfície do objeto em questão. Tratam-se de métodos normalmente com custo inferior, porém que nem sempre resultam em uma reconstrução tão precisa quanto a obtida com os métodos ativos. Há várias abordagens possíveis para este problema, das quais podemos citar: *photometric stereo* e *shape from shading*; *shape from texture*; *shape from focus/defocus*; e o *shape from motion*. Existem também métodos que combinam mais de uma abordagem de forma complementar, como *motion* e iluminação, por exemplo, tentando suprir mutuamente algumas deficiências. Todos esses métodos supracitados, usados como alternativa aos *range scanners*, têm suas vantagens e desvantagens para a reconstrução 3D, dependendo muito das características do objeto que se deseja reconstruir e da qualidade das imagens capturadas, sendo esses fatores cruciais na precisão e robustez das reconstruções 3D. A proposta deste trabalho é a de analisar métodos passivos de reconstrução para a posterior escolha do método mais adequado para a reconstrução 3D da maior gama possível de objetos relacionados à preservação digital, de forma a evitar o uso de *range scanners*, tanto pelo seu elevado custo quanto pela ineficiência de tais dispositivos para alguns tipos de objetos específicos (devido às propriedades das superfícies ou do tamanho do objeto). Um método híbrido que combina informações de movimentação do objeto com a reflexão de sua superfície é testado e verificamos que os resultados são interessantes,

apesar de haver ainda problemas a serem corrigidos para uma reconstrução de superfícies mais generalizada e acurada.

ABSTRACT

The acquisition methods that currently ensure best 3D (tridimensional) reconstruction make use of range scanners. Such methods are known as active optical methods, in which the surface information is obtained through a controlled projection of the light incident on the object. However, such solutions are expensive, mostly because of high price of the hardware designed for that purpose. Thereafter, the need of development of passive optical methods arises, which recovers the 3D information through the analysis of 2D intensity images resulting in the absolut depth of the scene or object's surface orientation. Passive methods are usually cheaper but not always result in a reconstruction as accurate as that obtained by active methods. There are several different approaches to settle this problem, among which: *photometric stereo* and *shape from shading*; *shape from texture*; *shape from focus/defocus*; and *shape from motion*. There are also methods that combine more than one single approach to improve the reconstruction accuracy, like motion and illumination, for example, trying to mutually overcome some shortcomings. All these aforementioned methods used as alternative to range scanners have their positive and negative points and the results are quite dependent on the object surface characteristics and can be affected by noise, decreasing the 3D reconstruction final quality and points accuracy. This work has as main objective to find the best passive method for 3D reconstruction that covers most different classes of objects related to digital preservation, to avoid the usage of range scanners because of its usually high cost or when they are not very suitable (due to object surface properties or object size). A hybrid method that combines the motion of the object and its surface reflectance is tested resulting in a good surface, although there are yet problems to be addressed for a more generalized and accurate reconstruction.

CAPÍTULO 1

INTRODUÇÃO

A construção realística de modelos digitais 3D de objetos físicos é um problema complexo, com grandes desafios ainda a serem vencidos [37, 57]. No entanto, avanços tecnológicos e científicos relacionados a reconstrução 3D têm sido observados nos últimos anos, proporcionando um crescimento significativo no número de projetos em diversas áreas de aplicação.

Recentemente, projetos para a preservação digital de acervos culturais [41, 73, 81] têm gerado uma demanda crescente por soluções computacionais eficientes e práticas para a reconstrução 3D de objetos físicos. Outro exemplo são os projetos voltados ao estudo da biodiversidade, nos quais coleções biológicas vêm sendo digitalizadas para fins de pesquisa, documentação e arquivamento seguro [58]. Estas iniciativas vêm ganhando popularidade e têm despertado interesse não somente da comunidade acadêmica, em especial das áreas de visão computacional, processamento de imagens e computação gráfica, como também das agências de fomento à pesquisa e entidades governamentais.

No processo de geração de modelos 3D de objetos físicos usualmente são empregados métodos ativos de sensores de profundidade, conhecidos como *range scanners* [11], capazes de obter a informação 3D das superfícies dos objetos observadas a partir de um ponto de vista. As imagens fornecidas por estes equipamentos são chamadas imagens de profundidade (*range images*). Os *scanners* mais comuns se valem de duas estratégias básicas para obterem informação de profundidade: tempo de voo, na qual a distância do sensor até a superfície do objeto é medida a partir do tempo gasto até que um pulso de laser emitido sobre a superfície seja visto pelo sensor; e triangulação, baseada na possibilidade de recuperar a informação de profundidade através da diferença da projeção de um laser a uma distância conhecida do sensor em uma superfície. Independentemente da

estratégia utilizada, um objeto físico não pode ser completamente observado com uma única imagem, fazendo-se necessárias então múltiplas imagens de diferentes vistas para a reconstrução de seu modelo 3D completo. O processo de integração dessas imagens para a reconstrução 3D do objeto envolve diversas etapas em um *pipeline* [81], como por exemplo o alinhamento (ou registro) das imagens das diversas vistas do objeto [69, 70, 71].

Atualmente os *scanners laser 3D* apresentam boa precisão (0.1 mm entre pontos amostrados) para a digitalização 3D de objetos de médio porte, com tamanhos entre 10 e 100 cm, como, por exemplo, esculturas e artefatos arqueológicos. Embora seja possível digitalizar uma estátua de 10 m de altura com um *scanner laser 3D*, o processo de captura e de integração das diversas imagens (geralmente cobrindo 50x50 cm do modelo aproximadamente) é bastante custoso como discutido por Levoy et al. em [46]. De fato, objetos de grande porte são um desafio a parte, e geralmente são empregados equipamentos da área de geologia [57] capazes de adquirir vistas com maior amplitude de campo, porém com menor precisão (1 cm entre pontos amostrados).

Vários projetos de preservação digital se utilizam de métodos ativos na reconstrução, como é o caso da preservação de antigas tabuletas de escrita cuneiforme usando *scanners* de triangulação [42], bem como para a reconstrução do Grande Buda no Japão [39] e das estátuas de Michelangelo [46], como Davi, em Florença, na Itália. Na reconstrução de relíquias escavadas na China, como visto em [89], os *range scanners* também foram escolhidos para o processo de reconstrução. Em alguns casos esses *scanners* acabam sendo usados em apenas algumas partes do processo, como para a reconstrução das áreas próximas à beirada do modelo da Roma Antiga, exposto no *Museo Della Civilità Romana*, em Roma, devido à dificuldade de utilizá-los nas partes centrais do modelo, por este ser demasiadamente grande (16x16 m), valendo-se de um *laser radar*, uma espécie de radar usando laser infravermelho de alta frequência, para a reconstrução da maior parte do modelo [26]. Já na reconstrução da obra em madeira de Donatello, Madalena, foi empregado um método ativo baseado em triangulação um pouco diferente, usando luz estruturada, ou seja, um padrão conhecido projetado na peça cuja deformação possibilita

a recuperação de uma nuvem de pontos [27]. Também na reconstrução da Pietà de Florentina, de Michelangelo, foi desenvolvido um sistema que usa luz estruturada na reconstrução com o auxílio de um algoritmo *stereo* baseado em várias vistas (*multi-view*) [10].

Com os avanços tecnológicos recentes nesta área, diversos sensores de imageamento têm sido lançados no mercado, com novas funcionalidades para capturar imagens com maior precisão, baixo nível de ruído e em menor tempo. Além disso, o custo para a aquisição destes equipamentos vem sendo reduzido consideravelmente nos últimos anos, trazendo oportunidades para o surgimento de novas aplicações, embora estes custos sejam ainda proibitivos em grande parte.

Para alguns tipos de objetos, como os de dimensões reduzidas, os *scanners laser 3D* não são adequados, pois superfícies muito pequenas são dificilmente digitalizadas com precisão devido a resolução do equipamento, problemas de reflexão ou difração do feixe de laser, ruído etc [81]. Neste caso, há a possibilidade do uso de equipamentos com agulha de toque (*touch probe scanners*) que possuem maior precisão, porém podem danificar o objeto.

Como alternativa de baixo custo para a reconstrução 3D, pode-se explorar técnicas de obtenção da informação 3D através da análise de características em imagens 2D adquiridas com uma câmera digital (imagens de intensidade), conhecidas como técnicas passivas. Além do baixo custo, esses métodos muitas vezes não exigem equipamentos especiais, ocupando menos espaço e facilitando seu transporte, dando mais liberdade durante a captura dos dados e algumas vezes até mesmo possibilitando a reconstrução de objetos que estão disponíveis apenas em fotos ou vídeos. O método mais adequado é escolhido normalmente levando-se em consideração características e limitações dos objetos que se deseja reconstruir.

Com o desenvolvimento de sistemas robustos e soluções alternativas de baixo custo para geração de modelos 3D abrem-se oportunidades de pesquisas multidisciplinares e inovadoras, com o intuito da preservação de patrimônios naturais e culturais, segurança,

educação, entretenimento e cultura.

Vários métodos de reconstrução vêm sendo estudados já há bastante tempo baseados em várias características diferentes das imagens. O *photometric stereo* é o processo no qual a orientação da superfície é encontrada a partir da observação da cena sob diferentes condições de iluminação, sendo o *shape from shading* um caso especial que se baseia na análise de apenas uma imagem. Em *shape from texture*, informações como tamanho e intensidade ou compressão espacial da textura são levadas em consideração para se descobrir a orientação da superfície. O *shape from focus* se apoia na variação de foco da câmera para descobrir a profundidade do objeto. *Shape from silhouette* combina silhuetas do objeto, segmentado do resto da cena, de forma a se obter um modelo sólido do objeto real que resulta na mesma silhueta para cada possível ponto de vista, chamado *visual hull*. O método de *shape from motion* baseia-se no uso de mais de uma câmera fotográfica, ou mesmo vários quadros da cena tirados em diferentes posições, para a recuperação de uma nuvem de pontos através da movimentação do objeto entre os quadros. Uma abordagem similar é a da visão estéreo (*stereo vision*), que analisa a disparidade entre duas imagens para a recuperação da profundidade.

Alguns métodos desenvolvidos especialmente para fins de preservação digital são facilmente encontrados, como em [19, 1], que usam *stereo vision* e *shape from motion*, junto de outras técnicas, para a reconstrução 3D de cenas arquiteturais. Em [63] encontramos um método que pode ser usado para a reconstrução de artefatos arqueológicos através de *shape from motion* a partir de uma sequência de vídeo de uma câmera não calibrada. O tempo de execução de tais algoritmos também passa a ser uma preocupação e soluções são sugeridas, como por Agarwal et al. [3], que desenvolve um sistema de *shape from motion* para coleções gigantescas de imagens retiradas da internet de forma a paralelizar tanto a busca por correspondências entre as imagens como a reconstrução da cena em si de forma a otimizar ao máximo o processo.

Dada a importância da reconstrução de modelos 3D de objetos reais e a abundância de alternativas, os objetivos principais deste trabalho são: (1) realizar um estudo dos

métodos passivos para aquisição e reconstrução 3D de objetos e (2) selecionar os métodos mais adequados à reconstrução de objetos de acervos culturais.

Nos próximos capítulos, serão explicadas algumas estratégias usadas nos métodos de reconstrução passivos, bem como serão citadas suas vantagens e restrições.

CAPÍTULO 2

MÉTODOS PASSIVOS DE RECONSTRUÇÃO

2.1 Shape from Motion

A ideia geral do *shape from motion* ou da visão estéreo é a de, a partir das imagens da cena projetadas em 2D, recuperar a posição da câmera com relação a um sistema de coordenadas e reconstruir a estrutura 3D do objeto em questão [66, 17, 64]. A visão estéreo é feita através de câmeras calibradas e a relação entre a projeção das informações de intensidade em cada cena capturada resulta na profundidade dos pontos.

A calibração de câmeras é o processo de se obter os parâmetros da câmera, tanto intrínsecos (internos à câmera) quanto extrínsecos (orientação da câmera com relação a um sistema de coordenadas) e utilizá-los na recuperação das informações 3D da cena. Dois métodos de calibração bem aceitos na comunidade de visão computacional são os propostos por Tsai [79] em 1987 e Zhang [88] em 1998, que assumem projeção em perspectiva (objetos mais distantes aparecem menores) e usam várias imagens, sem distorção, de um tabuleiro para, através dos pontos rastreados no plano, recuperar os parâmetros das câmeras antes da captura das imagens do objeto que se deseja reconstruir. Os resultados podem ser bem apurados e completos quando utilizamos câmeras calibradas, porém esse modelo de calibração nem sempre é passível de ser aplicado (quando usamos imagens previamente capturadas sem nenhum tipo de controle especial, por exemplo) e alguns métodos procuram calibrar as câmeras a partir das imagens já capturadas, sem o auxílio de objetos conhecidos presentes na cena [2].

Métodos de *shape from motion* serão tratados em maiores detalhes no próximo capítulo.

2.2 Shape from Shading e Photometric Stereo

A direção e intensidade das fontes de iluminação em uma cena são fatores importantes na formação da imagem, fazendo-se possível, a partir de tais informações, se estimar a orientação da superfície do objeto iluminado na cena. Para muitas superfícies a quantidade de luz refletida em uma certa direção depende somente da orientação da superfície, como podemos ver na *Figura 2.1.*

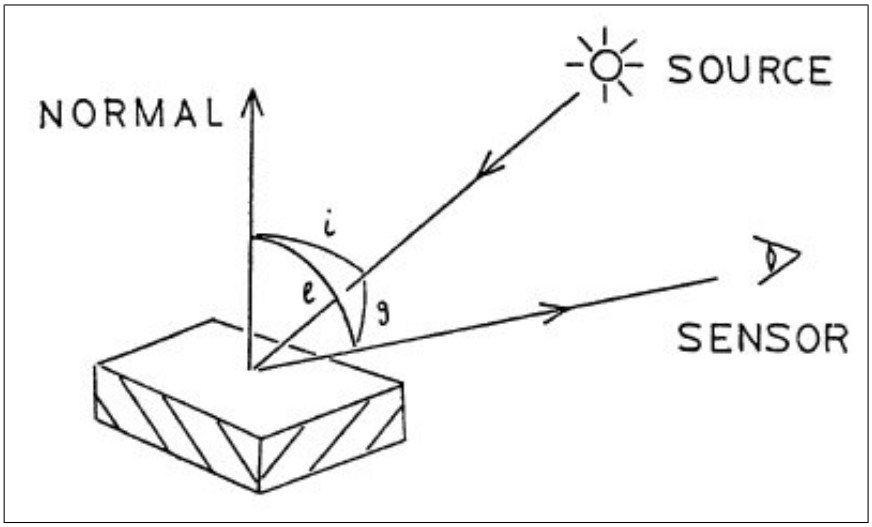

Figura 2.1: *O ângulo incidente i e o ângulo emergente e são definidos com relação à normal local da superfície. (fonte [35])*

O aparecimento de técnicas de reconstrução baseadas em iluminação remonta ao início dos anos 70, com Horn [35] desenvolvendo um trabalho completo em *shape from shading (SfS)*, técnica que consiste no uso de uma única imagem, dada a posição das fontes de iluminação e o uso de fotometria na superfície do objeto, para a recuperação de sua forma. Desde o aparecimento dessa técnica, muitas diferentes abordagens surgiram [86, 20]. Apesar da existência de vários modelos de refletividade, o modelo lambertiano é assumido na maioria das estratégias de *SfS*. Trata-se de um modelo simples, no qual os níveis de cinza em um pixel dependem da direção de origem da luz e do vetor normal da superfície, sendo que um determinado ponto tem a mesma radiância independentemente do ângulo em que está sendo visualizado - o que verificamos que nem sempre é verdade ao tomarmos como base objetos reais. Mesmo assumindo este modelo de refletividade, o

problema ainda não é simples, exigindo restrições adicionais para se obter uma solução única para o problema de *SfS*.

As técnicas de *SfS* se dividem em quatro grupos: abordagem de minimização, que procura minimizar uma função de energia baseada em algumas restrições de igualdade entre o brilho da superfície e da imagem formada, e de suavidade da superfície [38, 13] ou de similaridade de gradiente na vizinhança de cada pixel [90]; abordagem de propagação, baseada em faixas de características que são propagadas ao redor dos pontos de máxima intensidade [35]; abordagem local, em que a forma do objeto é recuperada a partir da intensidade e sua primeira e segunda derivadas, assumindo uma superfície localmente esférica em cada ponto [59, 43]; e abordagem linear, usando aproximação linear da função de refletividade com relação ao gradiente da superfície [61, 78]. Cada grupo tem vantagens e desvantagens para determinados objetos, mas no geral aqueles baseados em uma abordagem de minimização costumam ser mais robustos, enquanto os outros são, geralmente, mais rápidos.

Para casos gerais, técnicas de *shape from shading* não garantem solução única. Além disso, métodos de *SfS* não lidam com interreflexão, ou seja, com o brilho de uma face do objeto sendo refletido em outras de suas faces, o que ocorre com frequência no mundo real. Soluções para esses problemas passam a ser endereçadas a partir do surgimento do *photometric stereo*, no final dos anos 70 e início dos anos 80 por Woodham [82]. A ideia do *photometric stereo* habilita as vantagens de se poder utilizar informações de várias imagens sem ter de se preocupar com correspondência entre os pontos, já que a câmera se mantém estática enquanto a direção da luz incidente varia. Segundo o modelo lambertiano de reflexão temos

$$I_1 \;=\; n \cdot l_1,$$

sendo I a intensidade da imagem em um ponto, n a normal da superfície no ponto e l a direção da luz. Em [82] é mostrado que a obtenção de pelo menos mais duas imagens com a iluminação variando de direção

$$I_2 = n \cdot l_2$$
$$I_3 = n \cdot l_3$$

nos permite determinar tanto a orientação da superfície de forma única quanto o fator de refletividade em cada ponto, desde que se conheça a direção da fonte de iluminação.

Muitas variações de *photometric stereo* vêm sendo estudadas desde seu surgimento, tanto na área teórica quanto prática, partindo das mais diversas restrições e limitações de cena e superfície do objeto e usando diferentes estratégias. Ainda em 1994 Hayakawa surge com uma abordagem baseada em fatoração da matriz da imagem [29], usando *SVD*, em duas matrizes S de normais da superfície e L de direção e intensidade da fonte de iluminação (sem a necessidade de se saber essas informações a priori), apoiado no teorema do rank, de forma muito parecida com a proposta por Tomasi e Kanade para o problema de *shape from motion* [76] (método de *SfM* detalhado no próximo capítulo). Essa abordagem admite projeção ortográfica da cena e permite a recuperação da forma baseada em imagens tiradas em um ambiente não controlado. Basri e Jacobs, em 2001 [8], usa uma abordagem parecida, também admitindo projeção ortográfica e que permite a reconstrução 3D de objetos fora de laboratório, com as únicas restrições de que todas as fontes de iluminação sejam isotrópicas (ilumina de forma igual em todas as direções) e distantes do objeto. Há uma certa ambiguidade embutida no problema de *SfS* em ambientes não controlados e considerações extras devem ser feitas para a recuperação da superfície de forma única.

Hertzmann e Seitz em [33] usam uma abordagem de calibração do sistema através de objetos de geometria conhecida com material semelhante ao do objeto que se deseja reconstruir para reduzir as limitações do método. Com a câmera estática, admitindo projeção ortográfica, a iluminação varia em cada imagem e o objeto de calibração, presente na cena, serve como referência. Esta abordagem, além permitir qualquer número de fontes de iluminação na cena sem que isso atrapalhe na reconstrução, também faz a segmentação do objeto em diferentes materiais, caso necessário, e o resultado obtido é extremamente acurado. Apesar do método possuir também algumas desvantagens, como não considerar

sombras projetadas no objeto e interreflexão, o resultado da superfície obtida é comparável com aquele obtido usando *scanners laser*, como vemos na *Figura 2.2*.

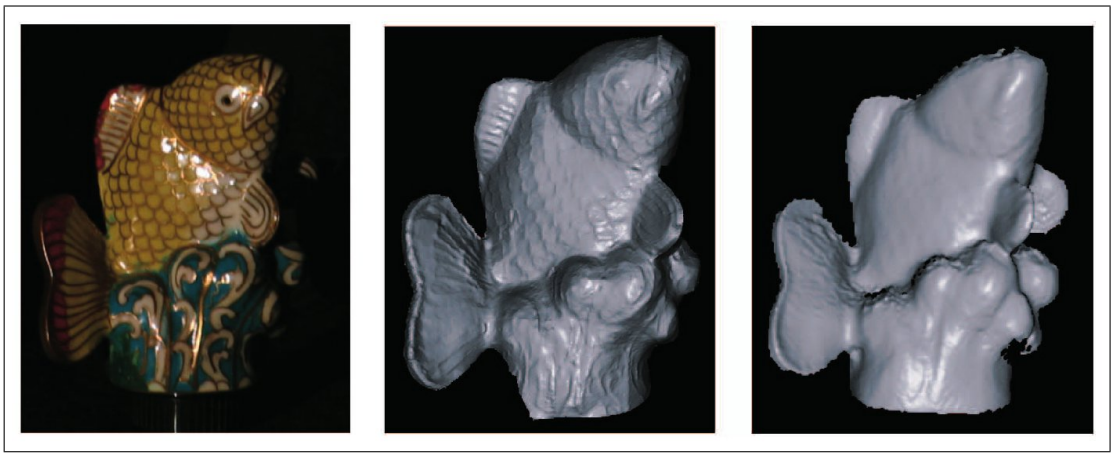

Figura 2.2: *Esquerda: objeto real; meio: reconstrução usando o método de Hertzmann e Seitz; direita: reconstrução usando scanner laser (o objeto teve de ser pintado com tinta difusa para poder ser escaneado). (fonte [33])*

Existem também abordagens que exigem pré-calibração, em que um objeto é usado para calibrar o sistema antes de se capturar a sequência do objeto que se deseja reconstruir. Em [31], Hernández et al. nos apresenta o método conhecido como *multispectral photometric stereo*. Nesse método, fontes de iluminação de cores diferentes (R, G e B) e distantes entre si são usadas na iluminação da cena, possibilitando assim a filtragem e separação de cada espectro de cor para a reconstrução da cena em cada quadro da sequência. Esse método de *photometric stereo* usa um objeto de calibração antes do início da captura e habilita a reconstrução de objetos deformáveis. Uma outra abordagem de calibração, chamada auto-calibração, descarta a necessidade do uso de outros objetos no processo. Em cada quadro, cada ponto é analisado de forma a se descobrir uma função de resposta radiométrica que mapeia a irradiância em um pixel através da análise das cores que esse pixel assume em cada imagem (o que limita essa abordagem para cenas ou imagens coloridas). Além disso, a ambiguidade das normais da superfície é eliminada a partir da análise dos pontos em que as normais são diferentes apesar de apresentarem a mesma refletividade, como mostrado por Shi et al. [67].

2.3 Shape from Texture

Apesar de sermos capazes de deduzir informações da forma de um objeto por combinar informações como a disparidade entre as imagens vistas por cada olho, a movimentação, as sombras, o contorno e a textura, é sabido que conseguimos reconhecer puramente a distorção da textura de um objeto 3D texturizado quando projetado em uma imagem 2D (*Figura 2.3*). Os métodos de *shape from texture* se focam, então, na distorção das texturas, normalmente quando o padrão da textura já é conhecido, para recuperar as informações de profundidade da cena. Porém, um número não muito grande de métodos foi desenvolvido para a resolução deste problema, e são divididos basicamente em abordagem global, na qual a textura é analisada como um todo, e abordagem local, baseada na recuperação dos parâmetros geométricos em pontos da superfície, como descrito em [23].

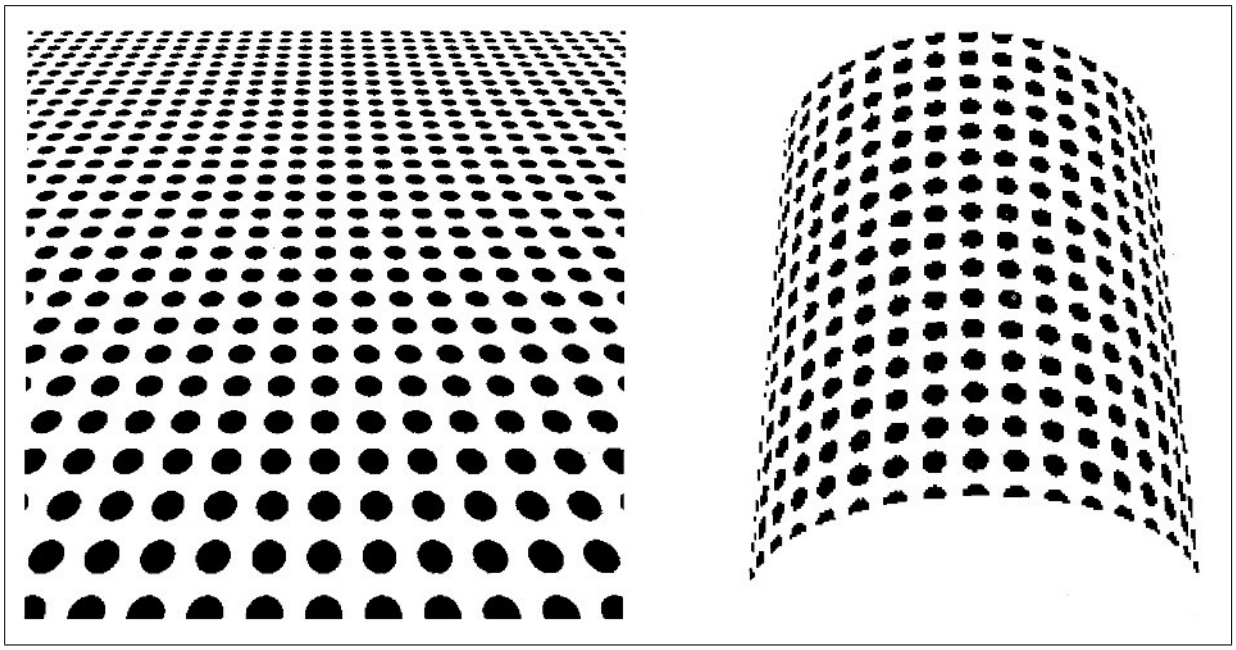

Figura 2.3: *Reconhecimento da forma do objeto usando apenas textura. (fonte [74])*

Bajcsy e Lieberman [7] desenvolveram um método global para medir o gradiente de textura de cenas formadas por uma superfície longitudinal com texturas que se repetem, como um gramado ou o oceano, por exemplo. Nesse método, a profundidade é inferida pelo tamanho relativo dos *texels* (elementos de textura) através do plano. O primeiro passo é encontrar e caracterizar os *texels*, processo feito através da transformada de *fou-*

rier, tornando possível uma descrição, como direção e tamanho de onda do espectro de força da textura. Essas informações são comparadas entre pequenos pedaços da imagem para que similiridades e diferenças sejam encontradas. Finalmente, um modelo de projeção é necessário para relacionar a mudança no tamanho da textura no plano da imagem com o mundo 3D. Esse é um método simplificado e serve basicamente para planos texturizados encontrados no mundo real assumindo uma projeção em perspectiva e texturas distribuídas uniformimente.

Um método de *shape from texture* exclusivamente para superfícies curvas foi proposto por Super e Bovik [74]. Nesse método, assumem-se texturas homogêneas e a projeção usada é a ortográfica, já que os efeitos da perspectiva são limitados em superfícies curvas. As informações usadas nesse método são extraídas localmente do espectro da textura através do filtro de *gabor*, ao invés das informações globais produzidas pela transformada de *fourier*. Essa informação é então usada diretamente em um modelo de projeção ortográfica de modo a recuperar a forma da superfície.

Aloimonos [5] desenvolveu um método para recuperar a orientação de planos texturizados usando projeção em paraperspectiva, assumindo, assim como em [7], densidade uniforme dos *texels*. Esse método permite que a orientação dos planos seja recuperada mesmo que os *texels* não sejam encontrados, desde que seus limites possam ser localizados.

2.4 Shape from Focus e Shape from Defocus

Em um modelo real de câmera apenas um plano de profundidade fica em foco em cada imagem tirada, como vemos na *Figura 2.4*, e, basedos nessa ideia simples, vários métodos têm surgido com a proposta de, a partir das medidas de foco, recuperar o *depth map* (mapa de profundidade) da cena. Esses métodos se baseiam na mudança da configuração de foco através do tamanho de abertura da lente ou da distância focal, sem a necessidade de se mover o dispositivo de captura ou o objeto de interesse, para cada imagem capturada. Essa diferença na aquisição das imagens já é uma vantagem em relação a vários outros

métodos que acabam adicionando a necessidade de se procurar correspondências entre as imagens.

O *shape from focus* é uma abordagem que propõe recuperar o mapa de profundidade da cena a partir da mudança ativa dos parâmetros óticos da câmera até que o ponto de interesse esteja em foco e, a partir disso, saber a profundidade de um ponto focado. O *shape from defocus* é o problema de reconstruir o mapa de profundidade da cena dado um conjunto de imagens tiradas da mesma câmera mudando os parâmetros óticos e é dividido em abordagem ativa (baseada em criar textura sobre todas as superfícies usando luz estruturada) e passiva (na qual não há intervenção na cena). [21]

Figura 2.4: *Distância focal diferenciada para a mesma cena*

Um método clássico de *shape from focus* é o de Nayar [53] para superfícies ásperas, no qual se introduz o SML (soma dos laplacianos modificados, em inglês) como o operador para calcular as medidas locais da qualidade de foco da imagem em detrimento ao laplaciano comum (como em Darrell e Wohn [18]), devido aos casos em que as segundas derivadas (que servem para detectar mudanças bruscas em imagens) em x e y têm sinais opostos e tendem a cancelar uma a outra. O SML resolve esse problema considerando o módulo das segundas derivadas em x e y para a medida de foco. Em seu artigo, Nayar capturou a pilha de imagens do objeto movendo o foco no objeto a cada 1μm e usando uma janela para medida de foco de 10x10 pixels. A medida de pronfundidade é obtida diretamente a partir dos pontos que maximizam a medida de foco em cada imagem, levando

em consideração o deslocamento de foco que foi feito no objeto. Entre as desvantagens dessa abordagem estão a presunção de que a superfície é altamente texturizada, para que se faça possível a medição do foco, e a dificuldade para a obtenção das imagens de foco, com o deslocamento exato desejado de foco, sem a ajuda de um sistema automático (proposto posteriormente por Nayar e Nakagawa em [54]). Tanto a resolução quanto a acurácia são limitadas com este método.

Pentland [60] foi um dos primeiros a usar a medida de desfoque para estimar profundidade usando um número reduzido de imagens, reduzindo assim também o trabalho na captura das imagens e o custo computacional para a resolução do problema. A função que descreve o quão borrado um ponto aparece, dada a distância que este se encontra da lente (*Figura 2.5*), chama-se função de espalhamento do ponto. Neste método, a função de espalhamento é aproximada por um gaussiano bidimensional, considerando-se uma constante espacial de borramento do ponto e a distância radial da lente. Essa constante usada no gaussiano deve ser medida, e duas formas são propostas por Pentland: observando-se locais da imagem com características conhecidas (como cantos); ou alterando-se algum aspecto do sistema de lentes em duas ou mais imagens, o que resultará em diferenças no foco entre elas, que podem ser comparadas. Entre as desvantagens do primeiro modo de se descobrir a constante, podemos citar: as características da cena devem ser conhecidas previamente; e pode ser usado basicamente para se fazer uma segmentação de planos da cena por profundidade, por não ser preciso o suficiente para produzir um mapa de profundidade. O segundo modo de se calcular a constante da função de espalhamento do ponto consiste em se capturar pelo menos duas imagens da cena, que serão idênticas, a não ser pelo gradiente de foco, através de diferentes tamanhos de abertura de lente. O mapa de disparidade gerado a partir da constante de espalhamento do ponto encontrado por essa segunda forma é bem mais preciso, porém, a grande dificuldade nesta abordagem é garantir imagens com texturas suficientes para a medição da mudança entre as imagens.

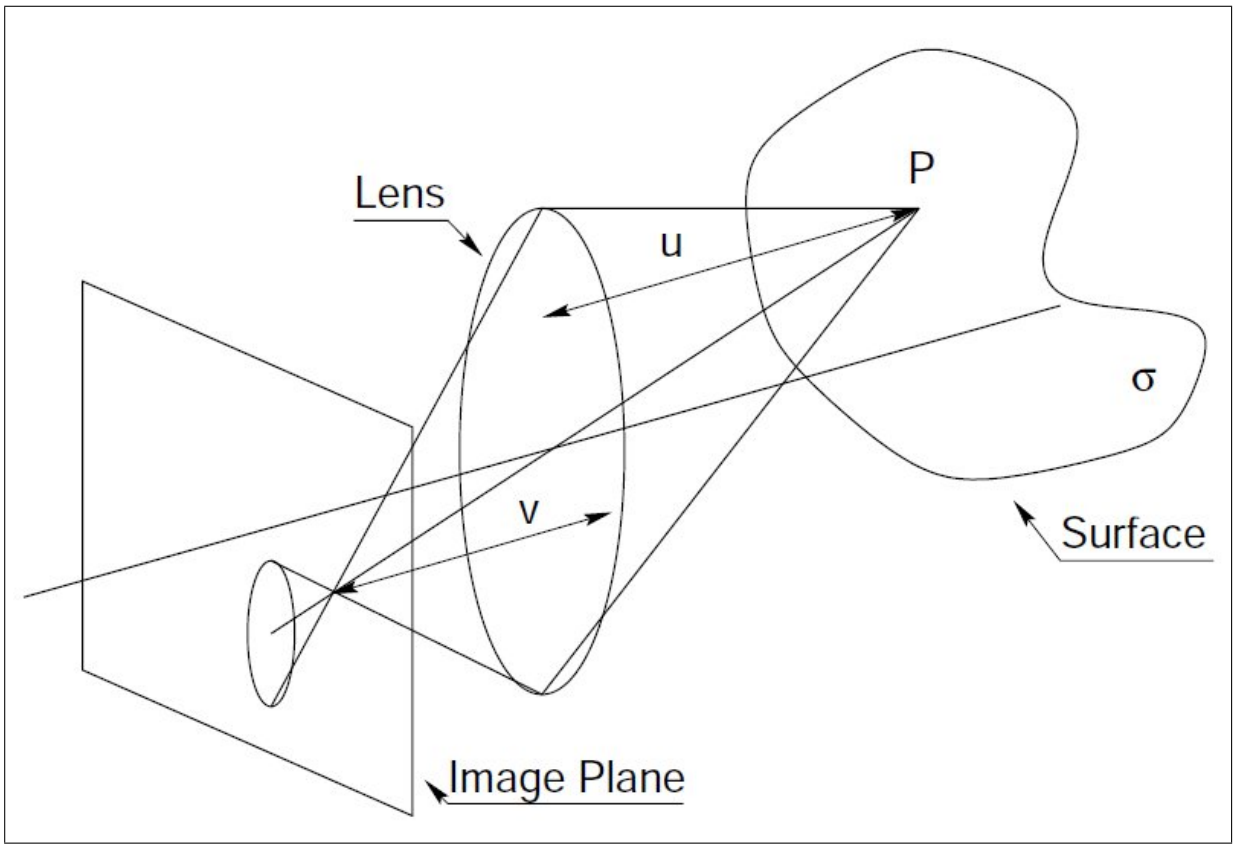

Figura 2.5: *Espalhamento do ponto (borramento) em um modelo real de câmera. (fonte [21])*

2.5 Métodos híbridos

Métodos híbridos combinam mais de um método na tentativa de obter uma reconstrução mais apurada. A ideia básica é, a partir de uma estimativa mais grosseira da geometria 3D, refinar o modelo com o auxílio de algum outro método a fim de obter um melhor resultado na reconstrução.

Hernandez e Schmitt [30] combinam *shape from silhouette* e *stereo* para a reconstrução da geometria e textura do objeto através de uma sequência geometricamente calibrada de imagens coloridas do objeto em uma mesa giratória (o que exige um ambiente controlado para a aquisição), obtendo resultados muito bons a partir da fusão dos dados obtidos através dos dois métodos em uma deformação evolutiva do *visual hull* (Figura 2.6).

Figura 2.6: *Esquerda: inicialização do visual hull; direita: modelo final. (fonte [30])*

Os métodos descritos em [16] e [32] usam *photometric stereo* como uma forma de refinar o *visual hull* obtido através de *shape from silhouette*. O primeiro rotaciona o objeto em um ângulo pré-estabelecido para a aquisição de cada imagem, o que exclui a necessidade de calibração do sistema, porém dificulta o processo de captura pela necessidade de um controle na rotação do objeto. O segundo também não necessita de calibração, já que tanto a posição da câmera quanto as direções e intensidades das fontes de iluminação são estimadas a partir das silhuetas do objeto, possibilitando uma reconstrução completa do modelo 3D de alta qualidade. Uma comparação entre um modelo reconstruído usando este último método e aquele usando *stereo* com silhueta em [30] pode ser visto na *Figura 2.7*.

Já em 2003, Zhang et al. [85] desenvolve um método que é capaz de estimar a movimentação da câmera, iluminação da cena, refletividade difusa da superfície e forma do objeto, tanto para superfícies muito texturizadas como para aquelas poucos texturizadas baseado na iluminação e movimentação da cena. A projeção ortográfica é considerada para a resolução do problema, que utiliza as informações de rotação da câmera e posição 3D dos *feature points* recuperadas através do método de *shape from motion* de fatoração

de Tomasi-Kanade [76] para inicializar o sistema. As normais da superfície são calculadas nesses pontos já reconstruídos e a iluminação da cena é também inicializada. A partir dessas primeiras estimativas de superfície e iluminação, iterativamente, para cada quadro da sequência, são calculadas a posição e as normais em cada pixel da imagem para integração com o modelo, bem como a variação da iluminação. É assumido um objeto rígido de superfície lambertiana iluminado por uma fonte de luz fixa distante. Sombras, oclusões e interreflexões não são consideradas.

Figura 2.7: *Esquerda: imagem original; meio: modelo usando [30]; direita: modelo usando [32]. (fonte [32])*

A ideia proposta por Lim et al. [47] é bem simples e produz também bons resultados. Basicamente, o objeto é movido em frente à câmera em uma cena (possivelmente com iluminação ambiente) enquanto é iluminado por uma fonte direcional de luz fixa, também distante. A movimentação do objeto é recuperada a partir do método de fatoração de Tomasi-Kanade, bem como uma superfície inicial é gerada a partir dos poucos pontos reconstruídos por este método. As normais do objeto são obtidas, através do método de *photometric stereo* da fatoração da matriz de intensidades usando *SVD*, e é feita a integração com a superfície já produzida, resultando em um novo mapa de profundidades. A cada iteração são feitas correções para que os valores de profundidade se aproximem cada vez mais daqueles obtidos pelos pontos rastreados (*Figura 2.8*). Joshi e Kriegman [40]

usam uma abordagem semelhante, também usando a fatoração de Tomasi-Kanade para a recuperação das matrizes de projeção da câmera, porém geram um mapa de profundidade denso (e não apenas alguns *feature points*) para a integração com as normais obtidas através do *photometric stereo*. Esse método possui bons resultados e algumas vantagens sobre o método de Zhang et al. [85], como necessitar de poucas imagens no processo, e de Lim et al. [47], que nem sempre converge para um bom resultado. Além disso, não é um método iterativo, já que computa cada etapa em apenas um passo e une os resultados em uma superfície de alta qualidade.

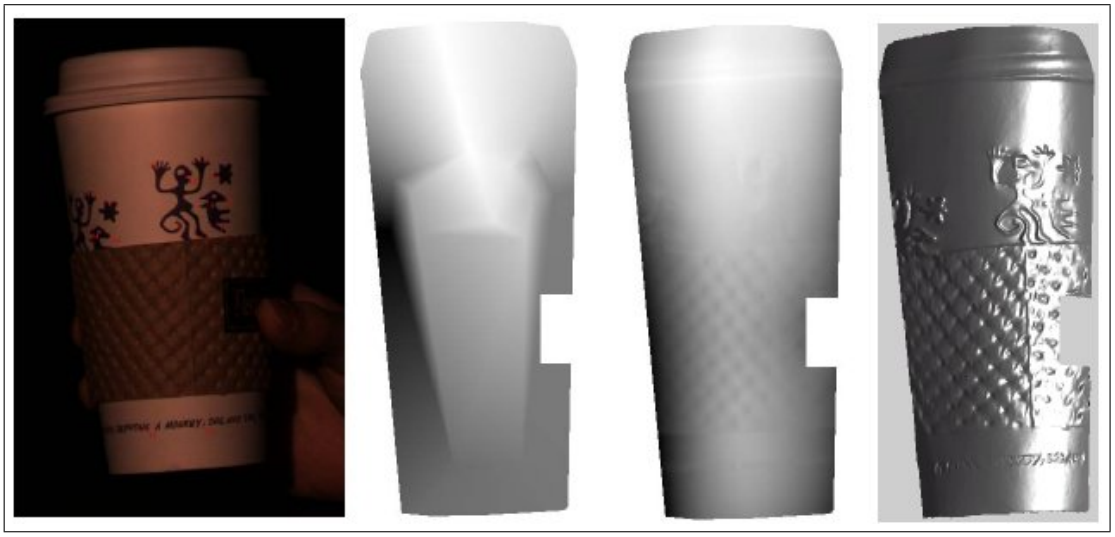

Figura 2.8: *Esquerda: imagem original tirada da sequência usada na reconstrução; meio: superfície inicial reconstruída a partir do shape from motion e depois de quatro iterações; direita: superfície reconstruída usando modelo de refletividade. (fonte [47])*

Higo et al. [34] formula o problema de forma mais realística, admitindo superfícies não lambertianas, fontes de luz próximas além de luz ambiente, e assumindo o modelo de projeção em perspectiva. O sistema assume uma configuração básica de um ponto de luz anexo à camera, mantido sempre ligado durante a captura, de forma à inserir uma restrição para auxiliar na sua resolução. A calibração dos parâmetros internos da câmera é feita ainda antes do início da aquisição dos dados e a calibração dos parâmetros externos é feita com o bundler [72].

2.6 Discussão sobre os métodos e linha seguida

Dentre os métodos estudados, alguns não são de grande valia para a reconstrução de objetos de acervos culturais devido a sua baixa resolução ou acurária, ou devido a grandes limitações dos métodos para certos tipos de objetos ou ambientes controlados.

O *shape from motion* é um dos métodos mais utilizados atualmente, seja devido ao seu baixo custo e portabilidade ou por ser um método que permite a recuperação das informações de textura do objeto ou da posição das câmeras utilizadas na captura, porém tem a capacidade de recuperar poucos pontos, devido sua dependência com relação a correspondência de pontos e regiões texturizadas, o que faz seu uso sem o auxílio de outros métodos inadequado à reconstrução precisa de objetos.

Encontramos em muitas técnicas de *shape from shading* e *photometric stereo* também a vantagem da portabilidade e do baixo custo, além de superar a limitação de reconstruir apenas objetos com superfícies altamente texturizadas, como é o caso da maioria dos métodos passivos. Como desvantagens podemos citar, por exemplo, no caso específico do *shape from shading*, a ambiguidade encontrada ao se recuperar a orientação de uma superfície ou, de forma mais ampla para os métodos que envolvem fotometria, considerações que devem ser feitas com relação a superfície que se deseja reconstruir, como assumir refletividade lambertiana ou a ausência de interreflexão na superfície do objeto. Também, quando a calibração é requerida, um ambiente controlado ou objetos de calibração passam a ser necessários, acabando com a portabilidade do método ou limitando a reconstrução para objetos com superfícies mais comuns, para as quais objetos de calibração possam ser facilmente adquiridos.

O *shape from texture* é sabidamente um método com maior aplicabilidade na recuperação da estimativa de deformação de superfícies não rígidas, como tecido e pele [23], não sendo muito utilizado na reconstrução de objetos rígidos. É também limitado pela necessidade de texturas que seguem padrões conhecidos ou determinados e recupera apenas orientações de planos ou formas simples. Por esses motivos, não é considerado um

método atrativo para a reconstrução realística de modelos para a preservação digital, não sendo explorado em maiores detalhes.

Os métodos que se baseiam na medição de foco são simples e fáceis de se aplicar, porém demandam o uso de lentes muito especiais para um controle adequado do foco e o principal uso acaba ficando limitado a imagens capturadas através de microscópios. O *shape from focus* possui uma grande dificuldade no deslocamento exato do foco quando este é feito manualmente e a resolução do modelo reconstruído é limitada pela distância que o foco varia a cada imagem capturada. Há também grande dependência, tanto em *shape from focus* quanto em *shape from defocus*, de uma superfície altamente texturizada para que as medidas do foco possam ser tiradas com certa precisão. Devido ao grande número de restrições e limitações, este método também demonstrou-se como não sendo adequado ao propósito deste trabalho.

Os métodos híbridos parecem ser a fonte de melhores resultados dentre os métodos estudados por vários motivos. Primeiramente, todos os métodos híbridos citados produzem resultados utilizáveis, ou seja, superfícies, e não apenas pontos esparsos, com boa precisão. Além disso, as informações recuperadas são uma mistura daquelas recuperadas pelos métodos que se combinam, por exemplo, movimentação da câmera para os métodos que utilizam *shape from motion* e iluminação da cena e refletividade difusa da superfície do objeto quando se usa *photometric stereo*. Devido a algumas restrições dos métodos que usam *shape from silhouette* (como o uso de uma mesa giratória ou da necessidadede controle no ângulo de rotação do objeto para cada imagem obtida, ou da necessidade de uma segmentação do objeto de interesse do fundo), os métodos híbridos que dependem de reconstrução por silhueta não serão estudados mais a fundo. Métodos que combinam *shape from motion* e *photometric stereo* parecem o principal foco de bons resultados de métodos passivos, com menos restrições e aplicabilidade na reconstrução de objetos de acervos culturais.

Neste trabalho são testados os métodos sugeridos de *shape from motion* que são a base dos métodos híbridos que parecem mais interessantes para nossa finalidade: o método

de fatoração de Tomasi-Kanade [76], um método clássico e já bastante estudado que caracteriza uma classe de métodos de reconstrução, e o bundler [72], método estado da arte de reconstrução baseado em fotos de várias vistas de uma mesma cena. Ambos os métodos propõem a recuperação de uma nuvem de pontos esparsos, que podem ser usados, em uma próxima etapa, na geração de uma superfície completa a partir dos dados extras obtidos sobre a orientação da superfície do objeto com um método de *photometric stereo*. Também um estudo sobre o *photometric stereo* utilizando a fatoração de matrizes será feito, de modo a salientar as restrições que devem ser assumidas para a reconstrução dos vetores normais da superfície e o albedo do objeto, bem como a iluminação da cena de forma única. Finalmente será feito um paralelo sobre como os métodos de *shape from motion* podem melhorar a qualidade das superfícies reconstruídas e auxiliar na resolução das ambiguidades obtidas no *photometric stereo*. Um método híbrido proposto por Lim et al. [47], baseado no método de Tomasi-Kanade e no *photometric stereo* através de fatoração da matriz de intensidade dos pontos, será também explicado e seus resultados expostos e comentados, bem como uma análise sobre sua utilidade final em preservação digital será feita.

CAPÍTULO 3

MÉTODOS DE STRUCT FROM MOTION

Dois métodos são explicados neste capítulo: o método da fatoração de Tomasi-Kanade [76], originalmente formulado para projeção ortográfica (um tipo de projeção paralela, mais simples que a projeção em perspectiva, no qual todas as linhas da projeção são ortogonais ao plano de projeção) e que dispensa o uso de calibração; e o bundler [72], um método estado da arte de *shape from motion* que formula o problema assumindo projeção em perspectiva e é voltado para coleções desordenadas de imagens retiradas da internet.

Os resultados dos testes feitos com esses métodos são devidamente expostos e comentados de acordo com a sua utilidade para a preservação digital de acervos culturais, como é a proposta deste trabalho.

3.1 Método da fatoração de Tomasi-Kanade

Em 1979, Ullman [80] prova a existência de uma solução para o problema de *struct from motion* em projeções ortográficas e Roach e Aggarwal [65] provam a existência de uma solução para o mesmo problema em projeções em perspectiva. Em ambos os trabalhos, as coordenadas são expressas em um sistema de referência centrado no objeto. Desde então, comumente passou-se a adotar o sistema de referência centrado na câmera, onde a posição dos pontos é representada pelas suas coordenadas na imagem e por sua profundidade, definida pela distância do centro da câmera até o ponto em questão, o que simplifica as equações para projeções em perspectiva. Apesar de simplificar as equações, esse sistema torna a estimativa da estrutura sensível a ruído e instável, já que, quando a movimentação da câmera é pequena, uma rotação e uma translação resultam em mudanças muito semelhantes na imagem e também os problemas são aumentados quando o

objeto é pequeno se comparado a distância da câmera. No método de Tomasi e Kanade [76] esse problema deixa de existir, já que volta-se a utilizar o sistema de coordenadas centrado no objeto, e não na câmera, e as informações de estrutura da cena, bem como de movimentação da câmera, são obtidas através da fatoração da matriz de medidas, que nada mais é do que uma matriz que contém as coordenadas 2D dos pontos rastreados em imagens através de uma sequência de vídeo.

A seguir, serão explicadas as etapas necessárias para a implementação do método originalmente proposto, baseado em projeção ortogonal, e sua extensão proposta para projeções em paraperspectiva.

3.1.1 Matriz de medidas

A primeira coisa a ser feita é o rastreamento dos pontos do objeto em cada quadro da sequência de vídeo e o armazenamento de suas coordenadas em duas matrizes, U e V. Foi utilizada uma implementação do algoritmo *KLT* para o rastreamento dos pontos [68] (o *SIFT* também poderia ter sido usado neste caso, porém, devido à natureza do problema, o uso do *KLT* é suficiente) e as coordenadas horizontais dos pontos rastreados foram salvas na matriz U, de tamanho F (número de quadros, ou *frames*, do vídeo) por P (número de pontos que puderam ser rastreados em todos os quadros), e na matriz V, de mesmo tamanho da matriz U, foram mantidas as coordenadas verticais desses pontos. O tamanho dessas matrizes é justificado, pois, utiliza-se uma linha por *frame* e uma coluna por ponto rastreado.

Essas duas matrizes podem ser representadas em apenas uma:

$$W = \left[\frac{U}{V}\right]. \tag{3.1}$$

W é a matriz de medidas. A metade superior da matriz representa as coordenadas horizontais e a metade inferior as coordenadas verticais dos pontos.

De forma a simplificar as equações, a matriz W é registrada movendo-se a origem do sistema para o centróide dos pontos rastreados.

$$\{(u_{fp}; v_{fp}) \mid f = 1, ..., F, p = 1, ..., P\}.$$

$$\tilde{u}_{fp} = u_{fp} - af \tag{3.2}$$

$$\tilde{v}_{fp} = v_{fp} - bf, \tag{3.3}$$

onde

$$af = \frac{1}{P}\sum_{p=1}^{P} u_{fp} \tag{3.4}$$

$$bf = \frac{1}{P}\sum_{p=1}^{P} v_{fp}. \tag{3.5}$$

A matriz obtida é

$$\tilde{W} = \left[\frac{\tilde{U}}{\tilde{V}}\right], \tag{3.6}$$

que passa a ser utilizada, ao invés da matriz de medidas original W.

3.1.2 Rank da matriz de medidas

A álgebra linear define que o *rank* de uma matriz é menor ou igual à menor de suas dimensões. Nossa matriz de medidas $\tilde{W}$ $(2F \times P)$ é altamente deficiente com relação ao *rank*.

A orientação da câmera em um *frame* f é representada por um par de vetores unitários i_f *e* j_f. Na projeção ortográfica os raios da projeção são todos paralelos ao produto externo de i_f *e* j_f:

$$k_f = i_f \times j_f. \tag{3.7}$$

Assim, as coordenadas 2D de cada ponto $\{s_p = (x_p, y_p, z_p)^T \mid p = 1, \ldots, P\}$ no *frame* f são representadas pelo produto interno

$$\tilde{u}_{fp} = i_f^T (s_p - t_f) \tag{3.8}$$

$$\tilde{v}_{fp} = j_f^T (s_p - t_f), \tag{3.9}$$

onde $t_f = (a_f, b_f, c_f)^T$ é o vetor de transformação da origem do universo para a origem da imagem no *frame* f. Porém, como em $\tilde{W}$ o centro do sistema de coordenadas de cada quadro do vídeo já foi considerado como o centróide dos pontos naquele quadro, é válido desconsiderarmos o vetor de translação t_f, podendo assim representar $\tilde{u}_{fp}$ e $\tilde{v}_{fp}$ da seguinte forma:

$$\tilde{u}f_p = i_f^T s_p \tag{3.10}$$

$$\tilde{v}f_p = j_f^T s_p. \tag{3.11}$$

Sendo assim, a matriz $\tilde{W}$ pode ser expressa por:

$$\tilde{W} = RS \tag{3.12}$$

onde

$$R = \begin{bmatrix} i_1^T \\ \vdots \\ i_f^T \\ j_1^T \\ \vdots \\ j_f^T \end{bmatrix} \tag{3.13}$$

representa a orientação da câmera através da seqüência e

$$S = \begin{bmatrix} s_1 & ... & s_p \end{bmatrix} \tag{3.14}$$

é a matriz contendo as coordenadas dos P pontos com respeito ao seu centróide.

Os vetores base i_f *e* j_f, em projeção ortográfica, estão sujeitos as seguintes restrições:

$$i_f^T i_f = j_f^T j_f = 1 \; i_f^T j_f = 0,$$

já que são ortonormais, ou seja, ortogonais e unitários.

O modelo de projeção ortográfica, onde a direção da visão é ortogonal ao plano de projeção, é como um modelo de projeção em perspectiva onde a câmera está a uma distância infinita do objeto e possui um tamanho focal também infinito, ou seja, as linhas seguem em direção ao plano de projeção paralelamente, e não se encontram nunca. Como este método foi feito assumindo-se a projeção ortográfica, para um bom resultado a distância da câmera até o objeto deve ser grande se comparada a sua profundidade, ou o resultado será pobre devido a falta de projeção real em perspectiva.

Como a matriz $\tilde{W}$ pode ser expressa em duas matrizes, R, de tamanho $2F \times 3$, e S, $3 \times P$, o teorema do *rank* implica que a matriz registrada de medidas $\tilde{W}$, sem ruído, é de no máximo *rank* três.

3.1.3 Resolvendo o rank para os casos de ruído

Devido ao ruído obtido durante o rastreamento dos pontos, a matriz de medidas $\tilde{W}$ não vai ter *rank* exatamente três. O teorema do *rank* é estendido para medidas ruidosas através do *Singular Value Decomposition (SVD)* [24]. O *SVD* exige matrizes completas, ou seja, pontos que sofrem oclusão devem ser excluídos da sequência ou tratados de alguma forma. Existem outros métodos que podem ser usados para a fatoração de matrizes incompletas [25]. Além disso, sendo a transformada do *SVD* um método de mínimos quadrados, ela é extremamente sensível a *outliers* (erros de rastreamento dos pontos, por exemplo).

$\tilde{W}$ pode ser decomposta, através do *SVD*, em uma matriz O_1 de tamanho $2F \times P$, uma matriz diagonal D de tamanho $P \times P$ e outra matriz de tamanho $P \times P$ chamada de O_2,

$$\tilde{W} = O_1 D O_2, \tag{3.15}$$

de modo que $O1^T O_1 = O_2^T O_2 = O_2 O_2^T = I$, onde I é a matriz identidade $P \times P$. Os valores singulares (*singular values*) são dispostos na matriz diagonal D, ordenados de forma não-crescente.

O produto que resulta em $\tilde{W}$ pode ser divido em duas partes:

$$O_1 D O_2 = O'_1 D' O'_2 + O''_1 D'' O''_2 \tag{3.16}$$

onde

$$O_1 = \left[\begin{array}{c|c} O'_1 & O''_1 \end{array} \right] \tag{3.17}$$

e O'_1 tem tamanho $2F \times 3$ e O''_1 tem tamanho $2F \times (P-3)$,

$$D = \left[\begin{array}{c|c} D' & 0 \\ \hline 0 & D'' \end{array} \right] \tag{3.18}$$

e D' tem tamanho 3×3 e D'' tem tamanho $(P-3) \times (P-3)$ e

$$O_2 = \left[\begin{array}{c} O_2' \\ \hline O_2'' \end{array}\right], \tag{3.19}$$

em que O_2' tem tamanho $3 \times P$ e O_2'' tem tamanho $(P-3) \times P$.

Chamaremos de $\hat{W}$ a matriz de medidas com ausência de ruído. Pelo teorema do *rank*, sabemos que $\hat{W}$ tem no máximo três valores singulares não nulos. Desta forma, D' deve conter todos o valores singulares de $\hat{W}$ que superam o nível de ruído quando fazemos

$$\hat{W} = O_1' D' O_2', \tag{3.20}$$

ou seja, toda a informação sobre a rotação e estrutura em $\tilde{W}$ está contida nos seus três maiores valores singulares, juntamente com seus autovetores esquerdo e direito.

Dessa forma, assumimos

$$\hat{R} = O_1' [D']^{1/2} \tag{3.21}$$

$$\hat{S} = [D']^{1/2} O_2' \tag{3.22}$$

para

$$\hat{W} = \hat{R}\hat{S}. \tag{3.23}$$

Fica óbvio que a decomposição de $\hat{W}$ não é única, pois, escolhendo-se qualquer matriz inversível Q, de tamanho 3×3, as matrizes $\hat{R}Q$ e $Q^{-1}\hat{S}$ são também uma decomposição válida de $\hat{W}$.

Felizmente é verdade que, exceto pela presença de ruído, a matriz $\hat{R}$ é uma transformação linear de R, sendo o mesmo válido para $\hat{S}$ e S. Basta, então, descobrir-se a matriz Q para que obtenhamos as matrizes R e S, nas quais encontram-se a rotação de

câmera e a estrutura da cena, respectivamente:

$$R = \hat{R}Q \tag{3.24}$$

$$S = Q^{-1}\hat{S}. \tag{3.25}$$

Considerando a projeção como ortogonal, podemos deduzir os elementos de Q a partir das seguintes restrições da matriz $\hat{R}$:

$$\hat{i}_f^T Q Q^T \hat{i}_f = 1,\ \hat{j}_f^T Q Q^T \hat{j}_f = 1,\ \hat{i}_f^T Q Q^T \hat{j}_f = 0.$$

Para resolver esse sistema não linear, temos duas opções: modelar o problema e utilizar um método que resolva sistemas não lineares, como o Levenberg-Marquardt [45, 50]; ou definir $L = QQ^T$, um sistema linear, e resolver um sistema de equações para L usando o método da pseudoinversa, como explicado por Morita e Kanade em [52].

3.1.4 Atualização métrica

Usamos o segundo método para a resolução do sistema envolvendo a matriz Q. Denotamos

$$L = \begin{bmatrix} l_1 & l_2 & l_3 \\ l_2 & l_4 & l_5 \\ l_3 & l_5 & l_6 \end{bmatrix}, \tag{3.26}$$

porém, o sistema pode ser reescrito

$$GI = c, \tag{3.27}$$

onde G tem o tamanho $3F \times 6$, I é um vetor 6×1 e c é um vetor de tamanho $3F \times 1$, definidos por

$$G = \begin{bmatrix} g^T(i_1, i_1) \\ \vdots \\ g^T(i_f, i_f) \\ g^T(j_1, j_1) \\ \vdots \\ g^T(j_f, j_f) \\ g^T(i_1, j_1) \\ \vdots \\ g^T(i_f, j_f) \end{bmatrix}, l = \begin{bmatrix} l_1 \\ \vdots \\ l_6 \end{bmatrix}, c = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3.28}$$

onde c tem $2F$ uns e f zeros e

$$g(a, b) = \begin{bmatrix} a1b1 \\ a1b2 + a2b1 \\ a1b3 + a3b1 \\ a2b2 \\ a2b3 + a3b2 \\ a3b3 \end{bmatrix}. \tag{3.29}$$

Para resolvermos o sistema e descobrirmos os valores no vetor l, utilizamos o método da pseudoinversa:

$$l = G^* c, \tag{3.30}$$

sendo

$$G^* = (G^T G)^{-1} G^T.$$

O vetor l determina a matriz simétrica L, que deve ser uma matriz definida positiva

para podermos determinar Q. Como sabemos, L já é uma matriz simétrica (requisito básico para a definição positiva), portanto podemos decompor L em autovetores (*eigen-decomposition*)

$$L = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathbf{T}},$$

e, para forçamos a definição positiva, colocamos todos os autovalores negativos na diagonal de $\mathbf{\Sigma}$ como 0, obtendo assim $\mathbf{\Sigma}_+$. A partir daí, podemos obter Q fazendo

$$Q = \mathbf{U}{\mathbf{\Sigma}_+}^{1/2}.$$

Tendo descoberto a matriz Q, torna-se então possível usar as equações (3.24) e (3.25) para se obter a rotação da câmera e a estrutura da cena.

3.1.5 Projeção em paraperspectiva

A extensão do método da fatoração para projeção em paraperspectiva [56] foi feita por Poelman e Kanade [62], passando a considerar alguns fatores do mundo real os quais a projeção ortográfica ignora.

A projeção em paraperspectiva se aproxima mais da projeção em perspectiva por considerar tanto o fator de escala (objetos mais distantes aparecem menores que os mais próximos) e o efeito de posição (objetos na periferia da imagem são visualizados de um ângulo diferente daqueles localizados mais próximos ao centro).

A diferença entre projeção em perspectiva e ortográfica pode ser verificada nas *Figuras 3.1* e *3.2*.

Na projeção em paraperspectiva, cada ponto do objeto é projetado em um plano de referência, paralelo ao plano da imagem e posto virtualmente entre este e o objeto, em uma direção paralela à linha que conecta um dado ponto de referência ao centro de foco da câmera. Deste plano de referência para o plano de imagem, os pontos são projetados

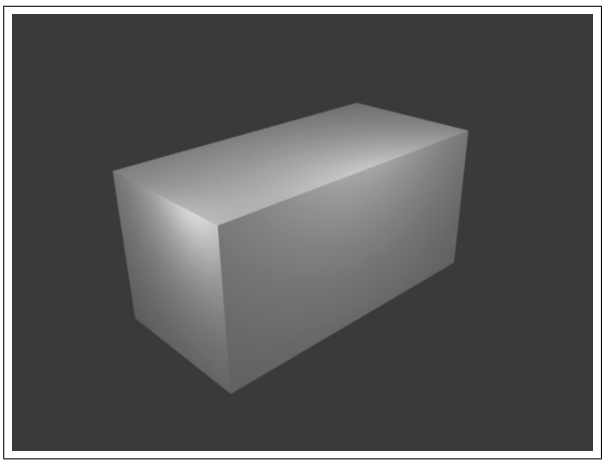

Figura 3.1: *Paralelepípedo projetado em perspectiva.*

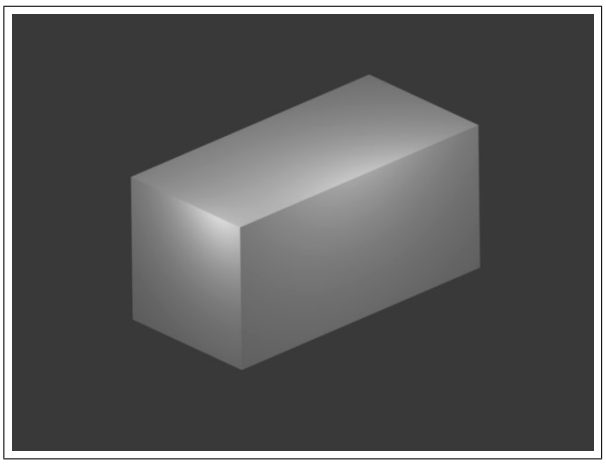

Figura 3.2: *Paralelepípedo em projeção ortográfica.*

através de uma projeção central. Por esse motivo, todo o processo é igual ao da projeção ortográfica, alterando-se apenas as restrições métricas.

M, neste caso, é a matriz estimada de movimento

$$M = \begin{bmatrix} m_1^T \\ \vdots \\ m_F^T \\ n_1^T \\ \vdots \\ n_F^T \end{bmatrix}, \tag{3.31}$$

e T é o vetor de translação

$$T = \begin{bmatrix} x_1 \\ \vdots \\ x_F \\ y_1 \\ \vdots \\ y_F \end{bmatrix}. \tag{3.32}$$

Sendo assim, as restrições para o caso de projeção em paraperspectiva são:

$$\frac{|m_f|^2}{1+x_f^2} - \frac{|n_f|^2}{1+y_f^2} = 0, \tag{3.33}$$

$$m_f \cdot n_f - x_f y_f \frac{1}{2}\left(\frac{|m_f|^2}{1+x_f^2} - \frac{|n_f|^2}{1+y_f^2}\right) = 0, \tag{3.34}$$

$$m_f = 1. \tag{3.35}$$

Da mesma forma que no caso de projeção ortográfica

$$|m_f|^2 = \hat{m}_f^T Q Q^T \hat{m}_f,\ |n_f|^2 = \hat{n}_f^T Q Q^T \hat{n}_f,\ m_f \cdot n_f = \hat{m}_f^T Q Q^T \hat{n}_f,$$

onde Q é uma matrix 3×3 e $\hat{m}_f$ e $\hat{n}_f$ são os elementos da matriz estimada $\hat{M}$. A matriz Q pode ser resolvida da mesma forma que no caso ortográfico.

3.2 Bundler para a reconstrução a partir de coleções desordenadas de imagens

Coleções de imagens são encontradas em abundância na internet, como, por exemplo, de fotos de monumentos turísticos famosos. Aproveitando-se da facilidade de se obter tais tipos de imagens, foi desenvolvido um método de *shape from motion* capaz de reconhecer e relacionar trechos de imagens, independentemente de suas diferenças de tamanho, resolução, iluminação, ou da forma como estão ordenadas, e, através de refinamento da estrutura 3D da cena a cada nova relação encontrada entre imagens, reconstruir a cena 3D com um número razoável de pontos e boa precisão. Este método deu origem ao bundler [72]. De acordo com a proposta deste trabalho, as etapas de correspondência e de recuperação das coordenadas 3D dos pontos são explicadas em maiores detalhes.

3.2.1 Correspondência de características

Diferentemente do método da fatoração, no qual optamos por fazer a correspondência dos pontos entre os pares de imagens através do *KLT* [68], que mostra bons resultados quando o rastreamento é feito quadro a quadro (cenas vistas de pontos de vista bem próximos), a correspondência entre os pontos na implementação que pode ser encontrada no site do projeto bundler [1], que foi a versão testada neste trabalho, é feita através do método *SIFT (Scale Invariant Feature Transform)* de Lowe [49]. Este método é melhor indicado para este caso, no qual as imagens foram tiradas de vistas bem diferentes entre si, pois tem boa invariância a transformações nas imagens.

A extração das características é o passo inicial do *SIFT*. Primeiro encontram-se os *feature points* (ou *keypoints*, que são pontos nas imagens com características distinguíveis), e cria-se, para cada ponto, um vetor de características como forma de descrevê-los. Cada vetor de características é completamente invariante quanto à translação, escala e rotação, e parcialmente invariante à mudanças de iluminação. O processo de *matching*, ou seja, de casamento dos *keypoints* de uma imagem com os de outra, no bundler, é feito com o auxílio de uma *kd-tree*. Para cada par de imagens, uma *kd-tree* é criada com todos os descritores de características encontrados na segunda imagem (imagem J), e, para cada *keypoint* na primeira imagem (I), busca-se o vizinho mais próximo (o vetor de características com a miníma distância euclidiana) na *kd-tree*, limitando a busca a 200 nós da árvore, de forma a manter a busca eficiente. No bundler foi usada a implementação de *kd-trees* encontrada na biblioteca *ANN* de Arya et al. [6]. A forma sugerida por Lowe para a detecção de falsos *matches* é a de encontrar os dois vizinhos mais próximos do *keypoint* de I em J (de distâncias d_1 e d_2) e, somente aceitar o de distânia d_1 como um *match* se $\frac{d_1}{d_2} < 0,6$, descartando-o caso contrário. Isso significa que se os dois pontos encontrados tiverem distâncias muito próximas, não podemos afirmar com segurança que um deles é realmente melhor que o outro. Outro caso em que o *match* é descartado é quando mais de um *feature* em I faz *match* com o mesmo ponto em J.

[1] http://phototour.cs.washington.edu/bundler/

Para cada par de imagens I e J a matriz fundamental de projeção é calculada a partir das correspondências de pontos encontradas na etapa anterior. A matriz fundamental F é uma matriz 3×3, de rank 2, estimada a partir de correspondências encontradas em um par de imagens, que limita as posições possíveis da projeção dos pontos nas duas imagens. Para cada ponto com coordenadas homogêneas x em uma imagem, Fx descreve uma linha epipolar (*Figura 3.3*) na qual o ponto de coordenada x' deve estar na outra imagem:

$$x'^T F x = 0, \tag{3.36}$$

sendo

$$x = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \; x' = \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \text{ e } F = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix}.$$

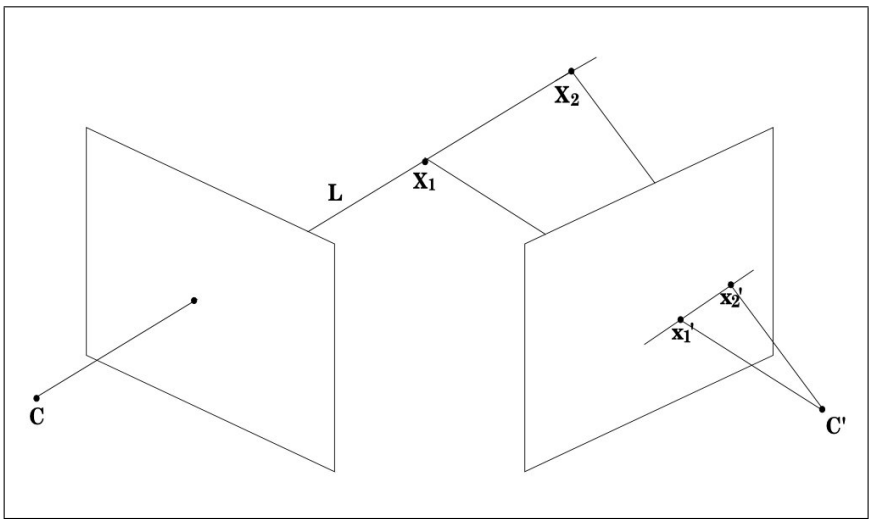

Figura 3.3: *Os pontos X_1 e X_2 são coincidentes quando vistos pela camera com centro em C e não coincidentes quando vistos pela camera com centro em C'. A linha que passa por x'_1 e x_2' é a imagem de L e é o que chamamos de linha epipolar.*

A matriz fundamental é computada através do algoritmo dos 8 pontos normalizado [22]. Este algoritmo consiste em encontrar a matriz fundamental F, a partir de oito pontos rastreados nas duas imagens, que minimize

$$f^T Y,$$

sendo cada coluna de Y composta pelos k vetores linearmente independentes y_k, f o vetor que contém os elementos de F, $k = 8$ e

$$y_k = \begin{bmatrix} u'_k u_k \\ u'_k v_k \\ u'_k \\ v'_k u_k \\ v'_k v_k \\ v'_k \\ u_k \\ v_k \\ 1 \end{bmatrix}. \tag{3.37}$$

Os pontos são normalizadas para aumentar a robustez do algoritmo, e isso é feito mudando-se a origem do sistema de coordenadas ao centróide dos pontos e a escala dos pontos é uniformizada fazendo-se com que a média da distância dos pontos até a origem seja igual a $\sqrt{2}$. A matriz F encontrada deveria ter rank 2, o que normalmente não acontece devido ao ruído, portanto utiliza-se o SVD na matriz F para se obter a matriz fundamental que será considerada, F'.

Como não se sabe, a priori, os oito melhores pontos para se computar a matriz fundamental, usa-se o RANSAC (*RANdom Sample Consensus*, de Fischler e Bolles [22]), que é um método para se estimar parâmetros de um modelo matemático a partir de um conjunto de dados que contém *outliers*, para calcular a matriz através do algoritmo dos 8 pontos. Uma matriz fundamental candidata é computada a cada iteração do RANSAC e o limiar considerado para se decidir os pontos *inliers* para a matriz é 0,6% da dimensão máxima da imagem. Após isso, o erro é minimizado em todos os *inliers* da matriz através

da aplicação de um método de refinamento não-linear (Levenberg-Marquardt [45, 50]) nos oito parâmetros da matriz fundamental candidata. Então, são removidos os *matches* que são *outliers* da matriz fundamental refinada usando o mesmo valor de limiar usado para o RANSAC. Todos os *matches* são desconsiderados se o número de remanescentes for menor do que 20 para o par de imagens.

O que se faz, por fim, é organizar os *matches* geometricamente consistentes em *tracks*, ou seja, trilhas que registram as ocorrências de um determinado *keypoint* nas imagens, sendo que cada *track* deve conter pelo menos duas ocorrências do *keypoint* para ser válido.

3.2.2 Recuperação das coordenadas 3D dos pontos

O bundler é um método de *struct from motion* que difere do método da fatoração de Tomasi-Kanade por usar uma abordagem incremental de pontos para otimização da cena.

Após o fim da etapa anterior, temos definidos os pontos que se relacionam entre si nas imagens. O passo inicial para a reconstrução, propriamente, é a seleção de um par adequado de imagens para a estimativa inicial dos pontos e configurações das câmeras. O par ideal é aquele que possui o maior número de *matches*, mas que não pode ser bem modelado por uma única homografia (o que significa câmeras coincidentes ou de pontos de vista muito próximos), para uma reconstrução inicial robusta. Uma homografia é representada pela seguinte equação:

$$x' = Hx. \tag{3.38}$$

A forma usada para se estimar a matriz H, de tamanho 3×3, que representa a homografia para um par de imagens usa um método com etapas semelhantes ao do algoritmo dos 8 pontos normalizado, porém com apenas quatro pontos: as coordenadas são normalizadas; organiza-se a equação 3.38 em uma matriz de tamanho $2n \times 9$, com $n = 4$; e

os elemenos do autovetor de tamanho nove dessa matriz, obtido com o *SVD*, contém os elementos de H. Os pontos correspondentes nas duas imagens a serem usados são selecionados com o auxílio do RANSAC (aqui com um limiar de 0,4% da dimensão máxima da imagem para definir os *outliers*).

Obtendo-se a matriz H para todos os pares de imagens que contém correspondências, conforme definido na seção anterior, o par inicial é aquele com no mínimo 100 *matches* e a menor porcentagem de *matches* que representem homografia. Os parâmetros de câmera para o par escolhido são então estimados através do método dos cinco pontos [55].

Com as as correspondências dos pontos e as câmeras calibradas, podemos fazer a triangulação de todos os *tracks* visíveis nas duas imagens, o que resulta em uma estimativa da localização dos pontos em um sistema de coordenadas 3D, e um refinamento desses dados é feito em seguida através de um algoritmo de otimização não linear voltado especificamente para a otimização da estrutura 3D e dos parâmetros de câmera de uma cena chamado *bundle adjustment* [77], baseado no Levenberg-Marquardt [45, 50]. Este algoritmo minimiza o erro de reprojeção dos pontos observados, que é expresso como a soma dos quadrados de uma grande quantidade de funções reais não lineares, através da busca pelos parâmetros que melhor predizem a localização desses pontos nas imagens. A implementação de *bundle adjustment* usada é a *sparse bundle adjustment* de Lourakis e Argyros [48] [2].

Depois da inicialização, o algoritmo adiciona à otimização a câmera com o maior número de *matches* com os pontos 3D já estimados e também todas as câmeras com pelo menos 75% do número de *matches* da câmera adicionada. Neste caso, como estamos trabalhando com pontos cujas coordenadas 3D já foram estimadas, podemos usar o DLT [28] (*direct linear transformation* ou transformada linear direta), dentro de um procedimento RANSAC com limiar de 0,4% da dimensão máxima da imagem, para inicializar os parâmetros externos da câmera. O DLT é um algoritmo que resolve um conjunto de variáveis para um conjunto de relações de similaridade, nesse caso a projeção 3D dos

[2] http://www.ics.forth.gr/∼lourakis/sba/

pontos em uma imagem 2D baseado na condição de colinearidade dos pontos que formam a imagem (o ponto no espaço, o centro da projeção e o ponto projetado no plano de imagem). Uma matriz triangular superior K, que pode ser usada como uma estimativa dos parâmetros intrínsecos da câmera, é também retornada pelo DLT. Essa matriz K, juntamente com a distância focal extraída da *EXIF tag* (uma tabela embutida no arquivo da imagem), é usada para inicializar a informação de distância focal da nova câmera.

O refinamento dos pontos da cena observados pelas novas câmeras é feito, então, através da execução do *bundle adjustment*. Depois disso, os pontos que são observados pelas novas câmeras em pelo menos mais uma câmera recuperada são adicionados à otimização, desde que o par de raios com maior ângulo de separação que possa ser usado na triangulação desse ponto forme um ângulo maior que um dado limiar (o limiar escolhido foi de 2 graus), de forma que a triangulação garanta uma boa estimativa de sua localização. Após essa adição de novos pontos, é executado um *bundle adjustment* global para refinar todo o modelo.

Esse processo de adição e otimização de novas câmeras é repetido até que não reste nenhuma câmera que observe um número suficiente de pontos na cena (vinte pontos é o número mínimo para o bundler). Além disso, para aumentar a robustez do algoritmo, depois de cada otimização os pontos que apresentam erro de reprojeção maior que 16 pixels são excluídos e a otimização é executada novamente, até que mais nenhum *outlier* seja encontrado.

3.3 Resultados do shape from motion

Para uma melhor compreensão dos fatores que influem em cada um dos métodos, tanto no método da fatoração de Tomasi-Kanade, usando projeção ortogonal, quanto no do bundler, realizamos testes com cada um. Eles foram feitos de forma controlada para que possamos verificar os vários aspectos que afetam a reconstrução no *shape from motion*, como o ângulo total de rotação do objeto e o número de imagens capturadas. Os problemas

decorrentes da reconstrução com cada um dos métodos são também levantados e discutidos ao fim dos testes.

3.3.1 Método da fatoração

Para os testes de reconstrução usando o método da fatoração de Tomasi-Kanade, gravamos três sequências de vídeo em diferentes resoluções: 640×480, 800×600 e 960×720. A captura das sequências foi feita mantendo-se a câmera estática enquanto o objeto é rotacionado por um número de graus pré-estabelecido. Podemos ver dois quadros de exemplo de uma das sequências capturadas na *Figura 3.4*. Como optamos por rotacionar o objeto e manter a câmera estática, pontos do cenário que não deveriam participar da reconstrução acabam sendo rastreados também, e, por esse motivo, precisamos fazer uma filtragem dos pontos estáticos, e fazemos isso selecionando um limiar que define uma distância em pixels que um ponto deve percorrer na sequência para ser considerado um ponto do objeto. Na *Figura 3.5*, os pontos vermelhos representam aqueles considerados estáticos até o momento e os verdes são os pontos que já percorreram o número de pixels definido pelo limiar. Na imagem da direita, o último quadro da sequência, verificamos que a maioria dos pontos rastreados pertencentes ao objeto são verdes. A rotação total máxima utilizada foi de 45° para que um número razoável de pontos pudesse ser rastreado em todos os quadros, já que quanto maior o ângulo percorrido pela câmera do primeiro ao último quadro, maior o número de pontos que se perdem por oclusão. Devido ao fato da rotação do objeto ter sido feita manualmente, não foi possível um controle maior para o número de quadros capturados para cada rotação total, de forma que o número de quadros varia para cada sequência. Todas as três sequências foram gravadas com rotação total de 45° e para os testes com rotações menores mantiveram-se apenas os quadros do intervalo para a rotação desejada.

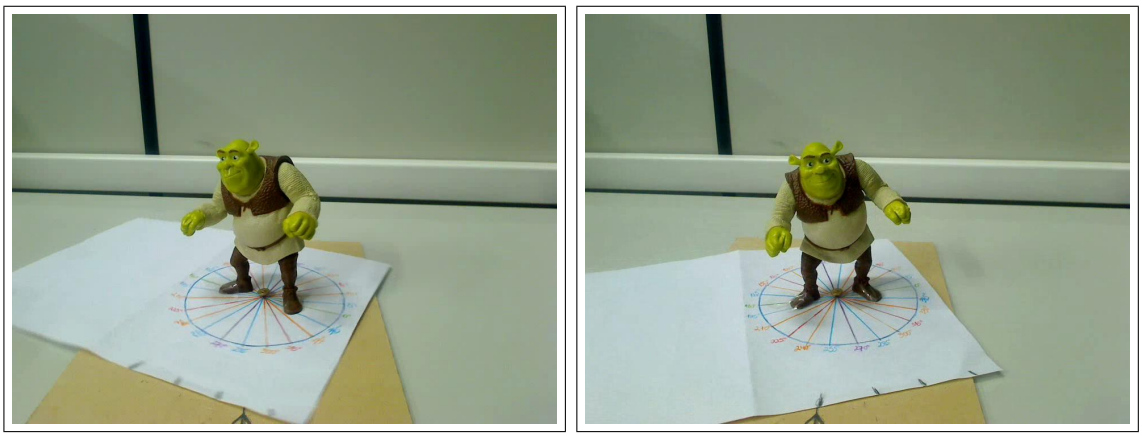

Figura 3.4: *Imagens da sequência com resolução 960×720*

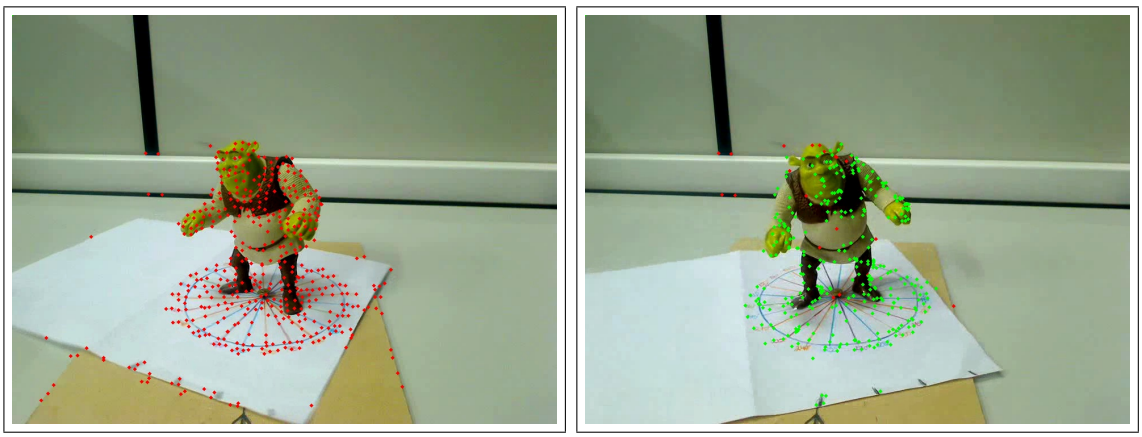

Figura 3.5: *Imagens com os pontos rastreados usando KLT*

O *KLT* foi usado para fazer o rastreamento dos pontos por ser um método simples e ao mesmo tempo suficiente, já que os quadros estão ordenados e contíguos, ou seja, apresentam pouca diferença entre cada quadro e seu próximo. Durante os testes, uma experiência feita de se rotacionar o objeto de 7,5° em 7,5° até o total de 45°, resultando em uma sequência com 7 quadros, gerou um péssimo resultado devido à dificuldade do *KLT* em rastrear os pontos por eles mudarem muito de posição de um quadro ao próximo do vídeo. Ao olhar as duas imagens na *Figura 3.5*, percebe-se facilmente que alguns pontos rastreados na primeira não estão presentes na segunda. A oclusão de pontos é algo comum, e como o *SVD* exige matrizes completas, devemos tratar de alguma forma este problema. Neste caso, especificamente, optamos apenas por ignorar os pontos que não puderam ser visualizados em todos os *frames* da sequência. O *KLT* nos permite selecionar o número n de pontos iniciais que desejamos rastrear, ou seja, os n pontos

	Rotação total							
	45°		30°		15°		7,5°	
Resolução	Frames	Vért.	Frames	Vért.	Frames	Vért.	Frames	Vért.
640×480	43	235	30	254	22	290	15	364*
800×600	36	233	24	241	16	273*	10	380
960×720	26	352	17	364	11	416	8	425

Tabela 3.1: Resolução×Rotação

mais diferenciáveis da imagem, segundo o *KLT*. No decorrer da sequência pontos que se tornam ruins ou que não são visualizados são excluídos do rastreamento e não voltam a ser considerados. Para todos os testes inicializamos o número n de pontos do *KLT* como 500.

Podemos ver o resultado numérico dos testes na *Tabela 3.1*. Os campos em que o número de vértices é seguido pelo * indicam os casos de reconstrução pobre, sem motivo aparente, em que a profundidade não fica muito evidente ou os pontos indicam um objeto deformado. Analisando os testes podemos verificar que, ao menos em se tratando do número de pontos, o resultado é muito próximo entre o vídeo de resolução 640×480 e o de 800×600. Apesar da diferença no número de quadros entre cada caso de teste dos vídeos dessas duas resoluções o número de vértices reconstruídos ao final é bem próximo para ambos. Quando comparamos os testes realizados com o vídeo de 960×720 com os outros percebemos que ele se saiu bem melhor, já que o número de vértices em todos os casos é bem superior. A rotação total do objeto não afetou a qualidade da reconstrução, ou seja, mesmo quadros que compreendem uma pequena rotação, como 7,5°, que é nossa rotação total mínima, conseguimos recuperar a informação de profundidade dos pontos. Outra informação que podemos extrair da tabela é que, indiscutivelmente, quanto menos quadros temos em nossa sequência mais pontos restam ao final, já que menos pontos são ocluídos ou se perdem durante o rastreamento. Dessa forma, o melhor dos casos dentre todos os nossos testes foi o vídeo com resolução 960×720, com oito quadros representando uma rotação total de 7,5° do objeto, que resultou em uma nuvem de 425 pontos, ou seja, apenas 75 pontos se perderam durante o rastreamento.

A nuvem de pontos resultante da reconstrução do nosso melhor caso pode ser vista

na *Figura 3.6* e a superfície 3D possível de se conseguir com a nuvem de pontos pode ser vista na *Figura 3.7*. Consideramos que o vetor vermelho está apontando para o eixo X, o azul para o Y e o verde para o Z, de forma a facilitar a visualização da orientação no espaço 3D.

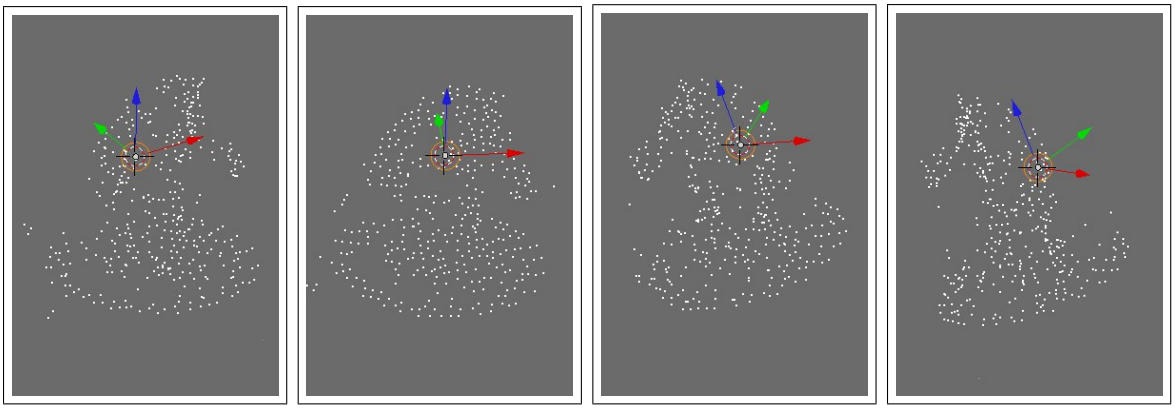

Figura 3.6: *Representação 3D dos pontos do objeto*

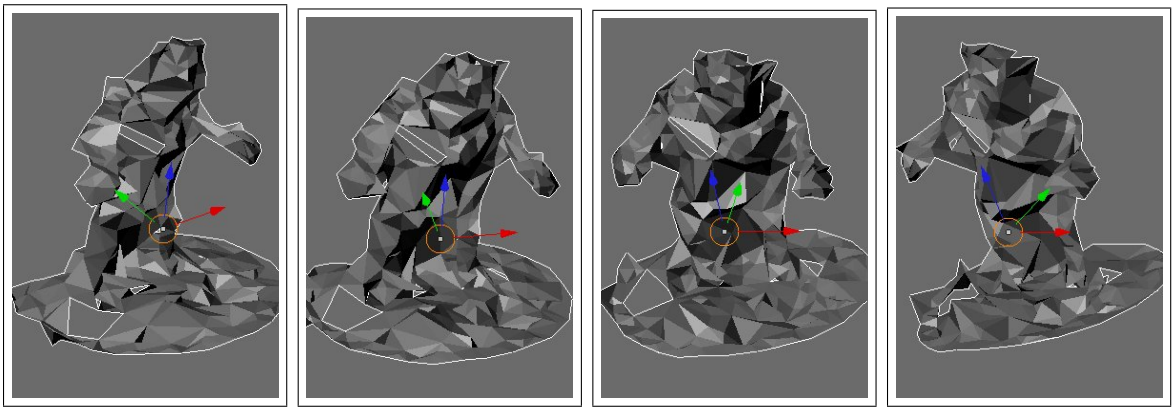

Figura 3.7: *Representação 3D da superfície do objeto*

Diversos fatores dificultam o uso do método da fatoração como o testado nesta seção para a obtenção de uma reconstrução realista, e os principais pontos que levam a isso estão na etapa do rastreamento. O resultado acaba de certa forma comprometido devido ao ruído comum dos pontos rastreados pelo *KLT*, como podemos observar na *Figura 3.8*, e também pela impossibilidade de rastreamento em regiões do objeto com pouca textura. Outro fator, menos relevante, é o fato de usarmos projeção ortogonal, que nos força a nos afastarmos do objeto de modo que a perspectiva não fique tão acentuada, o que nos obriga também a aumentar a resolução do vídeo para tornar possível o rastreamento dos pontos

do objeto, que acaba ficando distante e pequeno na cena. Além disso, a reconstrução de um objeto inteiro é trabalhosa, já que várias pequenas sequências são necessárias e, como os dados resultantes não são muito densos nem tão acurados, a junção das partes se torna difícil.

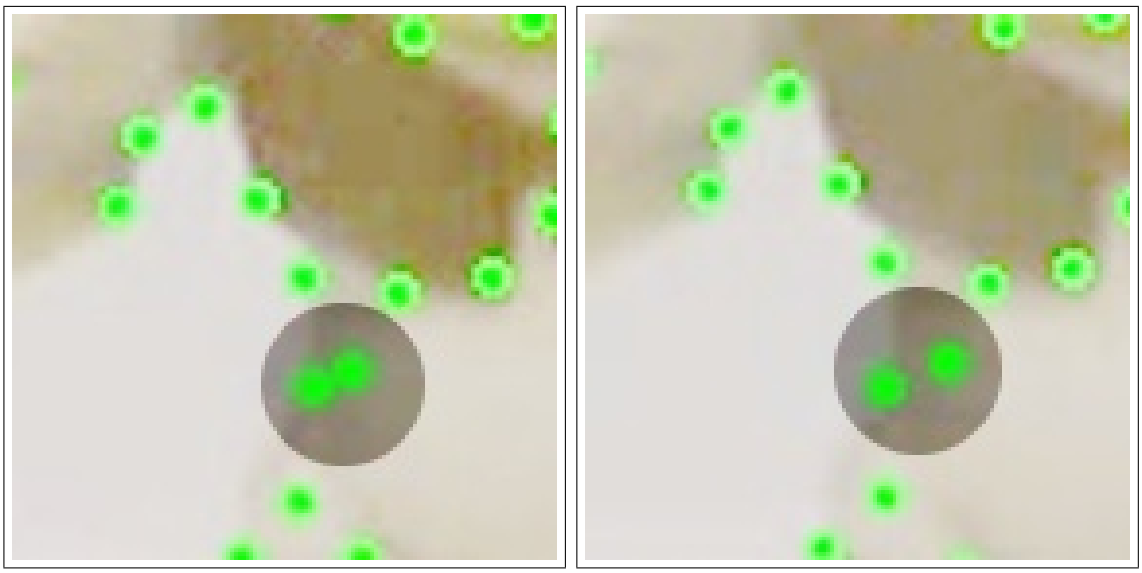

Figura 3.8: *Destaque do ruído durante o rastreamento, em dois quadros contíguos*

Apesar dos problemas citados o resultado obtido é interessante, sendo fácil reconhecer a forma do objeto nos pontos ou na superfície reconstruída. Porém, devido ao número reduzido de pontos obtidos no rastreamento pelos motivos já citados anteriormente, a possibilidade de usar tal resultado, sem o auxílio de qualquer outro método passivo de reconstrução 3D, como a representação de um modelo para preservação digital, é desconsiderada.

3.3.2 Bundler

Os testes para o método do bundler foram conduzidos de forma similar ao método da seção anterior. Três sequências de diferentes resoluções com diferentes número de quadros foram utilizadas, sendo a de 640×480 e a de 800×600 as mesmas utilizadas para o método da fatoração, e a terceira de resolução 3072×2304 fotografada quadro a quadro por uma câmera de melhor qualidade, o que acabou resultando em uma diferença na ilu-

mação entre os quadros impossibilitando o uso dessa sequência para o teste do método da fatoração, que usa o *KLT* que não é invariante quanto à iluminação (*Figura 3.9*).

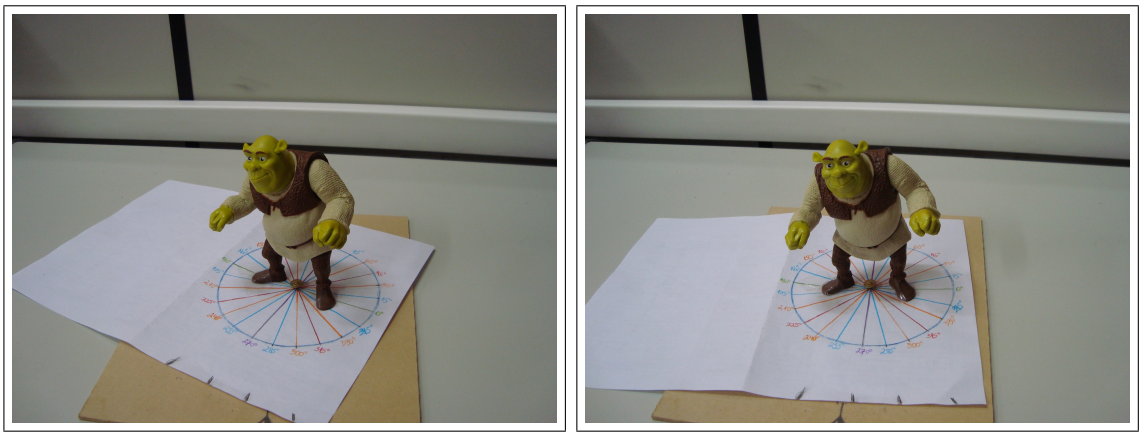

Figura 3.9: *Imagens da sequência com resolução 3072×2304*

O resultado numérico dos testes do bundler podem ser vistos nas *Tabelas 3.2 e 3.3*. Da mesma forma que nos testes da seção anterior, o * logo depois do número de vértices obtidos com a reconstrução indica um resultado ruim, como um objeto deformado ou com pouca profundidade. Quase todos os casos de teste com resolução menor resultaram em uma reconstrução ruim, possivelmente pela baixa qualidade das câmeras utilizadas no processo de captura, já que o algoritmo do bundler se baseia, por exemplo, em informações como distância focal fornecidas pela própria câmera com relação às imagens tiradas. Apesar disso, o número de vértices foi, quase em todos os casos, maior do que os obtidos com o método da fatoração, até porque o bundler só rejeita pontos que considera ruins para a otimização, e os pontos não precisam estar visíveis em todos os quadros para que sua coordenada 3D seja encontrada, bastando que esteja visível em pelo menos duas imagens. Novamente, os melhores resultados foram obtidos com resoluções mais altas e, ao contrário do método da fatoração, quanto mais quadros, maior é o número de vértices, já que novos pontos tendem a ser adicionados à cena a cada nova imagem utilizada na otimização. Um problema que encontramos aqui, porém, é que quanto maior a resolução das imagens e maior o número de imagens utilizadas, maior é o tempo gasto no processo de reconstrução. O tempo necessário para cada um dos casos usando as imagens de maior resolução é muito superior aos casos análogos com resoluções menores. A rotação total

	Rotação total					
	45°			30°		
Resolução	Frames	Vért.	Tempo	Frames	Vért.	Tempo
640×480	43	1401	6m45.686s	30	484*	2m35.767s
800×600	36	1094*	4m41.530s	24	1176*	2m22.405s
3072×2304	41	10387	72m54.762s	28	7170	37m9.105s

Tabela 3.2: Resolução×Rotação×Tempo (1)

	Rotação total					
	15°			7, 5°		
Resolução	Frames	Vért.	Tempo	Frames	Vért.	Tempo
640×480	22	712	1m33.664s	15	411*	0m50.057s
800×600	16	854*	1m17.673s	10	295*	0m41.298s
3072×2304	16	4596	27m58.803s	8	2485	11m41.067s

Tabela 3.3: Resolução×Rotação×Tempo (2)

não teve influência na profundidade do objeto, ou seja, para o caso de maior resolução, tanto com a rotação total de 45° quanto com a de 15°, a profundidade do objeto parece correta, interferindo no resultado apenas o número de quadros utilizados. Portanto, o melhor resultado obtido nos testes foi aquele que utilizou 41 imagens com resolução 3072×2304 cobrindo a rotação total de 45° que recuperou as coordenadas 3D de 10387 pontos, levando para tal um tempo de 72m54.762s.

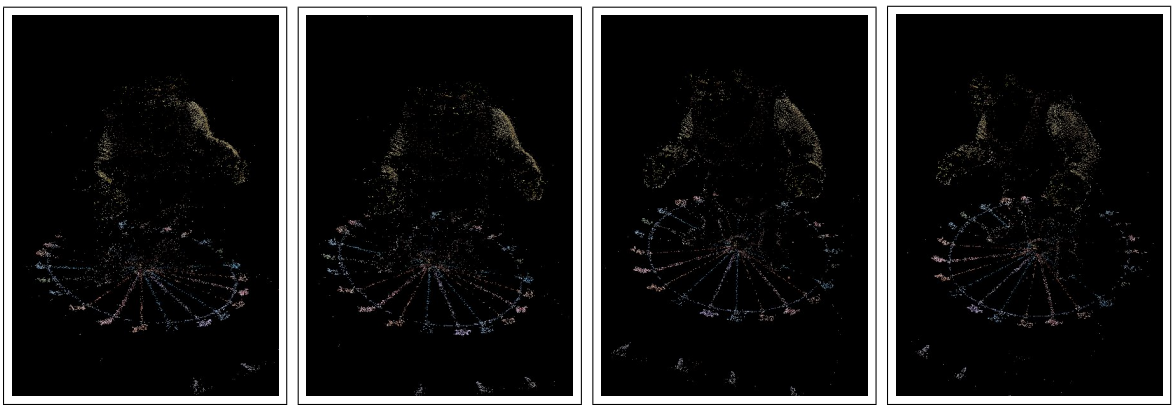

Figura 3.10: *Representação 3D dos pontos do objeto*

Podemos conferir o resultado da reconstrução do melhor caso na *Figura 3.10*. O arquivo gerado pelo bundler contém a informação da cor de cada vértice, portanto a visualização do resultado se torna mais fácil. Percebe-se que, apesar de termos regiões grandes sem pontos recuperados (a região próxima a barriga do modelo, por exemplo), algumas

outras regiões são formadas de nuvens bem densas e acuradas de pontos. O resultado obtido com o bundler é mais correto devido à utilização de um método de rastreamento mais robusto, à qualidade superior das imagens, à adoção de uma projeção em perspectiva e à natureza de otimização do método. O erro é reduzido a cada imagem nova introduzida no processo, e o número de pontos rastreado em cada imagem individualmente é extremamente superior ao obtido com o *KLT*. O tempo para este resultado é, porém, extremamente alto.

A mesma experiência feita para o método da fatoração de tirar uma sequência de 7 imagens rotacionando o objeto de 7,5° em 7,5° resultando em uma rotação total de 45° foi feito para o método do bundler, gerando neste caso, porém, um resultado positivo. Devido ao fato do *SIFT* ser um método de rastreamento mais robusto, não houve problemas no rastreamento dos pontos apesar da diferença brusca entre o ângulo de rotação do objeto nas imagens contíguas. O resultado, que recuperou 1833 vértices, pode ser visto na *Figura 3.11*. Apesar do grande número de pontos, estes encontram-se concentrados na face, nos braços e na base do modelo.

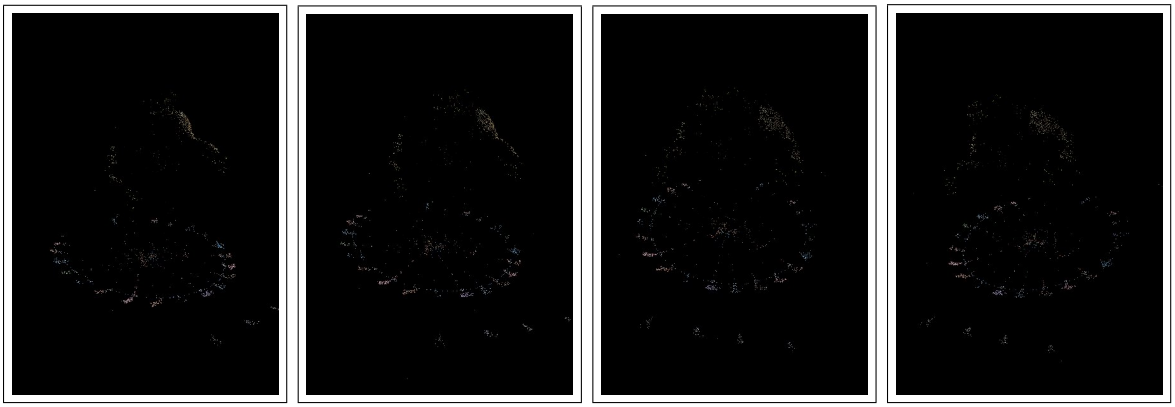

Figura 3.11: *Representação 3D dos pontos do objeto*

Da mesma forma que o método da fatoração, o bundler possui algumas deficiências. Grandes áreas sem textura continuam sem poder ser rastreadas, mesmo com a utilização de um método mais robusto, e também o tempo necessário para uma reconstrução é extremamente alto. Como pontos positivos temos o fato de que é possível se reconstruir o objeto inteiro de uma só vez, sem a necessidade de se fazer a união das várias vistas, pois

o bundler é um método que adiciona novos pontos na cena como um todo, baseado nos pontos em comum. A qualidade da reconstrução também é superior àquela obtida com o método da fatoração, porém sendo ainda apenas um bom método para uma estimativa inicial da superfície e não como o método que gera a superfície final do modelo real para sua preservação digital.

3.3.3 Considerações finais

Pelo motivo de ambos os métodos serem limitados pelo número de pontos rastreados, não vemos uma aplicação imediata de nenhum deles para o propósito de preservação digital. Como discutido, porém, no capítulo anterior, sobre os métodos híbridos, utilizaremos o método de *shape from motion* apenas como base para a reconstrução e auxílio na resolução das ambiguidades encontradas na superfície. Dessa forma, ambos os métodos são suficientes para este fim, sendo o método da fatoração mais simples e mais rápido na obtenção da localização 3D dos pontos rastreados. No próximo capítulo serão mostrados resultados quantitativos na comparação do método híbrido com o *ground truth* dos objetos testados.

CAPÍTULO 4

PHOTOMETRIC STEREO E O PHOTOMETRIC SHAPE FROM MOTION

A ideia de se decompor a matriz de intensidades da imagem (contendo a variação da intensidade de cada pixel no decorrer do tempo) nas matrizes B e L, de superfície e iluminação, utilizando-se do *SVD*, foi proposta primeiramente em 1994 por Hayakawa [29]. Esta ideia já não é novidade para nós. No capítulo anterior foi explicada a fatoração de uma matriz de medidas de pontos rastreados em matrizes de rotação e distribuição espacial desses mesmos pontos em uma cena tridimensional. Ora, a ideia básica é a mesma: sabemos que a intensidade dos pixels de um objeto lambertiano na cena é uma função do produto da geometria e refletividade de sua superfície com a direção e intensidade das luzes que atingem esta mesma superfície.

Neste capítulo iremos entender como a ideia da fatoração de uma matriz com informações da cena é estendida ao conceito da refração da luz na superfície de objetos de forma a recuperar-se a geometria e albedo dessa mesma superfície, bem como a utilização do *shape from motion* para a resolução das ambiguidades inerentes a esse método.

4.1 Photometric stereo

4.1.1 A matriz de intensidades

A matriz de intensidades I é montada de forma a representar a intensidade em cada ponto dentro da região que faz parte do objeto em cada quadro de uma sequência de tamanho F, de acordo com o modelo lambertiano (luz incidente é espalhada igualmente em qualquer ponto da superfície, não ocorrendo brilho), assumindo-se projeção ortogonal e fontes de

luz distantes do objeto:

$$I(\mathrm{x}, \mu) = a(\mathrm{x})\mathrm{n}(\mathrm{x}) \cdot \mathrm{l}(\mu) \equiv \mathrm{b}(\mathrm{x}) \cdot \mathrm{l}(\mu). \tag{4.1}$$

Podemos considerar b(x) como equivalente a a(x)n(x), sendo a(x) o albedo, ou quantidade de luz refletida em determinado ponto, e n(x) a normal, efetivamente, deste ponto. Já l(μ) é a representação do vetor de iluminação.

Apesar dos pontos serem dispostos em coordenadas bidimensionais em cada quadro da sequência, é mais adequado a nossas necessidades que aqui sejam representados como um vetor de intensidades, que comporá cada linha da nossa matriz I. Portanto, a matriz I tem dimensões $r \times F$, sendo r o número total de pontos que se encontra na região da imagem da superfície cujas normais se deseja encontrar (dentro de uma máscara definida, por exemplo) e F o número de quadros, refletindo portanto no número de diferentes orientações de iluminação, que varia a cada quadro com relação à essa superfície. No *photometric stereo* o objeto e a câmera se mantém estáticos enquanto a fonte de iluminação é movida para gerar a diferença de intensidade em cada ponto, necessária para a reconstrução 3D.

Da mesma forma que no método de Tomasi-Kanade para *shape from motion*, podemos considerar a matriz de intensidades como de *rank* 3, se admitirmos a superfície do objeto lambertiana e desconsiderarmos a iluminação ambiente (no caso de levarmos em conta essa iluminação, uma matriz de intensidades de *rank* 3 não modela bem o problema, e o *rank* da matriz aumenta para 4). Para o caso em que se conhece os vetores de iluminação previamente, os vetores b_i(x) são obtidos diretamente e de forma única, porém assumir tal conhecimento restringe em muito o cenário e as possibilidades durante a aquisição das imagens. Podemos, ao invés disso, através da minimização da seguinte função de energia

$$E[\mathrm{b},\mathrm{l}] = \sum_{\mu,\mathrm{x}} \{I(\mathrm{x},\mu) - \sum_{i=1}^{3} b_i(\mathrm{x}) l_i(\mu)\}^2 \tag{4.2}$$

em função de b e l, obtermos a matriz B de dimensões $r \times 3$, em que cada linha representa o vetor pseudo normal e o albedo de um ponto x na superfície do objeto e a matriz L de dimensões $3 \times F$, em que cada coluna representa o vetor da pseudo iluminação (direção e intensidade da luz). Essa minimização pode, portanto, ser feita até uma certa transformação linear 3×3, a matriz A (devido a uma ambiguidade na equação lambertiana representada na *Equação 4.3*).

$$\mathrm{b} \cdot \mathrm{l} = \mathrm{b}^{\mathrm{T}}\mathrm{l} = \mathrm{b}^{\mathrm{T}}\mathrm{A}\mathrm{A}^{-1}\mathrm{l}. \tag{4.3}$$

O melhor resultado para essa minimização pode ser obtido através do *SVD*. Podemos reescrever I, então, como a seguinte matriz de tamanho $r \times F$, J (admitindo que $r \geq F$, fazendo-se necessário transpor-se I em caso contrário):

$$\mathrm{J} = \mathrm{UDV} \tag{4.4}$$

onde $\mathrm{U}^{\mathrm{T}}\mathrm{U} = V^T V = V V^T = Id$, sendo Id a matriz identidade de tamanho $F \times F$, U de tamanho $r \times F$, e V e D de tamanho $F \times F$.

Da mesma forma que fizemos no método da fatoração já explicado anteriormente, dividiremos J em duas partes:

$$UDV = \mathrm{e}(\mathrm{x})D'\mathrm{f}(\mu) + U''D''V'' \tag{4.5}$$

onde

$$U = \left[\begin{array}{c|c} \text{e(x)} & U'' \end{array} \right] \tag{4.6}$$

e e(x) tem tamanho $r \times 3$ e U'' tem tamanho $P \times (F-3)$,

$$D = \left[\begin{array}{c|c} D' & 0 \\ \hline 0 & D'' \end{array} \right] \tag{4.7}$$

e D' tem tamanho 3×3 e D'' tem tamanho $(F-3) \times (F-3)$ e

$$V = \left[\begin{array}{c} \text{f}(\mu) \\ \hline V'' \end{array} \right], \tag{4.8}$$

em que f(μ) tem tamanho $3 \times F$ e V'' tem tamanho $(F-3) \times F$.

No caso perfeito sabemos que D', matriz diagonal de tamanho 3×3, contém os três valores singulares não nulos (apesar de isso não ser exatamente verdade devido aos ruídos, luz ambiente, brilho etc), portanto pegamos das matrizes U e V apenas as linhas e colunas referentes a esses valores de D. Dessa forma, definimos b(x) e s(μ) como se segue:

$$\text{b(x)} = \text{e(x)}\text{P}_3, \forall \text{x}, \tag{4.9}$$

$$\text{l}(\mu) = \text{Q}_3\text{f}(\mu), \forall \mu, \tag{4.10}$$

de forma que $P_3^T Q_3 = D'$. Porém, P_3 e Q_3 não são únicas, devido a ambiguidade já citada ($P_3 \mapsto P_3 A$ e $Q_3 \mapsto A^{-1} Q_3$). Precisamos, portanto, nos basearmos em alguma restrição para recuperar b(x), e consequentemente n(x), e f(μ) de forma única.

4.1.2 Restrições para recuperação das normais e iluminação de forma única

Em uma tentativa de se descobrir os vetores normais e de iluminação de forma única, Hayakawa sugere algumas restrições com relação à superfície ou à iluminação: encontrar-se ao menos seis pixels nos quais a refletividade é constante ou conhecida; ou encontrar-se seis quadros nos quais a intensidade da iluminação é, da mesma forma, constante ou conhecida [29]. Apesar dessas restrições, temos uma solução correta até uma certa rotação apenas, admitida por Hayakawa como sendo a matriz identidade. Essa suposição, porém, não é verdadeira para todos os conjuntos de dados, podendo então falhar.

Outra restrição, a de integrabilidade da superfície [12], implica que deve haver uma consistência das normais com a superfície do objeto. Nesta restrição, os vetores normais unitários da superfície do objeto ($\mathrm{n(x)} = (n_1(\mathrm{x}), n_2(\mathrm{x}), n_3(\mathrm{x}))$) devem estar consistentes de forma a gerar uma superfície. Sem perda de generalidade, podemos impor essa restrição de n(x) a b(x). No trabalho de Yuille e Snow [84] essa restrição, bem como seu uso no *photometric stereo*, é explicada:

$$\frac{\partial}{\partial x}\left(\frac{\mathrm{b_2(x)}}{\mathrm{b_3(x)}}\right) = \frac{\partial}{\partial y}\left(\frac{\mathrm{b_1(x)}}{\mathrm{b_3(x)}}\right), \tag{4.11}$$

que pode ser expandida para

$$b_3\frac{\partial b_2}{\partial x} - b_2\frac{\partial b_3}{\partial x} = b_3\frac{\partial b_1}{\partial y} - b_1\frac{\partial b_3}{\partial y}. \tag{4.12}$$

Assim, z, ou seja, a profundidade, será uma função da coordenada bidimensional do ponto, ou $f(x, y)$.

Essa restrição nos garante, mais uma vez, uma solução até uma determinada trans-

formação, representada por

$$z = \lambda f(x, y) + \mu x + \nu y, \tag{4.13}$$

ou seja, uma escala em z por um fator de λ e a adição de um plano $\mu x + \nu y$.

Esta é conhecida como ambiguidade de *bas-relief* generalizada (*GBR, generalized bas-relief ambiguity*, em inglês) [9], que consiste em não podermos identificar a real profundidade de um objeto quando o vemos de um ponto de vista único, ou com variação muito pequena, e não podemos inferir a posição da fonte de iluminação. Resumindamente, dois objetos com profundidades diferentes podem parecer iguais de um ponto de vista se as luzes forem posicionadas adequadamente para induzir esse efeito. A *Figura 4.1* ilustra o problema.

Algumas estratégias são utilizadas para se resolver essa ambiguidade, normalmente envolvendo suposições sobre a iluminação [84], a refletividade [29] ou a geometria [14] da superfície. O método sugerido por Alldrin et al. [4] se baseia no fato de que muitos objetos no mundo real são compostos de um pequeno grupo finito de valores de albedo, portanto é proposto um método que busca encontrar os valores de λ, μ e ν que minimizam a entropia da distribuição dos albedos.

A estratégia usada neste trabalho, porém, seguindo o método híbrido de Lim et al. [47], irá se basear em pontos previamente recuperados através de um método de *shape from motion* e na minimização da distância entre a superfície ambígua retornada pelo *photometric stereo* e esses pontos.

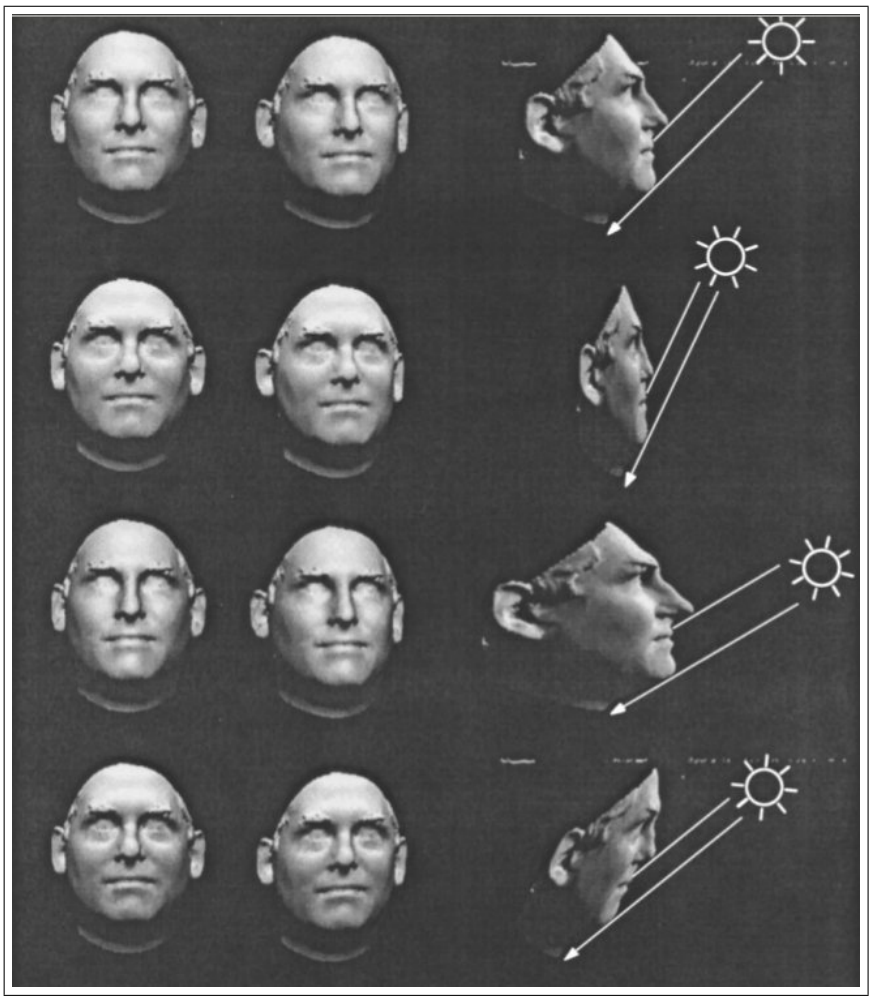

Figura 4.1: *Ambiguidade do GBR. (fonte [9])*

4.1.3 Como se chegar às superfícies a partir dos vetores normais

Partindo de b(x), devemos primeiro separar o albedo (a(x)) das normais (n(x)), e trabalharmos apenas com a última para a geração da superfície. O albedo pode ser obtido de maneira razoavelmente simples e eficaz, dividindo-se a normal de cada ponto pela média de sua intensidade no decorrer da sequência. Portanto, para cada vetor x representado em b, n pode ser extraído de b fazendo-se

$$a(\mathrm{x}) = \sqrt{\mathrm{b_1(x)}^2 + \mathrm{b_2(x)}^2 + \mathrm{b_3(x)}^2} \tag{4.14}$$

$$\mathrm{n(x)} = \frac{\mathrm{b(x)}}{a(\mathrm{x})}. \tag{4.15}$$

Em [8], Basri e Jacobs explicam um método usado para a recuperação da profundidade a partir das normais, contanto que elas sejam integráveis. Como vimos, ao seguirmos as restrições de intregrabilidade, n(x) é um conjunto integrável de normais. Definimos então $z(x, y)$ como a nossa superfície, sendo as direções das normais $n(x, y) = (p, q, -1)$, com $p = z_x$ e $q = z_y$, as derivadas parciais das coordenadas x e y no grid da imagem, respectivamente.

$$p \approx z(x + 1, y) - z(x, y) \tag{4.16}$$

$$q \approx z(x, y + 1) - z(x, y), \tag{4.17}$$

ou então, se considerarmos as normais

$$p = -\frac{n_x}{n_z} \tag{4.18}$$

$$q = -\frac{n_y}{n_z}. \tag{4.19}$$

A relação dos dois conjuntos de equação nos dá então as restrições a serem seguidas:

$$n_z z(x, y) - n_z z(x + 1, y) = n_x \tag{4.20}$$

$$n_z z(x, y) - n_z z(x, y + 1) = n_y. \tag{4.21}$$

Apesar de uma simples minimização linear, o problema deve normalmente ser resolvido computacionalmente com o auxílio do uso de um método iterativo de minimização linear para matrizes esparsas, já que o conjunto de dados é geralmente muito grande para ser mantido na memória RAM em uma matriz completa comum.

4.2 A combinação de shape from motion e photometric stereo para a reconstrução de superfícies

Como dito já no *Capítulo 2*, a combinação de mais de um método de reconstrução 3D funciona de forma que um método auxilie e corrija o outro para a obtenção de melhores resultados enquanto sobrepõem-se algumas das restrições e desvantagens de cada um dos métodos individualmente. O *shape from motion* se mostrou interessante para a reconstrução de poucos pontos e recuperação da matriz de projeção da câmera no decorrer dos quadros do vídeo. O método de Tomasi-Kanade [76] é um método rápido e esbarra em pouquíssimas restrições quanto a captura das imagens, não exigindo sequer qualquer calibração prévia, sendo limitado porém ao número de pontos rastreados em cada quadro (superfícies pouco texturizadas refletem também em poucos pontos rastreados). O método de *photometric stereo* para objetos de superfície lambertiana (sem refletividade especular) e fontes de iluminação distantes do objeto através do *SVD* [29, 84] sofrem da ambiguidade de *GBR*, porém resultam em superfícies densas, limitadas apenas pela resolução das imagens.

Implementamos então um método híbrido iterativo, proposto por Lim et al. [47], que combina as vantagens do método de Tomasi-Kanade e do *photometric stereo*, ao mesmo tempo em que aquele ajuda na correção da ambiguidade de *GBR* deste.

4.2.1 Photometric stereo from motion

O *photometric stereo* parte do princípio de que tanto o objeto quanto a câmera permanecem estáticos enquanto a direção da fonte de iluminação varia em cada imagem. Dessa forma, não temos problemas com o rastreamento dos pontos e podemos considerar todos os pontos da superfície, porém somos pegos pela ambiguidade inerente a este método. No *shape from motion* temos um objeto que se move em relação à câmera no decorrer dos quadros, possibilitando portanto a reconstrução 3D apenas de pontos passíveis de ser rastreados em pelo menos alguns quadros (no caso do método de Tomasi-Kanade como foi

implementado aqui os pontos devem ser rastreados em todos os quadros). Sendo assim, como se pode mesclar as duas estratégias? Ora, com o método de Tomasi-Kanade, recuperamos, além da posição 3D dos pontos rastreados, a matriz de projeção ortogonal em cada quadro. Isso viabiliza uma aproximação da reprojeção de qualquer ponto inserido em uma região (definida por uma máscara, por exemplo) do primeiro quadro, para sua posição 2D em qualquer quadro da sequência. A profundidade aproximada de cada ponto dentro da região definida pode ser facilmente obtida a partir das coordenadas 3D dos pontos já retornados pelo método.

4.2.2 Criando-se um mapa de profundidade aproximado a partir do shape from motion

Após a execução do *shape from motion* (como visto na *Seção 3.1* do capítulo anterior) temos uma matriz com os F vetores P_j, $1 \leq j \leq F$ de projeção ortogonal, sendo F o número de quadros da sequência (além disso, precisaremos também dos vetores t_j, que representam a translação feita no início do método para mover os pontos até seu centróide) e as posições 3D $[x_p, y_p, z_p]^T$ dos m pontos reconstruídos. Consideremos z_i como a profundidade do ponto (x_i, y_i) com relação ao primeiro quadro f_1 da sequência.

Admitimos R como sendo a região que contém todos os r pontos do objeto que serão usados na matriz de intensidade I. Na nossa implementação, a borda de R é composta pelos pixels em m mais distantes do centróide dos pontos, ou seja, a borda é também formada por pontos cuja profundidade é conhecida através do *shape from motion*. A superfície inicial S_0 representada pelo mapa de profundidade é criada então através de uma triangulação dos m pontos em f_1. A triangulação de Delaunay [44] 2D foi escolhida nesse caso (já que os pixels estão posicionados no grid 2D da imagem). As arestas mais externas ao objeto farão, portanto, parte da borda de R.

Temos, assim, a profundidade em alguns pontos de R; os vértices dos triângulos. Precisamos porém da profundidade de todos os pontos em R para sua reprojeção em

todos os F quadros da sequência. Para uma aproximação da profundidade (lembre-se que estamos apenas trabalhando no mapa de profundidade inicial S_0), fazemos uma interpolação bilinear da profundidade a partir da profundidade dos três vértices de cada triângulo. O resultado é, portanto, o mapa de profundidade inicial z_0.

4.2.3 Criando a matriz de intensidades e resolvendo o photometric stereo

A matriz de intensidades I, de tamanho $r \times F$, é construída da seguinte forma:

$$I_{ij} = f_j(P_j(x_i, y_i, z_i(x_i, y_i)) + t_j) \tag{4.22}$$

sendo $1 \leq i \leq r$.

Assim, cada coluna j de I é formada pela intensidade do respectivo ponto i reprojetado do primeiro quadro no quadro j.

Tendo a matriz I montada, podemos apenas aplicar o *photometric stereo* como visto na *Seção 4.1* e obteremos uma nova superfície $\overline{S_1}$, definida pelo mapa de profundidade $\overline{z_1}$. Esse mapa provavemente não representa a superfície correta do objeto devido à ambiguidade de *GBR*. Temos, porém, os m pontos recuperados pelo *shape from motion* que consideramos estar em conformidade com a profundidade real dos pontos do objeto e podemos corrigir $\overline{z_1}$ encontrando a transformação *GBR* que minimize a função D de distância dos pontos respectivos em $\overline{z_1}$ com relação à profundidade dos m pontos.

$$D = \sum_{i=1,p=1}^{m} \{(\lambda \overline{z_i}(x_i, y_i) + \mu x_i + \nu y_i) - z_p\}^2 \tag{4.23}$$

Dessa forma a ambiguidade *GBR* é resolvida e obtemos o mapa de profundidade

z_1 que representa a superfície S_1, com a profundidade dos pixels em R mais próximas da profundidade real. Assim, podemos obter uma nova matriz de intensidades I mais correta, já que a reprojeção dos pixels pode ser melhor calculada devido a correção nas profundidades. Essas etapas podem ser repetidas até a convergência da superfície S na superfície do objeto real, através de um limiar ou uma condição de otimização mínima entre uma iteração e outra. Um diagrama ilustrando as etapas do método híbrido pode ser visto na *Figura 4.2*.

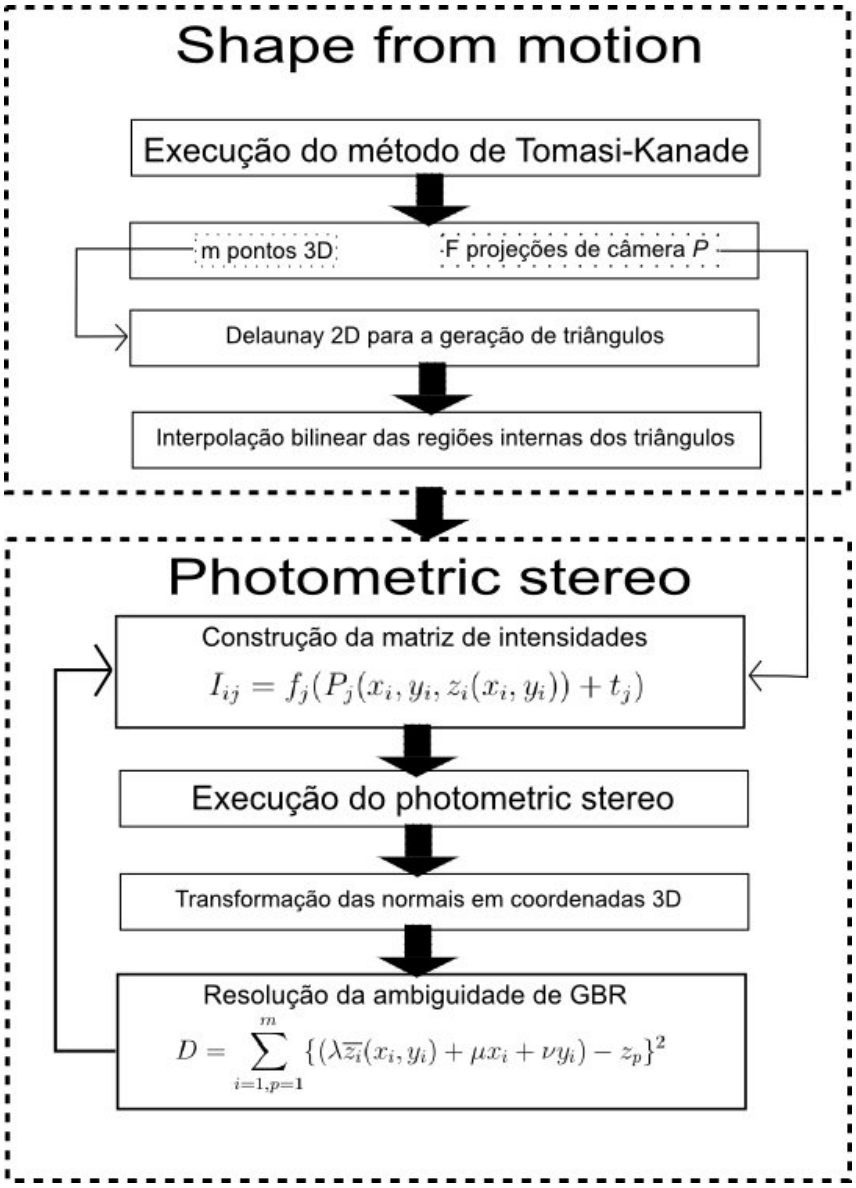

Figura 4.2: *Diagrama das etapas do método de photometric stereo from motion.*

4.3 Testes e resultados do photometric shape from motion

Nesta seção expomos alguns resultados experimentais do método implementado. O programa foi feito utilizando a linguagem *C* e o computador utilizado nos testes é um laptop dual-core de 2.53 GHz e 4GB de memória RAM. Os pontos iniciais, para o método ortográfico da fatoração de Tomasi-Kanade, foram selecionados manualmente em cada quadro, de forma a obtermos o melhor resultado possível para esses pontos e rotações de câmera mais corretas. Como nosso interesse, agora, é apenas testar as capacidades do método de Lim et al., a seleção manual dos pontos não representa um grande problema; porém, para a aplicação do método em casos reais, a seleção de pontos se torna inconveniente e devemos procurar um método automático adequado (como citado no capítulo anterior, o método *KLT* [68] se apresenta como um bom método de rastreamento para o método do Tomasi-Kanade). A quantidade de pontos não é um fator fundamental, servindo apenas para limitar as bordas da superfície e, os poucos pontos internos a essas bordas, para uma aproximação um pouco mais correta da superfície inicial, melhorando a reprojeção desde a primeira iteração. Esses pontos serão utilizados para se minimizar a distância com a superfície retornada pelo *photometric stereo*, sendo 4, portanto, o número mínimo de pontos devido ao número de variáveis envolvido na minimização (1 parâmetro de translação e 3 da transformação *GBR*).

As sequências de vídeo foram capturadas com uma câmera digital SLR (DSLR, *digital single-lens reflex*, em inglês) com uma lente de distância focal longa, permitindo assim a captura a partir de uma distância maior entre o objeto e a lente da câmera, mais correta à projeção ortográfica, e uma resolução aproximada de 640×480 (com exceção da sequência *lady*, obtida no website do método de Zhang et al. [85][1]). Em todas as sequências o objeto é movido em frente à câmera, que se mantém estática, bem como a fonte distante de iluminação, gerando assim a diferença de movimento necessária para o *shape from motion* e a diferença da direção de iluminação para o *photometric stereo*. Um exemplo dos quadros da sequência *flor* pode ser visto na *Figura 4.3*.

[1] grail.cs.washington.edu/projects/stshading/

Objeto	**Pts$_0$**	**Quad**	**Pts**	**Tempo/Iter**	**Iter**	**RMSD$_0$**	**RMSD$_\omega$**
flor	12	19	35161	0m18s	1	2.023138	2.023138
gato	19	21	54634	0m45s	6	20.157514	18.851648
lady	37	41	81713	2m21s	2	11.694878	10.382696
fruta	17	20	29044	0m10s	1	10.240192	10.240192
prato	11	16	50667	0m33s	3	9.691740	9.604567
ovelha	26	19	43383	0m25s	2	9.888885	6.670285

Tabela 4.1: Iterações do photometric stereo from motion

A *Tabela 4.1* nos mostra, para cada objeto, o número de pontos selecionados durante a fase de *shape from motion*, o número de quadros totais da sequência, o número de pontos da superfície retornada pelo *photometric stereo*, tempo por iteração, número de iterações e o erro medido através do RMSD ao final da primeira e da última iterações. O RMSD (*root-mean-square deviation*, em inglês) é a raíz quadrada da média da distância ao quadrado entre pontos do *SfM* e seus equivalentes na superfície do *photometric stereo*, em pixels (baseados na resolução do grid de imagem que usamos para as coordenadas 2D do objeto). Nossa condição de parada é *continuar enquanto o RMSD estiver diminuindo.*

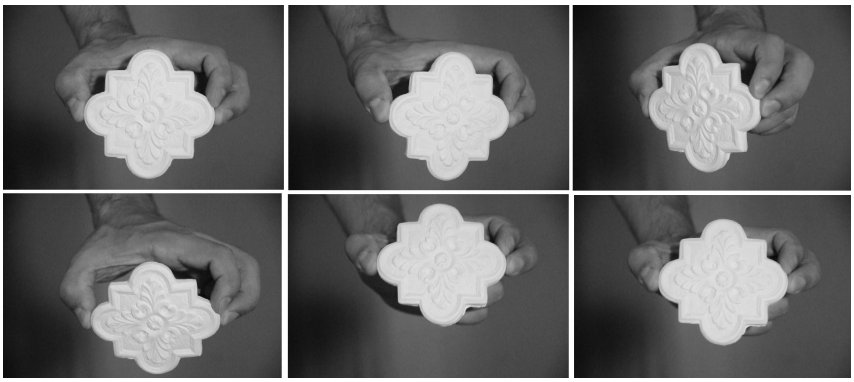

Figura 4.3: *Quadros de exemplo da sequência "flor".*

Analisando a tabela vemos que as iterações podem diminuir o RMSD, mas não o fazem, necessariamente. Para o caso da sequência *flor*, bem como para a sequência *fruta*, a primeira iteração é também a última, já que o erro cresce, ao invés de diminuir, na segunda iteração. Como ambos os objetos tem uma superfície praticamente plana, o conjunto de pontos inicial já é capaz de representar uma boa aproximação da superfície.

Na *Figura 4.4* temos uma boa ideia do resultado em cada etapa do algoritmo, como as

normais e os mapas de profundidade iniciais e finais obtidos com a execução do programa, para a sequência *flor*.

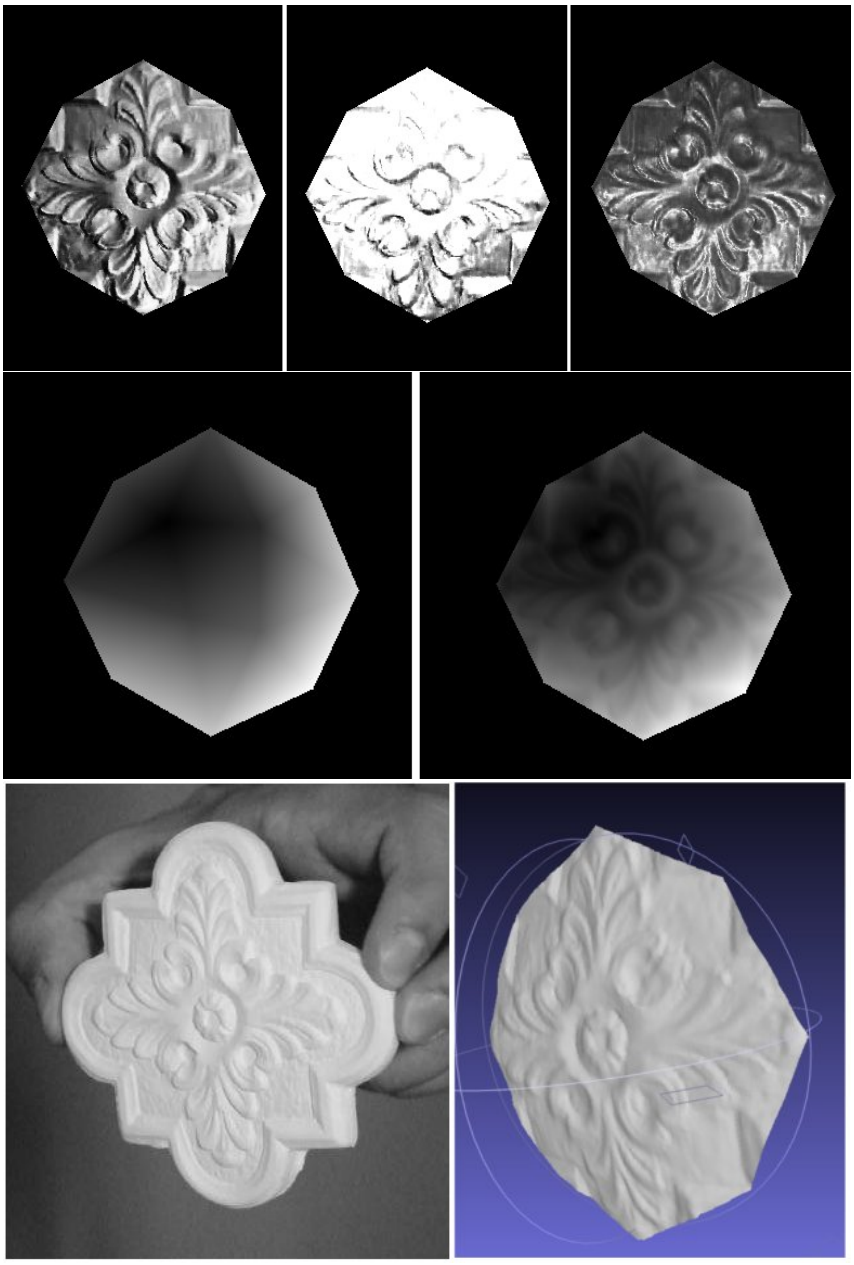

Figura 4.4: *Primeira linha: normais da superfície (em x, y e z, respectivamente) recuperadas pelo photometric stereo. Segunda linha: mapa de profundidade da superfície inicializada com os pontos do shape from motion e da superfície final através do photometric stereo. Terceira linha: parte de quadro da sequência "flor" e superfície final renderizada.*

A *Figura 4.5* nos mostra os piores resultados visuais. Esse problema ocorre principalmente pelo fato de estarmos, nesses casos, lidando com superfícies possuidoras de texturas, com diferentes cores, que confundem na remoção do albedo. Cores mais claras

representam um albedo mais alto, com maior refletividade. Na sequência *gato*, podemos perceber que as cores do focinho e dos olhos são escuras, e, no caso dos olhos especificamente, temos também o problema da refletividade não ser lambertiana, e sim altamente especular, gerando brilho. A seqência *lady* resulta em uma face que demonstra bem a profundidade do objeto, porém a superfície é extremamente ruidosa devido à textura, principalmente na região dos óculos e da barriga. Na sequência *fruta* o mesmo ocorre nos pontos claros que abundam a superfície.

As superfícies recuperadas nas sequências *flor, prato e ovelha*, por representarem objetos em gesso (com refletividade difusa) e sem textura, resultam nas melhores reconstruções. A *Figura 4.6* nos mostra o resultado da reconstrução para a sequência *prato*, que consiste de um prato circular com desenhos em relevo, representando cachos de uva dentro de uma espécie de cálice.

Na figura *Figura 4.7* podemos ver como a profundidade da superfície do objeto *ovelha* é bem representada na reconstrução utilizando o método de *photometric stereo from motion* (os pequenos pontos saltados na superfície são devido a marcas feitas na peça original para ajudar na marcação dos pontos manualmente). Pequenos detalhes foram preservados e podem ser percebidos em algumas marcações que fizemos na imagem, como no olho direito do modelo e também nas divisões pelo corpo. Nessa superfície, porém, podemos ver que há uma certa deformação no lado direito da figura, que é de onde vem a iluminação da cena. A sombra projetada em vários quadros ao redor do focinho parece confundir o método e acaba por achatar o focinho gravemente (um leve achatamento pode ser percebido também na superfície da sequência *lady*, na região do nariz, pelo mesmo motivo), gerando aí uma necessidade da correta detecção e exclusão das sombras projetadas no modelo ou a utilização de um método de estimativa mais robusto que o *SVD*, já que este, como dito no capítulo anterior, é extremamente sensível a *outliers*.

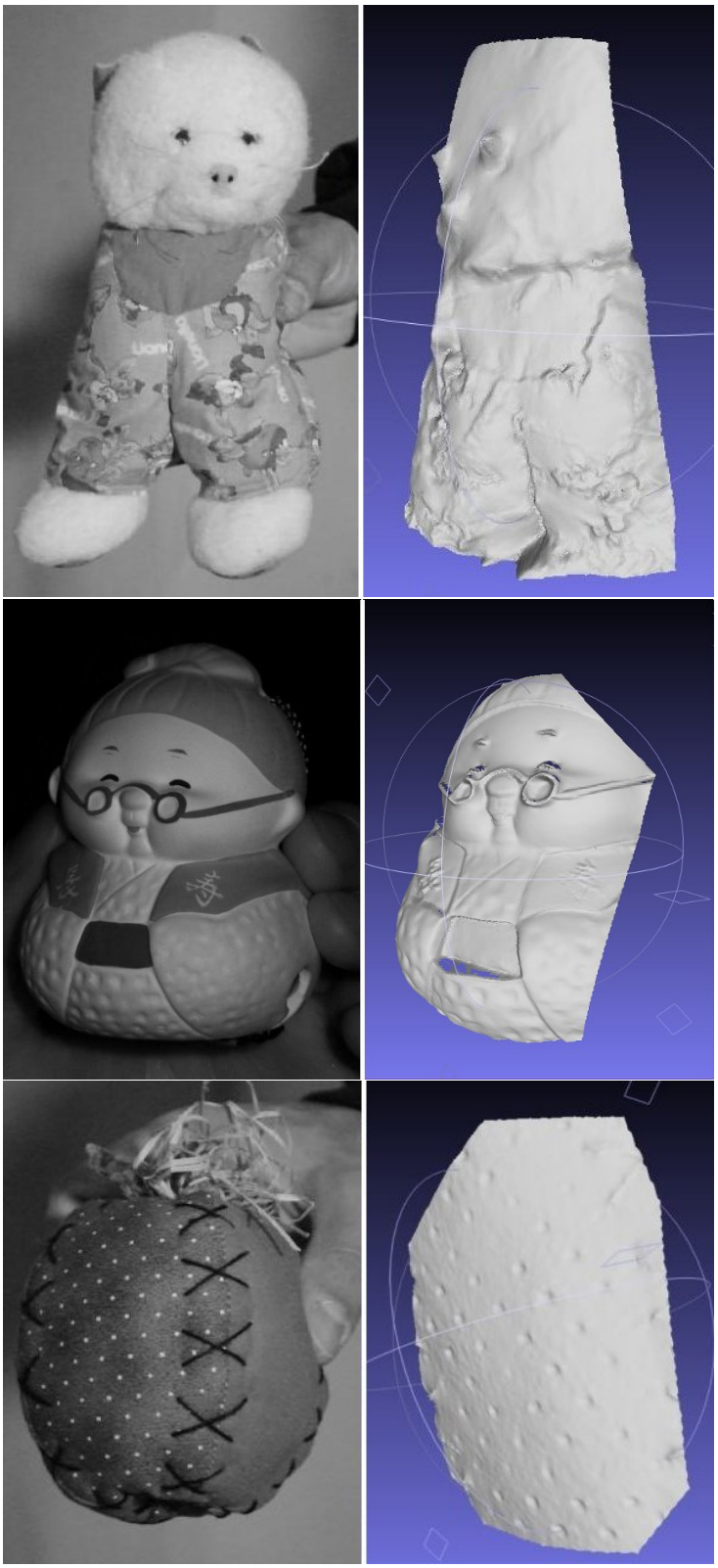

Figura 4.5: *Partes de quadro das sequências "gato", "lady" e "fruta", respectivamente, e superfície final renderizada.*

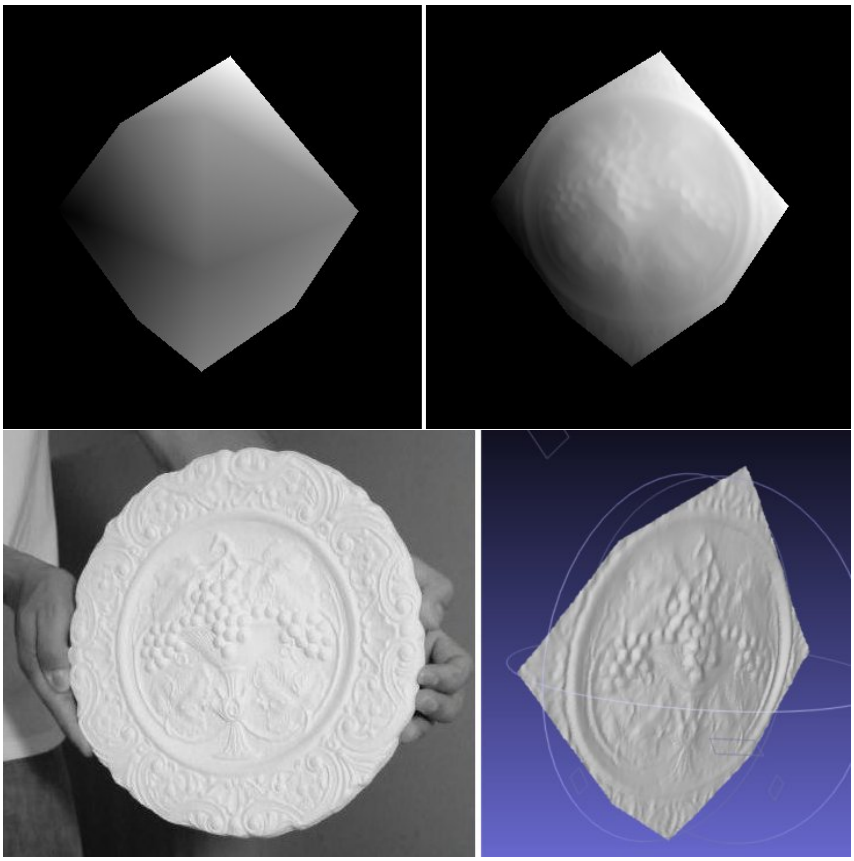

Figura 4.6: *Primeira linha: mapa de profundidade da superfície inicializada com os pontos do shape from motion e da superfície final através do photometric stereo. Segunda linha: parte de quadro da sequência "prato" e superfície final renderizada.*

Para uma comparação das superfícies as quais consideramos as melhores obtidas por nossa implementação, geramos *ground truths* para esses objetos, com o auxílio do *scanner laser 3D Vivid 910* da *Konica Minolta*[2], utilizado em projetos de preservação digital [81, 51]. Com essas superfícies em mãos, utilizamos o método de alinhamento *ICP* [87] (*Iterative Closest Points*, em inglês), um método iterativo que busca encontrar uma transformação que diminua a distância de regiões com interposição em duas superfícies, para alinharmos as duas superfícies de cada objeto e medirmos a média da distância entre elas, em milímetros. Essa média é calculada utilizando o conjunto de pontos selecionados pelo *ICP* nas duas superfícies para realizar o alinhamento (o número de pontos escolhido em nossos testes é 1000).

[2]http://www.konicaminolta.com/instruments/products/3d/non-contact/vivid910/index.html

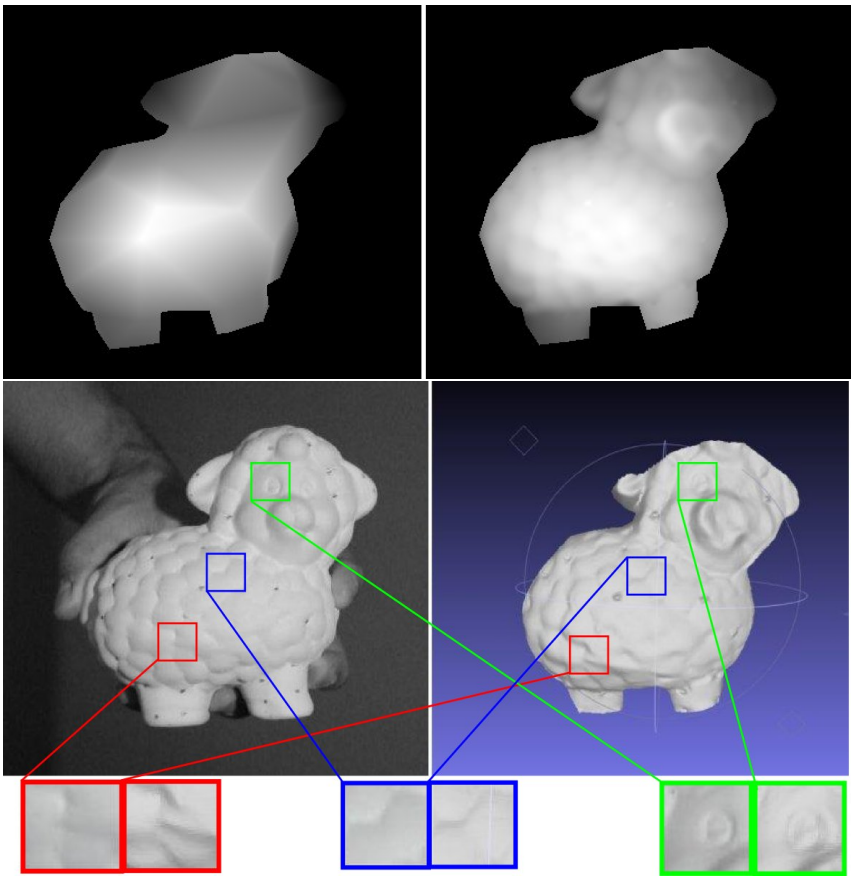

Figura 4.7: *Primeira linha: mapa de profundidade da superfície inicializada com os pontos do shape from motion e da superfície final através do photometric stereo. Segunda linha: parte de quadro da sequência "ovelha" e superfície final renderizada. As marcações coloridas dão destaque a detalhes preservados na superfície.*

Como podemos ver na *Tabela 4.2*, tanto as superfícies da sequência *flor* como da sequência *prato* têm um erro médio muito pequeno, segundo a medida do *ICP*. A superfície da sequência *ovelha*, por outro lado, acabou por gerar um erro bem grande quando alinhada com o *ground truth*. Isso se deve, como mencionado anteriormente, ao problema no lado direito da figura. O achatamento nesse lado específico da figura, que mal chega a tocar a superfície do *ground truth*, acabou elevando a média da distância, apesar do erro ser aparentemente bem reduzido ao olharmos no resto da figura (vide *Figura 4.8* para uma melhor ideia do alinhamento realizado).

Objeto	Flor	Prato	Ovelha
Dist. média	0.028930	1.490355	44.584443

Tabela 4.2: Média da distância entre as superfícies ao fim do alinhamento

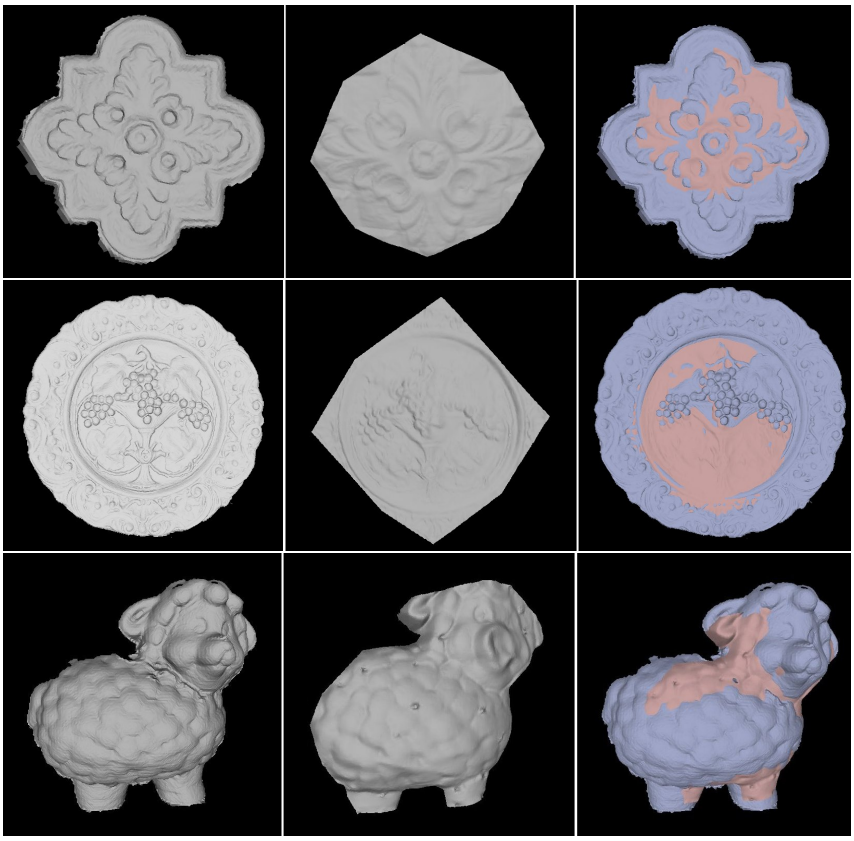

Figura 4.8: *À esquerda: ground truths. Ao centro: superfícies recuperadas pelo photometric stereo from motion. À direita: as duas superfícies alinhadas (em azul o ground truth e em rosa a superfície pelo método implementado) para cada sequência.*

Dados os resultados dos testes, percebemos que o método de *photometric stereo*, como está implementado, dependeria de métodos auxiliares para casos como os vistos na *Figura 4.5*. Tanto o albedo quanto reflexões especulares influem negativamente na qualidade da reconstrução, impossibilitando o uso desse método na preservação digital de objetos com tais superfícies altamente especulares ou texturizadas. Alguns métodos propõem a recuperação do albedo através de variações do *photometric stereo*, como [15, 36] (sendo este último um método de *photometric stereo* que usa uma sequência de imagens coloridas). Para o caso da reflexão especular podemos citar [83, 75], ambos propondo remover as especulares da superfície baseando-se em uma única imagem colorida. Resultados interessantes foram obtidos na ausência de especulares e texturas, sendo estes possivelmente

aplicáveis à preservação digital, porém, ainda nesses casos, um tratamento para o problema das sombras é necessário. Yuille e Snow propõem em [84] a remoção dos pontos onde $\mathrm{n(x)} \cdot \mathrm{l}(\mu) \leq T$, sendo T um limiar de sombra, para uma nova estimativa de b(x) e $\mathrm{l}(\mu)$, iterativamente, até que a sombra seja razoavelmente eliminada. Todos esses aspectos devem ser tratados para que possamos, assim, cobrir um número maior de grupos de superfícies com características diferenciadas e obtermos superfícies confiáveis para a preservação digital.

CAPÍTULO 5

CONCLUSÃO

O objetivo deste trabalho foi a análise de métodos passivos de aquisição para reconstrução 3D de objetos para preservação digital de acervos culturais. Várias abordagens completamente diferentes entre si foram analisadas para esse propósito, e a maioria delas descartada devido a certas restrições, como:

- necessidade de superfícies altamente texturizadas;
- grande limitação quanto ao tipo de superfície;
- limitações na captura, como exigência de mesas rotatórias, calibrações complicadas etc;
- resultados pobres com poucos números de pontos, ocasionando em superfícies extremamente simplificadas.

Visando bons resultados e captura razoavelmente descomplicada, decidimos optar por um método híbrido, que combina a corretude dos poucos pontos retornados pelo método da fatoração de Tomasi-Kanade [76], junto com a informação das rotações de câmera em cada quadro, com o resultado de superfícies completas e ricas em detalhes do método de *photometric stereo*. O método implementado e testado foi o de Lim et al. [47], também referido como *photometric stereo from motion*.

5.1 Análise do photometric stereo from motion

A união de dois métodos diferentes na tentativa de se obter uma combinação de seus pontos positivos acaba também por trazer algumas dificuldades de ambos. No caso do

método implementado, perdemos a vantagem do *photometric stereo* de não precisarmos nos preocupar com a correspondência dos pontos, mas em compensação ganhamos na facilidade de resolver as ambiguidades desse método. A vantagem sobre o *shape from motion* é a de obtermos uma superfície completa, com resolução limitada pelo tamanho das imagens com as quais decidimos trabalhar. Como foi visto no capítulo anterior, no caso ótimo conseguimos recuperar vários detalhes da superfície original, fator extremamente importante para a preservação digital.

Apesar da possibilidade de bons resultados, podemos citar alguns pontos negativos intransigíveis desse método:

1. ainda há a necessidade de pelo menos alguns pontos passíveis de rastreamento na superfície para o *shape from motion*;
2. cada iteração do método pode levar um tempo razoável para finalizar, dependendo do número de pontos dentro da região a ser reconstruída;
3. o resultado do método é apenas uma superfície de uma vista específica do objeto, havendo a necessidade de uma execução para cada vista e posterior combinação dessas em um objeto completo.

Além dos problemas citados acima, há também aqueles relacionados a características da superfície e condições da aquisição das imagens. O método de *photometric stereo* implementado é limitado a superfícies lambertianas, sem refletividade especular, e a maioria dos objetos do mundo real não cumpre com esse requisito. Também as texturas, que ajudam no rastreamente dos pontos, agem como fator dificultante na separação do albedo das normais da figura. A sombra, projetada facilmente por qualquer protuberância na superfície, é outro fator dificultante que acaba por afetar na profundidade da superfície. Felizmente, para estes problemas, há diversas técnicas sugeridas que tentam acabar, ou ao menos minimizar, com esses fatores que dificultam a reconstrução de uma superfície fidedigna.

5.2 Trabalhos futuros

As superfícies retornadas pelo *photometric stereo* se mostraram com qualidade suficiente para seu uso na preservação digital. Os problemas relacionados a sombras, refletividade especular e albedo, porém, devem ser ainda melhor investigados e resolvidos. Para ambos os estágios do método de *photometric stereo from motion*, há a possibilidade de melhora dos resultados com o uso de métodos de fatoração de matrizes incompletas, e a substituição do método dos mínimos quadrados por um método de estimativa robusto (menos suscetível a outliers, como erros de rastreamento ou sombras). Algumas outras soluções para os problemas na fase de *photometric stereo* já foram citadas ao final do capítulo anterior.

Como continuação do trabalho temos duas opções: continuar a utilização do *photometric stereo from motion*; ou passar a utilizar o *photometric stereo* sozinho com alguma alternativa para a resolução da ambiguidade, como o uso de calibração, por exemplo. Para ambos os casos, há também as opções de se começar a utilizar uma variação do método de *photometric stereo*, que modele o problema de maneira diferente e resolva os problemas encontrados neste trabalho; ou continuar usando o mesmo método, porém com algumas estratégias para corrigir os problemas de sombra, especular e albedo.

Desde que os problemas encontrados sejam todos resolvidos, o *photometric stereo from motion* é um método bastante promissor para o problema de preservação digital de acervos culturais.

BIBLIOGRAFIA

[1] Modelling and interpretation of architecture from several images. *International Journal on Computer Vision*, 60(2):111–134, 2004.

[2] Y. I. Abdel-Aziz e H. M. Karara. Direct linear transformation into object space coordinates in close-range photogrammetry. *Proceedings of Symposium Close-Range Photogrammetry*, páginas 1–18, 1971.

[3] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, e R. Szeliski. Building rome in a day. *International Conference on Computer Vision*, páginas 72–79, 2009.

[4] N. G. Alldrin, T. Zickler, e D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.

[5] J. Aloimonos. Shape from texture. *Biological Cybernetics*, 58:345–360, 1988.

[6] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, e A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *ACM-SIAM Symposium On Discrete Algorithms*, páginas 573–582, 1994.

[7] R. Bajcsy e L. Lieberman. Texture gradient as a depth cue. *Computer Graphics and Image Processing*, 5:52–67, 1976.

[8] R. Basri e David Jacobs. Photometric stereo with general, unknown lighting. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 374–381, 2001.

[9] P. N. Belhumeur, D. J. Kriegman, e A. L. Yuille. The bas-relief ambiguity. *International Journal on Computer Vision*, 35:33–44, November de 1999.

[10] F. Bernardini, H. Rushmeier, I. M. Martin, J. Mittleman, e G. Taubin. Building a digital model of Michelangelo's Florentine Pietá. *IEEE Computer Graphics and Applications*, 22:59–67, 2002.

[11] P. J. Besl e J. L. C. Sanz (Ed.). *Active optical range imaging sensors*. Springer-Verlag, New York, NY, 1988.

[12] M. J. Brooks. *Shape from shading*. MIT Press, Cambridge, MA, 1989.

[13] M. J. Brooks e B. K. P. Horn. Shape and source from shading. *Proceedings of International Joint Conference on Artificial Intelligence*, páginas 932–936, 1985.

[14] M. Chandraker, F. K., e D. Kriegman. Reflections on the generalized bas-relief ambiguity. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 788–795, June de 2005.

[15] C. Y. Chen, R. Klette, e R. Kakarala. Albedo recovery using a photometric stereo approach. *ICPR02*, páginas 700–703, 2002.

[16] C.Y. Chen, R. Klette, e C.F. Chen. 3d reconstruction using shape from photometric stereo and contours, 2003.

[17] G. Q. Chen e G. G. Medioni. Practical algorithms for stratified structure-from-motion. 20(2):103–123, 2002.

[18] T. Darrell e K. Wohn. Pyramid based depth from focus. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 504–509, 1988.

[19] P. E. Debevec, C. J. Taylor, e J. Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, páginas 11–20, New York, NY, 1996. ACM.

[20] J. D. Durou, M. Falcone, e M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43, 2008.

[21] P. Favaro. Shape from focus/defocus, 2002.

[22] M. A. Fischler e R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[23] D. A. Forsyth. Shape from texture without boundaries. *Proceedings of European Conference on Computer Vision*, páginas 225–239, 2002.

[24] G. H. Golub e C. Reinsch. *Singular value decomposition and least squares solutions. Handbook for Automatic Computation.* Springer-Verlag, New York, NY, 1971.

[25] P. F. U. Gotardo e A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2051–2065, 2011.

[26] G. Guidi, L. Micoli, M. Russo, B. Frischer, M. De Simone, A. Spinetti, e L. Carosso. 3d digitization of a large model of imperial rome. *Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling*, páginas 565–572, Washington, DC, 2005. IEEE Computer Society.

[27] G. Guidi, M. Pieraccini, S. Ciofi, V. Damato, J.-A. Beraldin, e C. Atzeni. Tridimensional digitizing of Donatello's Maddalena. *International Conference on Image Processing*, páginas 578–581, 2001.

[28] R. I. Hartley e A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, second edition, 2004.

[29] H. Hayakawa. Photometric stereo under a light-source with arbitrary motion. *Journal of the Optical Society of America A*, 11(11):3079–3089, November de 1994.

[30] C. Hernández e F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, December de 2004.

[31] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, e R. Cipolla. Non-rigid photometric stereo with colored lights. *Proceedings of International Conference on Computer Vision*, páginas 1–8, 2007.

[32] C. Hernández, G. Vogiatzis, e R. Cipolla. Multi-view photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:548–554, 2008.

[33] A. Hertzmann e S. M. Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1254–1264, 2005.

[34] T. Higo, Y. Matsushita, N. Joshi, e K. Ikeuchi. A hand-held photometric stereo camera for 3-d modeling. *International Conference on Computer Vision*, páginas 1234–1241, 2009.

[35] B. K. P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. *PhD thesis, MIT*, 1970.

[36] O. Ikeda e Y. Duan. Color photometric stereo for albedo and shape reconstruction. *WACV*, páginas 1–6. IEEE Computer Society, 2008.

[37] K. Ikeuchi e Y. Sato (Ed.). *Modeling From Reality*. Kluwer Academic Publishers, 2001.

[38] K. Ikeuchi e B. K. P. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17(1-3):141–184, 1981.

[39] K. Ikeuchi, T. Oishi, J. Takamatsu, R. Sagawa, A. Nakazawa, R. Kurazume, K. Nishino, M. Kamakura, e Y. Okamoto. The great buddha project: Digitally archiving, restoring, and analyzing cultural heritage objects. *International Journal on Computer Vision*, 75(1):189–208, 2007.

[40] N. Joshi e D. J. Kriegman. Shape from varying illumination and viewpoint. *International Conference on Computer Vision*, páginas 1–7, 2007.

[41] D. Koller, M. Turitzin, M. Levoy, M. Tarini, G. Croccia, P. Cignoni, e R. Scopigno. Protected interactive 3d graphics via remote rendering. *ACM Transactions on Graphics*, 23(3):695–703, 2004.

[42] S. Kumar, D. Snyder, D. Duncan, J. Cohen, e J. Cooper. Digital preservation of ancient cuneiform tablets using 3d-scanning. *International Conference on 3D Digital Imaging and Modeling*, 0:326, 2003.

[43] C. H. Lee e A. Rosenfeld. Improved methods of estimating shape from shading using the light source coordinate system. *Artificial Intelligence*, 26:125–143, 1985.

[44] D. T. Lee e B. J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer and Information Sciences*, 9(3):219–242, 1980.

[45] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathematics*, 2(2):164–168, 1944.

[46] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, e D. Fulk. The digital michelangelo project: 3d scanning of large statues. *Special Interest Group on GRAPHics and Interactive Techniques*, páginas 131–144, 2000.

[47] J. Lim, J. Ho, M. Yang, e D. J. Kriegman. Passive photometric stereo from motion. *International Conference on Computer Vision*, páginas 1635–1642, 2005.

[48] M. I. A. Lourakis e A. A. Argyros. Sba: A software package for generic sparse bundle adjustment. *ACM Transactions of Mathematical Software*, 36(1):1–30, 2009.

[49] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60:91–110, 2004.

[50] D. W. Marquardt. An algorithm for least-squares estimation of non-linear parameters. *Journal of the Society of Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[51] C. M. Mendes, L. Gomes, B. T. Andrade, R. S. Alburnio, W. A. Castelluber, J. O. S. Junior, e O. R. P. Bellon. Preservação digital do acervo nacional: de artes indígenas até obras coloniais de Antônio Francisco Lisboa, o Aleijadinho. *Special Session on*

Works in Progress - Conference on Graphics, Patterns and Images (XXIII SIBGRAPI), 2010.

[52] T. Morita e T. Kanade. A sequential factorization method for recovering shape and motion from image streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):858–867, 1997.

[53] S. K. Nayar. Shape from focus. Relatório Técnico CMU-RI-TR-89-27, Robotics Institute, Pittsburgh, PA, November de 1989.

[54] S. K. Nayar e Y. Nakagawa. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994.

[55] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–777, 2004.

[56] Y. Ohta, K. Maenobu, e T. Sakai. Obtaining surface orientation from texels under perspective projection. páginas 746–751, 1981.

[57] G. Pavlidis, A. Koutsoudis, F. Arnaoutoglou, V. Tsioukas, e C. Chamzas. Methods for 3d digitalization of cultural heritage. *Journal of Cultural Heritage*, 8(1):93–98, 2007.

[58] A. L. Peixoto, D. L. Canhos, L. Marinoni, e R. Vazzoler. Diretrizes e estratégias para a modernização de coleções biológicas brasileiras e a consolidação de sistemas integrados de informação sobre biodiversidade. 1, 2006.

[59] A. P. Pentland. Local shading analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:170–187, 1984.

[60] A. P. Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):523–531, 1987.

[61] A. P. Pentland. Shape information from shading: a theory about human perception. *Proceedings of International Conference on Computer Vision*, páginas 404–413, 1988.

[62] C. Poelman e T. Kanade. A paraperspective factorization method for shape and motion recovery. (CMU-CS-93-219), 1993.

[63] M. Pollefeys, L. Van Gool, M. Vergauwen, K. Cornelis, F. Verbiest, e J. Tops. Image-based 3d acquisition of archaeological heritage and applications. *Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage*, páginas 255–262, New York, NY, 2001. ACM.

[64] S. Ramalingam, S. K. Lodha, e P. F. Sturm. A generic structure from motion framework. *Computer Vision and Image Understanding*, 103(3):218–228, 2006.

[65] J. W. Roach e J. K. Aggarwal. Computer tracking of objects moving in space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, páginas 127–135, April, 1979.

[66] S. Seitz, B. Curless, J. Diebel, D. Scharstein, e R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *IEEE Conference on Computer Vision and Pattern Recognition*, 1:519–526, 2006.

[67] B. X. Shi, Y. Matsushita, Y. Wei, C. Xu, e P. Tan. Self-calibrating photometric stereo. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 1118–1125, 2010.

[68] J. Shi e C. Tomasi. Good features to track. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, páginas 593–600, 1994.

[69] L. Silva, O. R. P. Bellon, e K. L. Boyer. Precision range image registration using a robust surface interpenetration measure and enhanced genetic algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):762–776, 2005.

[70] L. Silva, O. R. P. Bellon, P. F. U. Gotardo, e K. L. Boyer. Low-overlap range image registration for archaeological applications. *Proceedings of IEEE/CVPR Workshop on Applications of Computer Vision in Archaeology*, 2003.

[71] L. Silva, O. R. P. Bellon, P. F. U. Gotardo, e K. L. Boyer. Range image registration using enhanced genetic algorithms. *Proceedings of the IEEE International Conference on Image Processing*, 2:711–714, 2003.

[72] N. Snavely, S. M. Seitz, e R. Szeliski. Modeling the world from internet photo collections. *International Journal on Computer Vision*, 80(2):189–210, 2008.

[73] I. J. A. Soares, A. Vrubel, D. R. Drees, M. C. L. Borntorin, L. Silva, e O. R. P. Bellon. Computação na preservação e difusão de patrimônios históricos e culturais. *Anais do XXVI Congresso da Sociedade Brasileira de Computação I Workshop de Computação e Aplicações*, 2006.

[74] B. J. Super e A. C. Bovik. Shape from texture using local spectral moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):333–343, 1995.

[75] R. T. Tan e K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:178–193, February de 2005.

[76] C. Tomasi e T. Kanade. Shape and motion from image streams: a factorization method - full report on the ortographic case. *Cornell TR 92-1270 and Carnegie Mellon CMU-CS-92-104*, March, 1992.

[77] B. Triggs, P. F. McLauchlan, R. I. Hartley, e A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. *Proceedings of the International Workshop on Vision Algorithms*, páginas 298–372, London, UK, 2000. Springer-Verlag.

[78] P. S. Tsai e M. Shah. Shape from shading using linear approximation. *Image and Vision Computing Journal*, 12(8):487–498, 1994.

[79] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3-d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automatoin*, 3(4):323–344, 1987.

[80] S. Ullman. *The Interpretation of Visual Motion*. The MIT Press, Cambridge, Ma, 1979.

[81] A. Vrubel. Pipeline para reconstrução digital de objetos com scanners 3d de triangulação a laser: aplicazccão na preservação de acervos naturais e culturais. *Dissertação de Mestrado do Programa de Pós-Graduação em Informática da Universidade Federal do Paraná*, 2008.

[82] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engeenering*, 19(1):139–144, January de 1980.

[83] Q. Yang, S. Wang, e N. Ahuja. Real-time specular highlight removal using bilateral filtering. *European Conference on Computer Vision*, 2010.

[84] A. L. Yuille e D. Snow. Shape and albedo from multiple images using integrability. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 158–164, 1997.

[85] L. Zhang, B. Curless, A. Hertzmann, e S. M. Seitz. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. *International Conference on Computer Vision*, 1:618, 2003.

[86] R. Zhang, P. Tsai, J. Cryer, e M. Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, Aug, 1999.

[87] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal on Computer Vision*, 13:119–152, October de 1994.

[88] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, Nov, 2000.

[89] J. Y. Zheng e Z. L. Zhang. Virtual recovery of excavated relics. *IEEE Computer Graphics and Applications*, 19:6–11, 1999.

[90] Q. Zheng e R. Chellapa. Estimation of illuminant direction, albedo, and shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):680–702, 1991.