UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LEONARDO ALVES MACHADO

Captura e Visualização de Vídeo 3D em Tempo Real

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Manuel Menezes Oliveira Neto
Orientador

Porto Alegre, janeiro de 2006

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Machado, Leonardo Alves

Captura e Visualização de Vídeo 3D em Tempo Real / Leonardo Alves Machado. – Porto Alegre: PPGC da UFRGS, 2006.

72 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2006. Orientador: Manuel Menezes Oliveira Neto.

1. Vídeos 3D. 2. Image-Based Rendering. 3. 3D Image Warping. I. Oliveira Neto, Manuel Menezes. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. José Carlos Ferraz Hennemann
Vice-Reitor: Prof. Pedro Cezar Dutra Fonseca
Pró-Reitora de Pós-Graduação: Profª. Valquíria Linck Bassani
Diretor do Instituto de Informática: Prof. Philippe Olivier Alexandre Navaux
Coordenador do PPGC: Prof. Carlos Alberto Heuser
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

“Imagination is more important than knowledge...”
— ALBERT EINSTEIN

AGRADECIMENTOS

Agradeço:
A Deus acima de tudo.
Ao meu orientador, por me ensinar praticamente tudo o que sei de computação gráfica, pela paciência comigo e por se mostrar sempre pronto a resolver qualquer dúvida ou problema.
A Ruigang Yang, que cooperou com o trabalho cedendo vídeos naturais em estéreo.
A todo o grupo de computação gráfica da UFRGS, tanto professores quanto alunos e colegas, que sempre foi muito receptivo e disposto a ajudar a resolver problemas e a desenvolver idéias.
À minha família, pelo apoio, por me sustentarem (pois só com a bolsa não me sustento sozinho...) e por se mostrarem interessados no trabalho mesmo sem entender nada das minhas explicações.
Aos meus amigos, pelo apoio e por simplesmente serem meus amigos
À Djenifer, minha namorada amada, pelo amor, carinho e dedicação
À CAPES, por ter me financiado com bolsa.
Ao CNPq, através do processo 477344/2003-8, que permitiu a aquisição de alguns equipamentos utilizados ao longo da realização desse trabalho. Esperamos, no futuro, também ser capazes poder utilizar vídeos 3D em aplicações médicas de vídeo cirurgia. ...

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Vídeos são dos principais meios de difusão de conhecimento, informação e entretenimento existentes. Todavia, apesar da boa qualidade e da boa aceitação do público, os vídeos atuais ainda restringem o espectador a um único ponto de vista. Atualmente, alguns estudos estão sendo desenvolvidos visando oferecer ao espectador maior liberdade para decidir de onde ele gostaria de assistir a cena. O tipo de vídeo a ser produzido por essas iniciativas tem sido chamado genericamente de *vídeo 3D*.

Esse trabalho propõe uma arquitetura para captura e exibição de vídeos 3D em tempo real utilizando as informações de cor e profundidade da cena, capturadas para cada pixel de cada quadro do vídeo. A informação de profundidade pode ser obtida utilizando-se câmeras 3D, algoritmos de extração de disparidade a partir de estéreo, ou com auxílio de luz estruturada. A partir da informação de profundidade é possível calcular novos pontos de vista da cena utilizando um algoritmo de *warping 3D*.

Devido a não disponibilidade de câmeras 3D durante a realização deste trabalho, a arquitetura proposta foi validada utilizando um ambiente sintético construído usando técnicas de computação gráfica. Este protótipo também foi utilizado para analisar diversos algoritmos de visão computacional que utilizam imagens estereoscópias para a extração da profundidade de cenas em tempo real. O uso de um ambiente controlado permitiu uma análise bastante criteriosa da qualidade dos mapas de profundidade produzidos por estes algoritmos, nos levando a concluir que eles ainda não são apropriados para uso de aplicações que necessitem da captura de vídeo 3D em tempo real.

Palavras-chave: Vídeos 3D, Image-Based Rendering, 3D Image Warping.

Real-Time 3D Video Capture and Visualization

ABSTRACT

Videos are one of the most important ways of communicating knowledge, and information and entertainment. However, despite of their high-quality and good acceptance, conventional videos restrict the spectator's viewpoint to the camera's viewpoint. Currently, some efforts are being conducted to give the possibility to choose the viewpoint for watching the scene. Videos with this feature are generically referred to as 3D videos.

This dissertation proposes an architecture for real-time capture and exhibition of 3D videos using scene depth and color information captured at pixel level. Depth information can be obtained using depth-from-stereo algorithms, structured light and 3D cameras. Given the scene's depth information, one can use a 3D warping algorithm to generate new viewpoints of the scene.

Since there were no 3D cameras available during the development of this dissertation, the proposed architecture was validated using a synthetic environment built using computer graphics techniques. This prototype was also used to analyze several real-time depth-from-stereo algorithms. The use of a controlled environment allowed a detailed analysis of the quality of the generated depth maps' quality, leading us to conclude that these algorithms are still not suitable for real-time 3D video applications.

Keywords: 3D video, 3D image warping, real-time depth extraction.

1 INTRODUÇÃO

No início do século passado, John Logie Baird, um dos primeiros a estudar sistemas de televisão, pensava em maneiras de construir sistemas capazes de reproduzir uma cena da forma mais natural possível (HILLS, 2002). Menos de uma década depois, começaram a surgir as televisões em preto e branco. Atualmente, os aparelhos de televisão mostram vídeos coloridos com uma boa resolução.

A televisão pode ser considerada como um dos principais meios de informação e entretenimento existentes. Estudos citam que, na Alemanha, as pessoas assistem em média duas horas de televisão por dia (WELLE, 2002), e os japoneses passam em média três horas e meia na frente da televisão, o que é a maior audiência entre os meios de comunicação de massa para esse país (HORIMURA, 2005). Todavia, apesar da boa qualidade dos vídeos transmitidos e da boa aceitação do público, as imagens recebidas atualmente pelos televisores ainda restringem o espectador a um único ponto de vista. Atualmente, alguns estudos estão sendo desenvolvidos visando oferecer ao espectador maior liberdade para decidir de onde ele gostaria de assistir a cena. O tipo de vídeo a ser produzido por essas iniciativas tem sido chamado genericamente de *vídeo 3D*.

Essa dissertação apresenta uma arquitetura que pode ser utilizada para a captura e principalmente para permitir a um espectador assistir a vídeos 3D. A idéia é, basicamente, capturar a informação de cor e profundidade em tempo real da cena para que, através de um algoritmo de *warping 3D* (MCMILLAN, 1997), novos pontos de vista da cena possam ser gerados. Tendo em vista que a extração da informação de profundidade de uma cena não é um processo trivial, serão analisadas diversas alternativas para este fim. Além disso, é apresentada uma análise de alguns algoritmos para extração de profundidade da cena em tempo real a partir de pares de imagens estéreo.

1.1 Motivação

Os vídeos convencionais de duas dimensões são atualmente uma tecnologia consagrada para aplicações profissionais e para entretenimento. Vários esquemas de codificação de vídeos bastante distintos já foram desenvolvidos para os mais variados propósitos, como por exemplo videoconferências (LEE, 1998), além de vídeos com boa compressão e boa qualidade (MITCHELL; PENNEBAKER; FOGG, 1997). Todavia, qualquer desses esquemas de codificação são capazes de capturar apenas as mudanças temporais na cena. Mudanças espaciais, como alteração no ponto de vista do telespectador, não são possíveis. Um vídeo 3D dá ao usuário a possibilidade de alterar o ponto de vista a partir do qual a cena é observada.

Efeitos especiais espaço-temporais (como rotacionar uma cena em um determinado instante fixo de tempo ou em câmera lenta) foram usados em vários vídeos (MVFX,

2002). Todavia, para realização desses efeitos, um número bastante grande de câmeras e uma quantidade considerável de edição manual foi usada em alguns filmes e séries televisivas recentes (MVFX, 2002). Com a disponibilidade de vídeo 3D, esta tarefa poderia ser bastante simplificada.

Outro ponto importante a ser salientado diz respeito à liberdade do usuário escolher o local de onde ele quer ver a cena. Com um vídeo 3D, o próprio usuário poderia escolher o ângulo de visão para assistir uma determinada cena e até mesmo produzir efeitos como "congelar e rotacionar" a mesma.

Esse trabalho propõe uma arquitetura para geração e exibição de um vídeo 3D em tempo real utilizando técnicas de rendering baseado em imagens. Um algoritmo de *warping 3D* é utilizado para a produção de novos pontos de vista da cena a partir de informações de cor e profundidade presentes no vídeo. Um exemplo de quadro de vídeo 3D sintético criado nesses moldes pode ser visto na Figura 1.1, na qual é mostrado o quadro original (acima e a esquerda) e três visualizações dessa cena a partir de novos pontos de vista.

Figura 1.1: Um quadro de um vídeo 3D. No canto superior esquerdo, o quadro visto do seu ponto de vista original. Nos demais, pontos de vista escolhidos pelo usuário.

Existem diversas alternativas para extração da informação de profundidade da cena. No Capítulo 3, serão analisadas as alternativas atuais para obter a informação de profundidade referente a cada pixel da cena. Essas alternativas estão classificadas em três grupos: cálculo de disparidade a partir de estéreo, câmeras 3D e obtenção de informação tridimensional com o auxílio de luz estruturada. Uma análise das vantagens e desvantagens de cada abordagem é apresentada na secção 3.4.

A fim de verificar a qualidade de técnicas atuais de extração de profundidade em tempo real, foi criada uma cena sintética que permite a livre exploração do usuário pelo ambiente. Como atualmente, não há meios de se extrair a profundidade de cenas naturais com total precisão, a utilização da cena sintética justifica-se, pois tem-se a possibilidade de se obter a informação de profundidade real da cena a partir dos dados armazenados no Z-buffer. O mapa de profundidades assim obtido é usado para comparação entre técnicas de visão computacional para extração de profundidade de cenas em tempo real. Comparou-se a qualidade de reconstrução de novos pontos de vista, a precisão dos mapas de disparidade e a velocidade destes algoritmos.

1.2 Organização do Texto

Esse trabalho está estruturado da seguinte forma. O Capítulo 2 descreve as idéias principais de vários trabalhos relacionados. Os trabalhos foram classificados em quatro grupos: *vídeos com profundidade*, que são técnicas que utilizam a informação de profundidade para cada pixel da cena para geração de vídeos 3D; *sistemas multi-câmeras*, que fazem uso de "vetores" de câmeras calibradas para produzir efeitos de tridimensionalidade como paralaxe e visão estéreo; *sistemas com modelagem completa dos objetos*, que propõe a utilização de modelos geométricos completos da cena para a produção dos vídeos; e *sistemas imersivos*, que buscam proporcionar ao espectador a sensação de estar presente na cena.

No Capítulo 3 são mostradas diversas técnicas para se obter a informação de profundidade de uma cena. Três grupos de técnicas foram abordados: a *extração de disparidade à partir de estéreo*, *luz estruturada* e *câmeras 3D*. As técnicas foram descritas e comparadas.

O Capítulo 4 descreve a arquitetura proposta para a captura e visualização de vídeo 3D em tempo real. Inicialmente, é apresentada uma descrição em alto nível da arquitetura, discutindo cada etapa na captura e exibição do vídeo 3D. Logo após, é explicado o algoritmo de *3D Video Warping* da cena para a geração de novos pontos de vista.

O Capítulo 5 faz uma análise dos resultados através de uma análise da qualidade de métodos de extração de profundidade em tempo real. Além de comparar a velocidade dos algoritmos e os mapas de disparidades gerados pixel a pixel com a profundidade real da cena, foi comparada a precisão das reconstruções desses algoritmos.

O Capítulo 6 apresenta as conclusões obtidas além de possíveis direções para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Apesar dos estudos sobre vídeos tridimensionais serem em sua maioria bastante recentes, existem várias publicações sobre o assunto em diversos campos de estudo, como codificações de vídeo (WüRMLIN et al., 2002), *displays* estéreos (MATUSIK; PFISTER, 2004), entre outros. Esse capítulo mostra uma breve revisão de trabalhos já publicados, começando com um resumo de trabalhos que utilizaram a informação de profundidade para produzir efeitos tridimensionais e logo após é feito um breve estudo sobre outras formas de se produzir vídeos 3D (por exemplo, utilizando várias câmeras, fazendo a modelagem completa dos objetos da cena, ou através de sistemas imersivos).

2.1 Vídeos com Profundidade

Com o surgimento de técnicas como o *3D warping* (MCMILLAN, 1997), o uso de *Image-Based Rendering* (*IBR*) chegou a ser estendido para vídeos. Christoph Fehn (FEHN, 2003) propôs uma técnica que acrescenta profundidade a vídeos convencionais. Nessa abordagem, é utilizado um vídeo no formato MPEG-2 (MITCHELL; PENNEBAKER; FOGG, 1997) para a cor - para ser compatível com os aparelhos de televisão de hoje - e um vídeo no formato *H.264/AVC* (ISO/IEC JTC 1/SC 29/WG 11 JOINT VIDEO SPECIFICATION , ITU-T REC. H.264 - ISO/IEC 14496-10 AVC) para a compressão da profundidade. As duas seqüências (*streams*) precisam ser devidamente sincronizados. O objetivo principal de Fehn (FEHN, 2003) é conseguir efeitos como paralaxe (a impressão de profundidade adquirida quando movemos a cabeça lateralmente) e a reprodução de profundidade binocular (diferença de visão que temos de um olho para o outro que dá a percepção tridimensional). A abordagem de Fehn (FEHN, 2003) diferencia-se do trabalho apresentado nessa dissertação por objetivar apenas pequenos deslocamentos como geração de paralaxe e reprodução de profundidade binocular, não se preocupando com a geração de novos pontos de vista da cena. Além disso, Fehn (FEHN, 2003) não discute como a informação de profundidade da cena será obtida

O Formato MPEG-4, dentro da *Animation Framework Extension* (*AFX*), também prevê em sua especificação (ISO/IEC JTC1/SC29/WG11 N4415: PDAM OF ISO/IEC 14496-1 / AMD4, 2001) a possibilidade de se usar *IBR* para representar uma cena. De acordo com Y. Bayakovski et. al. (BAYAKOVSKI et al., 2002), pode-se usar em vídeos MPEG-4 recursos como texturas associadas a uma profundidade (*DepthImage*), por *Layered Depth Images* (SHADE et al., 1998) (*LDI* - pixels que podem conter múltiplos valores de cor, cada uma associada a uma profundidade), ou *OctreeImage* (conjunto de voxels para representar um objeto organizados em uma octree) para a produção de vídeos 3D. A arquitetura apresentada nessa dissertação, diferentemente do trabalho de Bayakovski et al. (BAYAKOVSKI et al., 2002), independe de qualquer especificação de compressão e exi-

bição de dados, e, com os métodos de extração de profundidade descritos no Capítulo 3, pode ser usada para imagens naturais em tempo real independentemente da complexidade da cena.

Phil Harman (HARMAN, 2000) também propôs uma forma de geração de vídeos 3D que utiliza mapas de profundidade e tem como objetivo gerar vídeos para entretenimento que possam ser visualizado mesmo em casa (sem equipamentos especiais sofisticados). Sua proposta é converter os vídeos atuais de 2D para 3D. Para isso, gera-se um mapa de profundidades separando o fundo da cena dos objetos em primeiro plano com um software proprietário, para que um operador humano possa delimitar o contorno do objeto. Convertendo esse contorno em uma seqüência de pontos que dão origem à curvas *Bezier* para melhor armazenamento, o operador pode, se necessário, corrigir erros no processo de segmentação e, então, associar uma profundidade ao objeto. Feito isso, basta "seguir" o objeto selecionado nos outros quadros de vídeo para manter a informação de profundidade. Esse processo também é feito com o sob supervisão humana. A Figura 2.1 demonstra o processo de seleção, ajuste e associação da profundidade ao objeto. Diferentemente da abordagem de Harman (HARMAN, 2000), a arquitetura proposta nessa dissertação não necessita de operadores humanos na geração do vídeo, mas sim objetiva a produção e exibição do vídeo em tempo real.

Figura 2.1: À esquerda, delimitação do objeto, no centro, o ajuste ao seu contorno e à direita associação do mapa de profundidade correspondente ao objeto, Imagem extraída de (HARMAN, 2000).

Wurmlin et al. (WüRMLIN et al., 2002) propuseram recentemente uma forma de gerar vídeos 3D utilizando informação tridimensional para cada pixel e os armazenando em uma estrutura de dados orientada a pontos. O processo é feito da seguinte forma: primeiro os vídeos são capturados utilizando múltiplas câmeras sincronizadas em tempo real. Logo após, cada pixel é associado a um ponto 3D através de uma variante da técnica de *Image-Based Visual Hulls (IBVH)* (MATUSIK et al., 2000), com a diferença que em (WüRMLIN et al., 2002) a reconstrução 3D é feita para cada um dos pontos de vista das câmeras originais. Em (WüRMLIN et al., 2002), é feita a extração dos contornos dos objetos da cena, associando os pixels da silhueta com pontos de borda. Os pontos gerados então são armazenados em uma *PRk-tree* (WüRMLIN et al., 2002), criada especificamente para esse propósito semelhante a uma *quadtree*. Finalmente sofrem um processo de compressão para que possam ser transmitidos e decodificados em tempo real pelo receptor. A Figura 2.2 descreve de maneira simplificada o processo completo desde a aquisição até a visualização do vídeo. A abordagem descrita nessa dissertação, diferentemente de Wurmlin et al. (WüRMLIN et al., 2002), possui tanto a fase de captura quanto a de exibição das imagens em tempo real, não sendo necessárias etapas de armazenamento dos dados.

William Mark et al. (MARK; MCMILLAN; BISHOP, 1997) propuseram um sistema que utiliza imagens com profundidade e 3D image warping da cena para facilitar a trans-

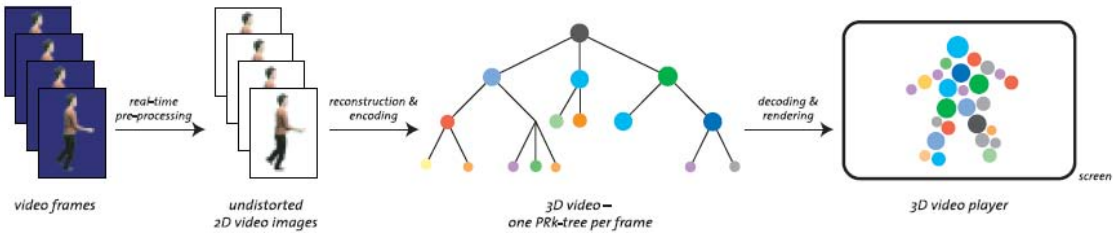

Figura 2.2: Pipeline do sistema de Wurmlin et al.: Os quadros 2D são processados para obtenção de informações 3D, as quais são armazenadas em uma *PRk-tree*. Os resultados são então transmitidos para os usuários, onde são decodificados e exibidos. Imagem extraída de (WüRMLIN et al., 2002).

missão de dados e aumentar o número de quadros por segundo de cenas sintéticas armazenadas remotamente ou localmente. William Mark et al. (MARK; MCMILLAN; BISHOP, 1997) afirmam que, em cenas complexas, a descrição geométrica dos objetos é pesada demais para uma transmissão e exibição do modelo em tempo real, então sugerem que se use o *3D warping* (MCMILLAN, 1997) para renderizar alguns quadros da cena, diminuindo assim a quantidade de dados a ser transmitidos e aumentando a taxa de quadros por segundo. A abordagem adotada nessa dissertação diferencia-se da adotada por William Mark et al. (MARK; MCMILLAN; BISHOP, 1997) pois, ao contrário de (MARK; MCMILLAN; BISHOP, 1997) nesse trabalho é sugerido que se faça o *warping 3D* de todos os quadros do vídeo, que pode ter origem natural ou sintética.

2.2 Outras Abordagens

2.2.1 Sistemas Multi-Câmeras

Alguns autores fazem uso de "vetores" de câmeras calibradas para produzir efeitos de tridimensionalidade como paralaxe e visão estéreo. Hartmut Schirmacher et al. (SCHIRMACHER; MING; SEIDEL, 2001) propuseram um método para reconstruir pontos de vista arbitrários a partir de múltiplas imagens utilizando várias câmeras. Cada câmera é tratada como um sensor que captura imagens 2D. As imagens recebidas são projetadas em um plano virtual da imagem posicionado no centro da cena para que esta seja particionada em regiões de interesse de cada sensor. Essas regiões correspondem ao conjunto de pixels que pode colaborar de algum modo para a reconstrução da imagem em todas as regiões nas quais o sensor está associado. Então os pixels são reprojetados em cada região de interesse de cada sensor levando em conta a posição do "olho" do usuário. As imagens são associadas com um teste de *Z-buffer* para visualização e então são exibidas para o usuário. Esse processo deve ser paralelizado para que resultados sejam obtidos em taxas de amostragem interativas.

Os pesquisadores da *MERL* (Mitsubishi Electronic Research Laboratories) (MATUSIK; PFISTER, 2004) propuseram um desses sistemas multi-câmeras com dezesseis câmeras de alta resolução calibradas e sincronizadas para captura de cenas. Essas câmeras foram ligadas a oito microcomputadores que comprimem os vídeos e os enviam através de uma conexão de banda larga para os receptores. Para a exibição das imagens adquiridas, um cluster recebe os dados de todas as câmeras e renderiza os quadros para o *display 3D*. Essas informações são então passadas a dezesseis projetores que geram cada um uma

visão diferente no *display*, que é formado de um material que atua como um "multiplexador" da luz dos projetores e acaba gerando a impressão de tridimensionalidade da cena. O vídeo resultante não é muito nítido - as imagens aparecem desfocadas - mas, conforme dizem os próprios pesquisadores, é o primeiro sistema a produzir uma experiência 3D sem o uso de óculos especiais com uma resolução de boa qualidade. A Figura 2.3 ilustra o funcionamento do sistema.

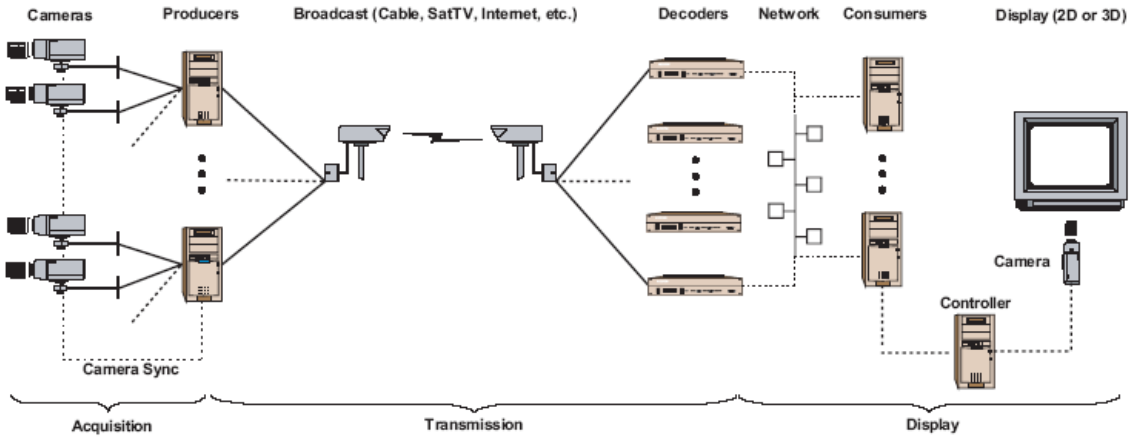

Figura 2.3: Esquema completo do sistema 3D TV proposto pelos pesquisadores da MERL. Cada 2 câmeras de captura estão ligadas a um computador que os comprimem e as repassam para os decodificadores que extraem as imagens e enviam para um cluster que as processa e as repassam para exibição em tempo real. Figura extraída de (MATUSIK; PFISTER, 2004)

2.2.2 Sistemas com Modelagem Completa dos Objetos

Uma alternativa para construção de vídeos 3D é a modelagem geométrica completa dos personagens e dos objetos da cena, os quais podem posteriormente ser visualizados utilizando técnicas de computação gráfica. A principal vantagem desse método é a possibilidade de se ver a cena de qualquer ponto de vista, sem qualquer problema de oclusão. Dentre os artigos pesquisados, dois deles optaram pelo uso do MPEG-4 (PRICE; THOMAS, 2000) (RITTERMANN; SCHULDT, 2003), que suporta a inserção da descrição dos objetos que compõem a cena.

Price et al. (PRICE; THOMAS, 2000) propuseram um método para produção de um sistema de televisão 3D que faz uso da descrição geométrica completa dos objetos utilizando MPEG-4. Esse projeto está em fase inicial e é descrito da seguinte forma:

- *Captura dos "atores"* - Inicialmente é feita a captura dos modelos geométricos dos atores (seres humanos na cena). Esse é um processo um pouco demorado, já que é necessário um conjunto de filmagens anteriores de todos os ângulos dos atores para que as estruturas das articulações dos personagens sejam capturadas. Após a obtenção dos modelos realistas dos atores, os seus movimentos e posições são capturadas, para que se possa animar os modelos.

- *Captura de faces* - Para capturar as faces dos atores, é assumido que os rostos estão claramente visíveis para uma câmera. Isso pode requerer que a posição de cada ator seja fixa ou que se use uma câmera que siga o rosto do personagem.

Outra alternativa proposta em (PRICE; THOMAS, 2000) se baseia no uso de uma modelagem simplificada. Apenas com um sistema que capture a posição do ator constantemente, pega-se a sua posição, posiciona-se então um plano transparente e mapeia-se a textura do ator sobre o objeto. Essa abordagem simplificada é demonstrada na Figura 2.4, onde o ator é representado por um retângulo com um mapeamento de textura.

Figura 2.4: Modelagem completa simplificada de Price et al. Imagem extraída de (PRICE; THOMAS, 2000)

Outro projeto semelhante foi proposto por Rittemann et al. (RITTERMANN; SCHULDT, 2003), o *IAVAS (Interactive Audiovisual Aplication Systems)*. Em sua publicação, os autores mencionam a possibilidade de se conseguir uma visualização 3D de uma cena natural utilizando as ferramentas presentes no formato MPEG-4 (ISO/IEC JTC 1/SC 29/WG 11 JOINT VIDEO SPECIFICATION , ITU-T REC. H.264 - ISO/IEC 14496-10 AVC). É sugerido o uso de ferramentas como câmeras que geram cenas omni direcionais, ou um conjunto de câmeras gravando várias visualizações da cena e dos objetos para pegar a descrição geométrica deles.

Existem projetos também que, ao invés de procurar obter toda a geometria da cena, se baseiam em aproximações das representações geométricas baseadas em imagens (*image-based*). Matusik et al. (MATUSIK et al., 2000) propuseram uma técnica chamada *Image-Based Visual Hulls (IBVH)*, que permite reconstruir, sem precisar montar uma representação geométrica ou volumétrica da cena, a silhueta dos objetos de forma rápida a ponto de permitir a observação do objeto a partir de um ponto de vista arbitrário em tempo real. Basicamente o cálculo consiste em três passos, como mostra a Figura 2.5: primeiro, um raio de "visualização" de uma câmera referência é projetado sobre a imagem visualizada por ela; a seguir, é feita a determinação dos intervalos onde o raio atravessa a silhueta 2D do objeto por meio da intersecção do raio projetado com a silhueta. Por último cada intervalo é então levado de volta ao espaço 3D com mapeamentos projetivos e intersecionados com os resultados obtidos com os raios projetivos vindos de outras câmeras. Essa abordagem, todavia, não consegue representar superfícies côncavas e a geometria da cena precisa ser previamente conhecida, por isso, não foi a abordagem adotada nessa dissertação para a produção de novos pontos de vista da cena.

2.2.3 Sistemas Imersivos

Alguns grupos de pesquisa, ao tentar criar vídeos 3D, optaram por utilizar técnicas de imersão. Essa abordagem consiste em proporcionar ao usuário a sensação de estar pre-

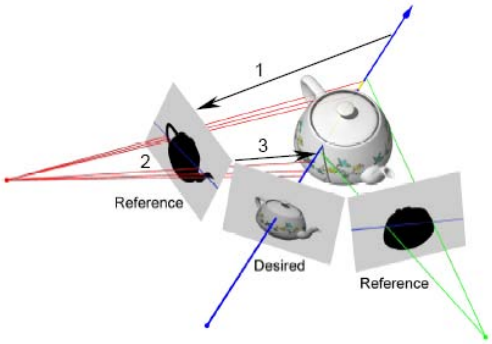

Figura 2.5: Image Based Visual Hulls - 1 raio desejado projetado nas imagens referência, 2 intervalos onde o raio projetado atravessa a silhueta do objeto são definidos, 3 mapeamentos projetivos. Extraído de (MATUSIK et al., 2000)

sente na cena, como se fosse mais um "personagem" do vídeo. Geralmente esses sistemas utilizam equipamentos especializados para proporcionar ao usuário o maior realismo possível.

Kauff et al. (FEHN; KAUFF, 2002) propuseram o *IVVV (Interactive Virtual View Vídeo)* (FEHN et al., 2001), que busca criar um sistema de televisão imersivo para distribuição em massa que procura combinar ferramentas audiovisuais do cinema, como visualização panorâmica e efeitos de som ambientais (*surround sound*), com ferramentas de realidade virtual, que proporciona bastante interatividade enquanto o usuário está imerso na cena (FEHN et al., 2001) (FEHN et al., 2001). Para isso, o vídeo é capturado por um conjunto de câmeras direcionadas para o centro do local de interesse. Os vários pontos de vista então são analisados e fundidos em uma representação compacta baseada em imagens da cena. Para realizar tudo isso, primeiro são capturadas imagens dos elementos estáticos da cena (por exemplo, o estádio em uma partida de futebol, ou o palco em uma peça de teatro) para que sejam analisados em pré-processamento e, se necessário, corrigindo os resultados manualmente. Já os objetos dinâmicos são processados em tempo real e enviados ao receptor que os compõem com as informações previamente adquiridas dos objetos estáticos. Os resultados são representações 3D constantemente atualizadas que podem ser usadas para gerar qualquer ponto de vista da cena.

Gross et.al. (GROSS et al., 2003) também possui um projeto de produção de vídeo 3D, o *blue-c*. Esse projeto propõe usar um ambiente de imersão total combinado simultaneamente com aquisição em tempo real de vídeo 3D e *rendering* de várias câmeras. Para a sincronização dos quadros, é usado um circuito eletrônico desenvolvido especialmente para combinar e gerar as imagens referentes aos olhos esquerdo e direito do usuário. Esses circuitos podem "ligar e desligar" a opacidade das paredes de projeção, para que as câmeras possam "olhar" através das paredes do ambiente e capturar as informações de posição do usuário. São usados LEDs brancos (aproximadamente dez mil) para a iluminação, algumas câmeras e alto-falantes. Fora do ambiente de imersão são posicionadas câmeras firewire e projetores LCD que produzem visões estéreo. A aquisição e a projeção é feita de maneira muito rápida, apesar de ser um processo com várias etapas. Primeiro são capturadas as visões estéreo do ambiente remoto. Em seguida, as paredes projetoras ficam transparentes e os LEDs e as câmeras firewire são ligadas para captura dos dados (posição, forma, cor, etc.) do usuário. Então a cena é reconstruída baseada nas informações adquiridas. A Figura 2.6 ilustra o ambiente do *blue-c*.

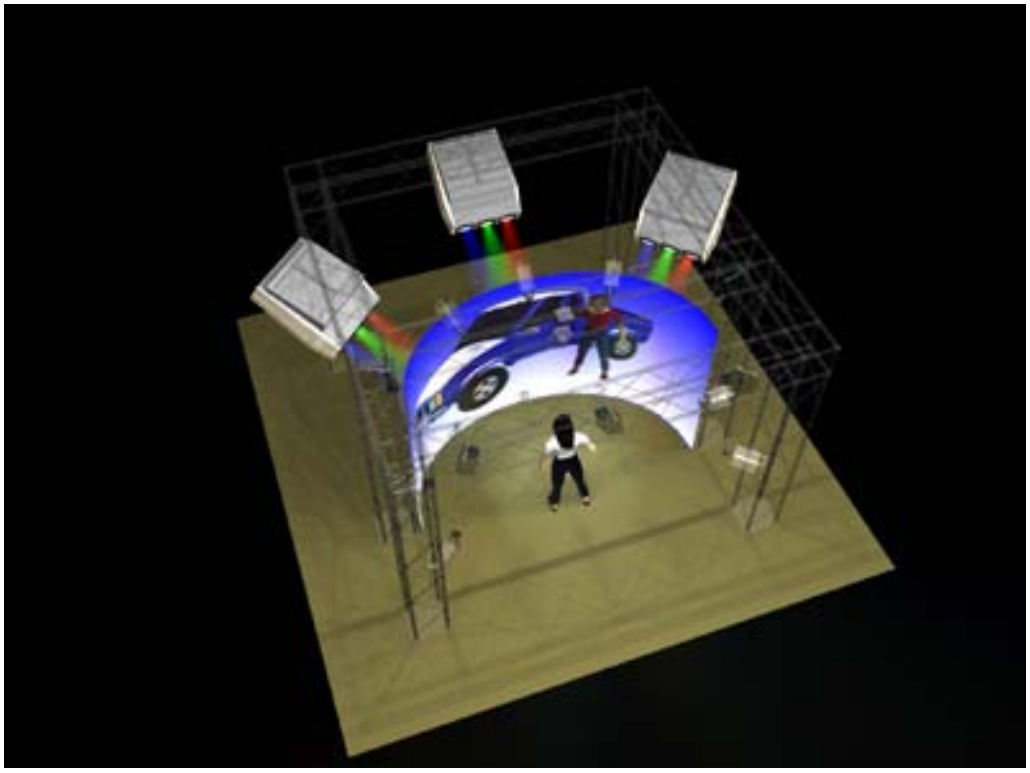

Figura 2.6: Demonstração do ambiente do sistema blue-c. Figura extraída de (GROSS et al., 2003)

2.3 Discussão

Nesse capítulo foram analisados os trabalhos relacionados a iniciativas para a criação e visualização de vídeos 3D. Esses trabalhos, podem ser classificados em quatro grupos: Vídeos com Profundidade, Sistemas Multi-Câmeras, Sistemas com Modelagem Completa dos Objetos e Sistemas Imersivos.

Vídeos com profundidade são propostas que, semelhante à técnica apresentada no Capítulo 4, associam dados de cor e profundidade a cada pixel de cada quadro do vídeo. As principais diferenças com relação ao presente trabalho são o tratamento que é dado a essa informação (que, no caso de Phil Harman (HARMAN, 2000) requer um processamento manual das informações) e os objetivos do vídeo (que, no caso de Fehn et al. (FEHN, 2003), é a geração de efeitos como paralaxe e a reprodução de profundidade binocular, para simular visão estéreo sem a necessidade do uso de duas câmeras). No contexto dessa dissertação, o interesse concentra-se na geração de novos pontos de vista da cena.

As outras abordagens diferenciam-se consideravelmente do contexto apresentado nesse trabalho, todavia, servem como alternativas válidas para a produção de vídeos 3D. Os Sistemas Multi-Câmeras fazem uso de "vetores" de câmeras calibradas para produzir, através de combinações e reamostragens das imagens capturadas, efeitos de tridimensionalidade como paralaxe e visão estéreo. Os Sistemas Imersivos buscam oferecer ao usuário a sensação de estar presente na cena, como se fosse mais um personagem do vídeo. Geralmente esses sistemas utilizam equipamentos especializados para proporcionar ao usuário o maior realismo possível.

A forma mais tradicional de se construir vídeos 3D consiste em fazer a modelagem geométrica completa dos personagens e dos objetos da cena. A principal vantagem desse

método é a possibilidade de se ver a cena de qualquer ponto de vista, sem qualquer problema de oclusão. Todavia, esse método é dependente da complexidade da cena, o que pode acarretar em uma perda de performance caso a cena possua muitos vértices. Outra restrição dessa abordagem é a dificuldade de se criar e manter uma malha poligonal da cena em tempo real, principalmente quando a cena possui objetos e superfícies deformáveis que podem forçar uma recriação de toda a malha geométrica com um movimento não previsto inicialmente na configuração da malha. Como um dos objetivos do trabalho é buscar que a geração e exibição de vídeo 3D seja em tempo real, a possibilidade da criação e manutenção de uma malha para descrever toda cena torna-se inviável. Isso justifica-se pois em grande parte dos vídeos, como eventos esportivos e transmissões ao vivo, a necessidade de informação é imediata, e o espectador não estaria disposto a esperar uma etapa de pré-processamento para a geração do vídeo.

3 OBTENDO A PROFUNDIDADE DE UMA CENA

Para criar um protótipo de sistema para geração de vídeos 3D de acordo com a arquitetura proposta no Capítulo 4, é necessário que haja meios de extrair a profundidade da cena. Para isso, é possível fazer uso de hardware especializado e técnicas de visão computacional. Alguns exemplos dessas técnicas são a utilização de mapas de disparidade extraído a partir de pares de imagens estéreo e o uso de luz estruturada. Este capítulo apresenta essas técnicas, começando com uma análise de técnicas para extração de profundidade de uma cena a partir de pares estéreo. Em seguida há uma descrição sobre câmeras 3D e das técnicas que se baseiam no uso de luz estruturada. No final do capítulo, é apresentada uma comparação entre os diversos métodos, evidenciando suas vantagens e desvantagens.

3.1 Extração de Profundidade Através de Imagens Estereoscópicas

Uma maneira para obtenção da profundidade de uma cena é a utilização de mapas de disparidade calculados a partir de pares de imagens estéreo. O algoritmo 3.1 ilustra o princípio utilizado para calcular o mapa de disparidade entre as duas imagens: esquerda (*ImE*) e direita (*ImD*). Para o cálculo de disparidade, *ImE* e *ImD* devem estar retificadas, isto é, as imagens devem estar alinhadas horizontalmente uma com a outra linha a linha. Para calcular a disparidade de um determinado pixel *p* da imagem ImE, basta procurar o pixel correspondente em ImD, que, se presente, vai estar na mesma linha de *p*. Encontrado o pixel correspondente, disparidade é representada pelo módulo da diferença entre as posições (*i.e.* colunas) dos pixels nas duas imagens.

```
para cada linha ImE[l] da imagem ImE
   para cada pixel ImE[l][p] da linha ImE[l]
      para i = 0 até fim_da_linha ImD[l]
       se (ImE[l][p] == ImD[l][i])
          Disparidade[ImE[l][p]] = |i-p|;
```

Algoritmo 3.1: Algoritmo para o cálculo de disparidade a partir de um par de imagens (ImE e ImD) estéreo retificadas.

Com o mapa de disparidade, é possível obter a profundidade Z_p para cada pixel da cena. Para isso, é usada a Equação 3.1 (TRUCCO; VERRI, 1998), onde Z_p é a profundidade referente ao pixel, f é a distância focal das câmeras, b é a distância entre os centros de projeção das duas câmeras (conhecida como *baseline*) e d é a disparidade.

Todavia, na prática, existem vários fatores que podem introduzir erros (como oclusão,

$$Z_p = \frac{fb}{d} \tag{3.1}$$

ou ambigüidade entre pixels, causadas, por exemplo, pela ausência de textura) e precisam ser tratados para evitar uma quantidade muito grande de ruído no mapa de disparidades gerado. Para resolver o problema, pode-se usar programação dinâmica, usar janela variável de pixels para comparação, ou utilizar abordagens *"Graph Cuts"* (HONG; CHEN, 2004). As próximas seções discutem cada uma das abordagens.

3.1.1 Usando Janela Variável

Algoritmos que usam janela variável para o cálculo do mapa de disparidade da cena procuram fazer a associação de pixels entre as duas imagens de maneira rápida e com pouco ruído. Para isso, na hora de fazer a comparação entre as imagens, não só o valor do pixel buscado é levado em consideração, mas também uma "janela" de vizinhos é considerada no cálculo. A premissa por trás dessa abordagem é que a disparidade entre pixels da mesma janela é aproximadamente igual. Isso reduz a ambigüidade na associação dos pixels entre as imagens.

Olga Veksler (VEKSLER, 2003) propôs um modo para cálculo de mapas de disparidade usando janela variável com bastante precisão. Inicialmente, é definida uma função que "mapeia" pixels para números reais. A partir desse resultado, calcula-se o tamanho ótimo da janela para cada um dos pixels através do método de imagens integrais (*Integral Image*, criada por Crow (CROW, 1984)) e então é feita a associação entre as imagens. Neste passo há um diferencial com relação às demais técnicas de janela de busca na hora de comparar os pixels (no caso, janelas de pixels). É feita uma interpolação linear entre as intensidades dos pixels da imagem da direita e então é medido onde o pixel da esquerda melhor se encaixa. Apesar dos bons resultados, esse algoritmo é inapropriado para vídeos por causa de seu elevado tempo de processamento.

Ruigang Yang et al. (YANG; POLLEFEYS, 2003) (YANG; WELCH; BISHOP, 2003) (YANG; POLLEFEYS; LI, 2004) também propuseram métodos de extração da disparidade de cenas usando janelas variáveis. Seu diferencial é o uso do *hardware* gráfico atual, que possibilita o cálculo em tempo real. Um dos métodos propostos faz uso de registradores especiais (*Register Combiners*) (YANG; POLLEFEYS, 2003)(YANG; WELCH; BISHOP, 2003) presentes nas placas da *NVidia*. Já o outro (YANG; POLLEFEYS; LI, 2004) utiliza o *fragment shader* das GPUs atuais.

3.1.1.1 Usando Register Combiners

Ruigang Yang et al. (YANG; POLLEFEYS, 2003) (YANG; WELCH; BISHOP, 2003) propuseram um método de extração da profundidade das cenas em tempo real utilizando o *hardware* gráfico que utiliza *register combiners*. A comparação entre os pixels das duas imagens é feita utilizando *mipmaping* (WILLIAMS, 1983) (*i.e.* uma "pirâmide" de imagens filtradas, sendo que a cada novo nível do mipmap, o valor de um pixel é dado pela média de quatro pixels do nível anterior). Isto permite diminuir a ambigüidade, visto que o uso de *mipmaping* permite a emulação de várias janelas de tamanho variável para comparação. Para cada nível de *mipmaping*, cada pixel de uma das imagens é associado ao seu correspondente na outra imagem. As comparações são feitas usando a soma de diferenças dos quadrados (*SSD*), ou seja, é feito uma soma de todos os quadrados das

diferenças obtidas em todos os níveis de mipmap calculados para cada pixel. O menor resultado de *SSD* encontrado é considerado o par do pixel para o qual se buscava a disparidade. O algoritmo 3.2 descreve simplificadamente a abordagem de Yang et al. (YANG; POLLEFEYS, 2003) (YANG; WELCH; BISHOP, 2003).

```
niveis_de_mipmap = le_usuario();
para cada linha ImE[l] da imagem ImE
   para cada pixel ImE[l][p] da linha ImE[l]
      para i = p até fim_da_linha ImD[l]
        SSD = 0;
        menor_diferença = maior_valor_possivel;
        para j = 0 até j == niveis_de_mipmap
           SSD = SSD + (ImE[l][p][j] - ImD[l][i][j])^2;
        fim para
       se (menor_diferença > SSD)
          Disparidade[ImE[l][p]] = |i-p|;
```

Algoritmo 3.2: Abordagem de Yang et al. (YANG; POLLEFEYS, 2003) (YANG; WELCH; BISHOP, 2003) para extração de profundidade usando *Register Combiners* em tempo real.

3.1.1.2 Usando Shaders

Com o objetivo de melhorar a qualidade dos resultados e aproveitando o rápido desenvolvimento do *hardware* gráfico, Ruigang Yang et al. (YANG; POLLEFEYS; LI, 2004) propuseram melhorias no algoritmo publicado anteriormente (YANG; POLLEFEYS, 2003) e passaram a usar os *fragment shaders* das placas programáveis mais modernas.

Ruigang Yang et al. (YANG; POLLEFEYS; LI, 2004) descrevem um processo de "janela adaptativa" para substituir o uso de *mipmaping*, pois este limita o tamanho da janela de comparação. Inicialmente, é selecionada uma janela de 4×4 pixels para fazer a comparação entre as imagens. Então, o procedimento é repetido para as quatro janelas 4×4 mais próximas (imediatamente acima, abaixo, à esquerda e à direita) e as duas janelas que derem a menor diferença serão consideradas no cálculo final. A comparação entre os pixels é feita usando a soma da diferença dos valores absolutos das intensidades de cada pixel (*SAD - sum of the absolute differences*), dada pela equação $SAD = \sum |p_e - p_d|$, onde p_e é o pixel referência na imagem à esquerda e p_d é o pixel procurado na imagem à direita.

Outro passo opcional proposto nesse artigo é o *Cross-Checking*, ou seja, trocar a imagem referência para gerar o mapa de disparidades. Como os mapas de disparidade podem não ser idênticos (por causa de problemas de oclusão e de amostragem), é proposto a remoção dos valores inconsistentes, substituindo-os pelos valores que dentre seus vizinhos mais se repetem, para preservar a suavidade do mapa. O algoritmo 3.3 descreve em linhas gerais a abordagem de Yang et al. (YANG; POLLEFEYS; LI, 2004).

3.1.2 Usando Programação Dinâmica

Outra proposta para extrair mapas de disparidade de uma cena a partir de estéreo é a utilização de programação dinâmica (GONG; YANG, 2005). Nela, a imagem é dividida

```
para cada linha ImE[l] da imagem ImE
   para cada pixel ImE[l][p] da linha ImE[l]
      janela4x4[ImE[l][p]].cria(centro);
      janela4x4[ImE[l][p]].cria(acima);
      janela4x4[ImE[l][p]].cria(abaixo);
      janela4x4[ImE[l][p]].cria(esquerda);
      janela4x4[ImE[l][p]].cria(direita);
      menor_diferença = maior_valor_possivel;
      para i = p até fim_da_linha ImD[l]
        janela4x4[ImD[l][i]].cria(tmp);
        computa_SAD(centro, tmp);
        Seleciona_2_menores_SAD_vizinhos_e_soma(tmp);
       se (menor_diferença > SAD)
          Disparidade[ImE[l][p]] = |i-p|;
```

Algoritmo 3.3: Abordagem de Yang et al. (YANG; POLLEFEYS; LI, 2004) para extração de profundidade utilizando *shaders* em tempo real.

em diversas linhas e o problema é resolvido para cada uma delas.

Basicamente, programação dinâmica (GONG; YANG, 2005) é uma forma de simplificar a abordagem adotada pelos algoritmos de otimização global (por exemplo, *Graph Cuts* (HONG; CHEN, 2004), veja subseção 3.1.3). Estes analisam a precisão da extração do mapa de disparidade de acordo com a função de energia objetiva descrita na Equação 3.2 (HONG; CHEN, 2004). A energia do mapa (*E(d)*), que deve ser minimizada, é dada pela soma entre a energia dos pixels que o mapa *d* coloca em correspondência ($E_{data}(d)$) e a suavidade do mapa ($E_{smooth}(d)$) (*i.e.* pequena variação de disparidade entre pixels vizinhos).

$$E(d) = E_{data}(d) + E_{smooth}(d) \tag{3.2}$$

Os algoritmos de programação dinâmica também utilizam a Equação 3.2, no entanto, para simplificar a abordagem e obter resultados em tempo menor, eles dividem a imagem em *scanlines* (que, na maioria dos casos, são as linhas da imagem) e procuram otimizar o mapa para cada uma delas. Essa estratégia é passível de erros como falta de coerência nas direções verticais. Contudo, as últimas publicações no assunto propõem meios de contornar esse problema (VEKSLER, 2005) (GONG; YANG, 2005) (GONG; YANG, 2003).

Olga Veksler (VEKSLER, 2005) propôs um modo de calcular um mapa de disparidade usando programação dinâmica. Para resolver o problema da coerência vertical, ela, em vez de utilizar as linhas da imagem (*scanlines)* para resolver o problema, propôs a criação de uma árvore na qual, cada vértice (pixel da imagem) é ligado aos vértices próximos (que pertencem a uma área de intensidade homogênea) e estes são considerados na hora de calcular a energia de suavização. Para simplificar a árvore, os vértices com diferença de intensidade entre o pixel e seus vizinhos imediatos muito altas são descartados, já que contribuem pouco para o aumento do valor da energia do mapa de disparidade e, por isso, não precisam ser "minimizados".

Minglun Gong et al. (GONG; YANG, 2003) também possuem um método bastante eficiente para se obter mapas de disparidade a partir de estéreo utilizando programação dinâmica. Para resolver os problemas clássicos da programação dinâmica descritos no início da secção, é feito um processo de *programação dinâmica baseada na confiabilidade* (*reliability-based dynamic programming*, RDP (GONG; YANG, 2003)). Por esse processo, a disparidade associada para cada pixel só é atribuída se a diferença de energia (ver Equação 3.2) entre o menor valor e o segundo menor valor obtido for maior que um dado limiar (*threshold*). Para fazer a associação entre os pixels, é usada uma janela 3×3. Para produzir bons resultados, o algoritmo é repetido várias vezes, e, a cada 3 a 5 iterações, o peso da energia de descontinuidade é aumentado para evitar mapas super-suavizados.

Em outro artigo, Minglun Gong et al. (GONG; YANG, 2005) propuseram um método que se vale da programação dinâmica com o auxílio do *hardware* gráfico para calcular um mapa de disparidades de forma bastante rápida. Para calcular o mapa de disparidades, também é feito um processo de *programação dinâmica baseada na confiabilidade* (*reliability-based dynamic programming*, RDP (GONG; YANG, 2003)). Esse processo atribui um valor de confiança para a disparidade encontrada para cada pixel. Após uma busca na horizontal, que calcula a disparidade tomando como referência ambas as imagens, é estabelecido um limiar λ e é feita a análise da suavidade do mapa considerando a direção vertical. Esse processo é repetido contínuas vezes (geralmente 3) aumentando gradualmente o valor de λ para produzir mais pixels "confiáveis". O processo pode ser todo feito no *hardware* gráfico, todavia, experimentalmente os autores concluíram que fazendo um processo misto (*GPU/CPU*) os resultados eram obtidos mais rapidamente devido ao número excessivo de passos de renderização necessários para fazer todo o processo na GPU. O algoritmo 3.4 descreve em linhas gerais a técnica.

```
Calcula o custo da associação entre as imagens
Se (usa CPU)
   copia os custos obtidos da placa para a CPU
   enquanto (muitos buracos)
      procura pares confiáveis na direção horizontal
      procura pares confiáveis na direção vertical
      aumenta o limiar
senão
   enquanto (muitos buracos)
      procura pares confiáveis na direção horizontal
      procura pares confiáveis na direção vertical
      aumenta o limiar
```

Algoritmo 3.4: Algoritmo de Gong et al. em linhas gerais (GONG; YANG, 2005)

3.1.3 Usando *Graph Cuts*

O uso de *Graph Cuts* para gerar mapas de disparidade a partir de estéreo tem obtido os melhores resultados de acordo com a classificação da Middlebury (SCHARSTEIN; SZELISKI, 2002a). Essa estratégia procura minimizar globalmente a energia do mapa de disparidade (veja Equação 3.2) construindo um grafo no qual cada vértice representa um pixel da imagem e cada aresta possui o valor da energia associada aos pixels ligados a ela. São criadas tabelas com os possíveis valores de disparidade associada a cada vértice e, então é feito um processo de escolha de melhor valor de disparidade para todos os pixels

da imagem. Esse processo, apesar de produzir excelentes resultados quando comparado com as outras técnicas de geração de mapas de disparidade a partir de estéreo, tende a ser bastante lento e, por isso, atualmente não são apropriados para criação de vídeos 3D em tempo real.

Li Hong et al. (HONG; CHEN, 2004) desenvolveram uma forma de obter um mapa de disparidades a partir de estéreo usando *Graph Cuts*. Com um processo de segmentação de cores, a imagem é subdividida em regiões homogêneas e a energia $E_{smooth}(d)$ (veja Equação 3.2) é avaliada apenas para os pixels dessas regiões, simplificando assim o processo e preservando bordas da imagem.

3.1.4 Estéreo Levando em Consideração Variações Temporais

Uma abordagem alternativa, que é própria para vídeos e aumenta consideravelmente a qualidade dos mapas de disparidade obtidos, é levar em consideração as vizinhanças espaciais e temporais de um pixel. Li Zhang et al. (ZHANG et al., 2004) propuseram um método de reconstrução de faces a partir de estéreo, combinando com luz estruturada (ver subseção 3.3), que leva em consideração as variações temporais da cena.

A idéia básica é assumir que em uma janela 3D de pixels *x, y, t* (onde *x* e *y* são as coordenadas do pixel e *t* representa de tempo) a variação de disparidade é mínima. Li Zhang et al. (ZHANG et al., 2004) ainda sugerem o uso de luz estruturada para melhorar a precisão dos resultados para que seja possível fazer uma reconstrução poligonal de faces capturadas pelo seu sistema. Todavia, é possível usar os primeiros passos do sistema proposto por eles (ou seja, o cálculo do mapa de disparidades) para auxiliar na construção de um vídeo 3D usando técnicas de rendering baseados em imagens. No entanto, este cálculo de disparidade que leva em consideração variações no tempo não pode ser implementado em tempo real, já que vários mapas de disparidade da cena são influenciados pelo resultado do mapa de cada quadro, inclusive de quadros "futuros" do vídeo.

A Figura 3.1 faz uma comparação entre a reconstrução de uma face com estéreo utilizando técnicas convencionais e a técnica de Li Zhang et al. (ZHANG et al., 2004). As duas técnicas utilizam luz estruturada para acentuar as diferenças entre pixels, auxiliando no cálculo do mapa de disparidades.

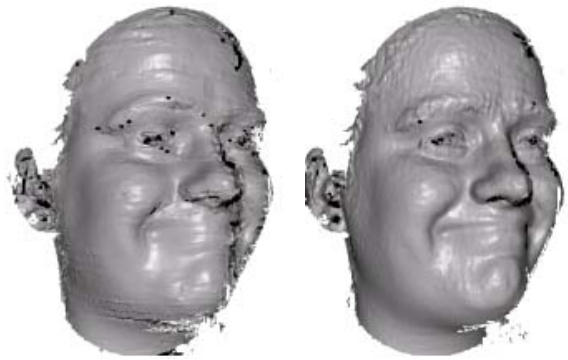

Figura 3.1: Comparação entre a reconstrução 3D de uma face com técnicas estéreo convencionais (esquerda) e a técnica de Li Zhang et al. (ZHANG et al., 2004) (direita). É possivel notar nitidamente pelos olhos, nariz, boca e pelo lado direito do rosto que a técnica de Li Zhang et al. (ZHANG et al., 2004) produz resultados menos ruidosos. Figura extraída de (ZHANG et al., 2004).

3.2 Câmeras 3D

Outra maneira de se extrair a profundidade de cenas naturais é utilizando *hardware* especial desenvolvido com essa finalidade. Keigo Iizuka e Masahiro Kawakita (IIZUKA; KAWAKITA, 2002) descrevem a *Axi-Vision camera*. Esse equipamento usa uma fonte de luz infravermelho para "iluminar" o ambiente e capturar a informação de profundidade através da intensidade dessa luz recebida por cada objeto da cena. A *Axi-Vision camera* consegue uma resolução de 920.000 pixels a uma taxa de 29,97Hz e 1,7cm precisão com um objeto posicionado a 2 metros da câmera (IIZUKA; KAWAKITA, 2002). A Figura 3.2 mostra um exemplo de extração de informação de cor e profundidade obtidas com a *Axi-Vision camera*.

Figura 3.2: Demonstração das informações de cor e profundidade extraídas com a Axi-Vision camera; Figura extraída de (IIZUKA; KAWAKITA, 2002).

A precisão do equipamento decresce com o aumento da distância, conforme mostra o gráfico da Figura3.3. A câmera possui limitações com relação ao número de vetores LEDs para captura da informação de profundidade que Iizuka et al. (KAWAKITA et al., 2002) estão tentando resolver. Espera-se uma melhora sensível na precisão dos equipamentos quando um maior número de LEDs para a captura de informação 3D for usado.

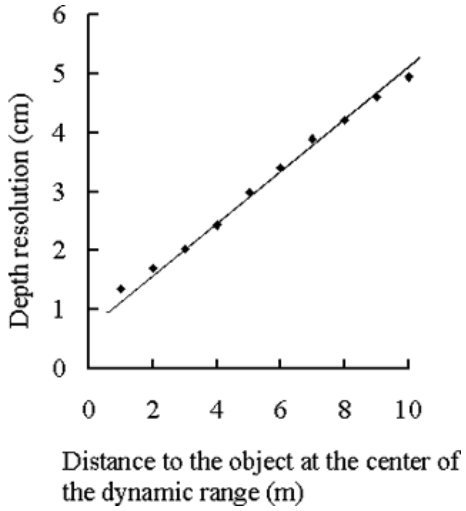

Figura 3.3: Axi-Vision camera - gráfico de precisão do equipamento. No eixo X, a distância do objeto para a câmera, no eixo Y, a precisão do mapa de profundidades. Figura extraída de (KAWAKITA et al., 2002).

G.J. Iddan e G. Yahav (IDDAN; G., 2001) também desenvolveram uma câmera capaz

de capturar além das informações RGB a profundidade da cena (Figura3.5). A *Zcam* (como foi chamada por eles) tem como princípio a geração de uma "parede luminosa" que se move ao longo do campo de visão e, quando atinge um objeto, essa luz é refletida e a informação usada para calcular o mapa de profundidade. A "parede de luz" usada é um pulso laser de curta duração com a forma de um quadrado. A *ZCam* consegue até 0.5cm de precisão, com alcance de 1 a 10m. A Figura 3.4 mostra um exemplo de extração de informação de cor e profundidade com a *Zcam*. A principal restrição dessa abordagem é a limitação referente a reflexão da luz, que, por exemplo, não ocorre em superfícies pretas.

Figura 3.4: Demonstração das informações de cor e profundidade extraídas com a ZCam; extraído de (IDDAN; G., 2001).

Figura 3.5: Demonstração da ZCam. Figura extraída de (IDDAN; G., 2001).

Por serem aparelhos portáveis e por obterem qualidade satisfatória em seus mapas de profundidade, o uso de Câmeras 3D em um sistema baseado na arquitetura descrita no Capítulo 4 pode ser uma boa alternativa. Deve-se, todavia, atentar às restrições desses equipamentos, pois nenhum dos dois exemplares possui longo alcance para obtenção de profundidade.

3.3 Obtenção de Informação Tridimensional com o Auxílio de Luz Estruturada

Uma forma de se obter informações tridimensionais de uma cena de forma rápida e eficiente é a utilização de *luz estruturada*. Nesta abordagem, as informações tridimensionais da cena podem ser obtidas da seguinte forma: um padrão de iluminação é projetado sobre

os objetos de uma cena e a luz refletida é então capturada por uma câmera. A distância relativa entre um ponto no padrão de iluminação e a posição em que se encontra na imagem capturada pela câmera é inversamente proporcional à profundidade do ponto (BOYER; KAK, 1987). Logo, conhecendo-se os parâmetros da câmera e do projetor, é possível posicionar o ponto no espaço. A Figura 3.6 ilustra exemplos de aplicação de luz estruturada em uma cena.

Figura 3.6: Demonstração do uso de luz estruturada para obter as informações tridimensionais de uma cena, extraído de (ZHANG; CURLESS; SEITZ, 2002).

O aspecto crítico nesse tipo de técnica é a obtenção da correspondência entre os pontos no padrão de luz projetado e os pixels da imagem capturada pela câmera. Se a imagem obtida pela câmera está *retificada* (ou seja, sem as distorções introduzidas pela lente da câmera), a correspondência pode ser obtida apenas com uma busca em 1D (ZHANG; CURLESS; SEITZ, 2002).

Existem problemas nesse tipo de técnica de obtenção de informações tridimensionais de uma cena. Por exemplo, a cor do pixel na imagem capturada pela câmera não depende apenas da cor do raio de luz que origina do padrão de luz projetado, mas também das propriedades de reflectância da superfície, do ângulo de visualização da câmera, entre outros fatores. Outro desafio é o tratamento de oclusões, já que raramente o padrão projetado será integralmente visível pela câmera. Ainda devemos levar em consideração a informação de cor da cena, que, na maioria das publicações atuais, é desconsiderada.

Normalmente, utiliza-se luz estruturada para extrair um modelo poligonal da cena, usando como vértices os pontos no espaço 3D encontrados com os algoritmos de busca. Nada impede, porém, que se utilize esse tipo de técnica para criar um mapa de disparidades a ser utilizado em uma aplicação de *Image-based Rendering*.

O uso de luz estruturada é bastante comum em aplicações de visão computacional. Li Zhang et al. (ZHANG; CURLESS; SEITZ, 2002) propuseram um meio de se obter rapidamente a geometria de objetos utilizando luz estruturada. É utilizado um padrão de luz com faixas coloridas bem definidas para que a transição entre as cores seja bastante simples. Então, as transições entre as cores são usadas para obter a correspondência entre o pixel da imagem capturada pela câmera e o padrão de luz projetado sobre a cena. Para minimizar os problemas decorrentes do uso de luz estruturada, é usada programação dinâmica para fazer uma otimização global, associando a cada faixa projetada uma probabilidade de associação entre padrão e pixel. Isso auxilia no tratamento de oclusões, sombras e descontinuidades. A Figura 3.7 mostra resultados dessa abordagem.

Song Zhang et al. (ZHANG; HUANG, 2004) também descrevem um método para captura da geometria de objetos em tempo real utilizando luz estruturada. Inicialmente, um PC gera um padrão RGB que varia rapidamente de forma senoidal para que haja uma variação previsível e não ambígua de intensidades luminosas a serem projetadas. O padrão passado para um projetor de alta-velocidade (que projeta padrões seqüênciais com a freqüência de 240Hz) que projeta o padrão em níveis de cinza. Uma câmera CCD, "preto-e-branco" de alta velocidade de captura (em torno de 240 quadros por segundo) é usada para capturar as imagens de cada canal de cor separadamente, de onde as informações 3D do objeto são obtidas. Outra câmera CCD é sincronizada com o projetor e alinhada com a câmera preto-e-branco é usada para tirar fotos 2D da cena para fazer mapeamento de textura. Experimentos também foram desenvolvidos com essa técnica para um padrão RGB variando trapezoidalmente, concluindo que com ondas trapezoidais os resultados são até seis vezes mais precisos. A Figura 3.8 mostra como o sistema funciona, e a Figura 3.9 exibe alguns resultados da técnica.

Vieira et al. (VIEIRA et al., 2004) propuseram uma forma baseada em luz estruturada para a aquisição de profundidade da cena em tempo real. Com uma câmera de vídeo e um projetor devidamente sincronizados e calibrados, é possível obter a profundidade de pontos da cena através de uma triangulação feita entre a câmera e o projetor. Essa técnica é uma das poucas que recupera as informações de cor da cena, conforme mostra a Figura 3.10.

A utilização de luz estruturada é atualmente a forma mais precisa de se obter informação de profundidade de uma cena, e poderia ser sem maiores problemas integrado a sistemas que seguem a arquitetura de vídeo 3D descrita no Capítulo 4. Seu maior problema é a necessidade de calibração entre a câmera e o projetor, que precisa ser refeita toda a vez que um dos equipamentos é movimentado. Isso restringe a câmera que captura a cena a um único ponto de vista e também o ambiente de filmagem do vídeo à um estúdio.

3.4 Comparação Entre os Métodos

Ao longo desse capítulo, foram analisados vários métodos de extração da informação referente à profundidade de uma cena. Com essa informação, é possível utilizar métodos para possibilitar ao usuário a visualização da cena a partir de pontos de vista diferentes dos capturados pela(s) câmera(s). Todavia, todos métodos possuem limitações. Nessa

Figura 3.7: Demonstração do resultado da técnica de Li Zhang et al. (ZHANG; CURLESS; SEITZ, 2002), extraído de (ZHANG; CURLESS; SEITZ, 2002).

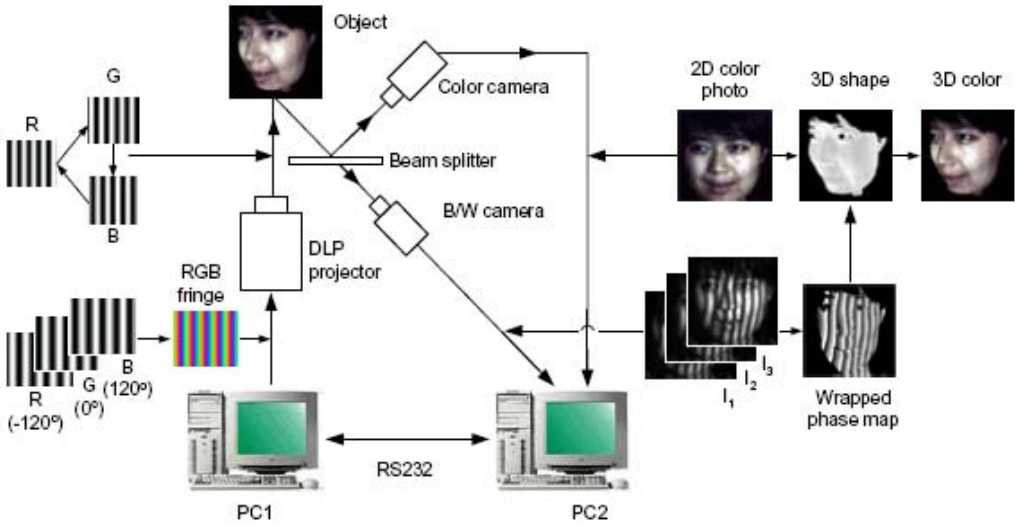

Figura 3.8: Esquema explicativo da técnica de Song Zhang et al. (ZHANG; HUANG, 2004): um padrão de cores RGB é criado pelo PC1 e projetado em escala de cinza sobre a cena; uma câmera preto-e-branco (*B/W*) captura as imagens da cada canal de cor para extrair as informações 3D da cena e outra câmera colorida captura imagens 2D para mapeamento de textura; extraído de (ZHANG; HUANG, 2004).

secção é feita uma comparação entre os métodos, citando as vantagens e limitações de cada um deles. É comparada a velocidade de aquisição de cada método, portabilidade do equipamento necessário para capturar as informações necessárias da cena, qualidade dos resultados e restrições com relação a ambientes em que as técnicas podem ser aplicadas. Ao final da secção, a tabela 3.1 sumariza essa análise.

3.4.1 Velocidade de Aquisição

A velocidade de aquisição de um mapa de disparidade é um ponto crucial para a criação de um vídeo 3D. No contexto dessa dissertação, é extremamente importante que as informações necessárias para compor o vídeo sejam extraídas em tempo real. Dentre as

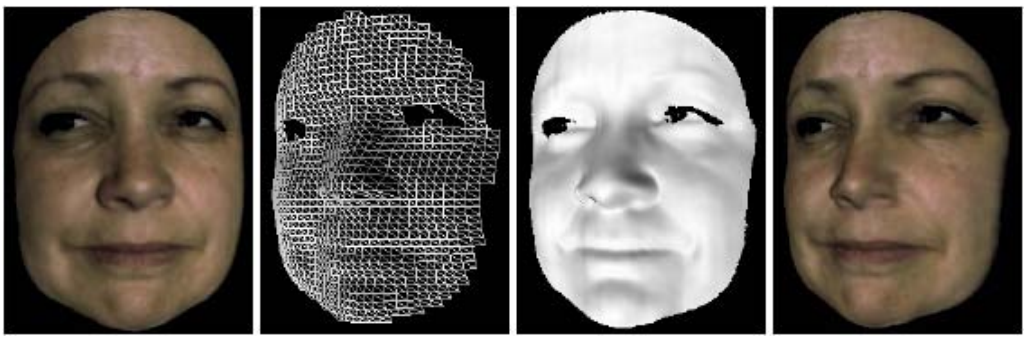

Figura 3.9: Resultados da técnica de Song Zhang et al. (ZHANG; HUANG, 2004): à esquerda a imagem colorida para mapeamento de textura, logo a seguir o modelo 3D em *wireframe*, com a iluminação aplicada e à direita o modelo 3D com a textura mapeada; extraído de (ZHANG; HUANG, 2004).

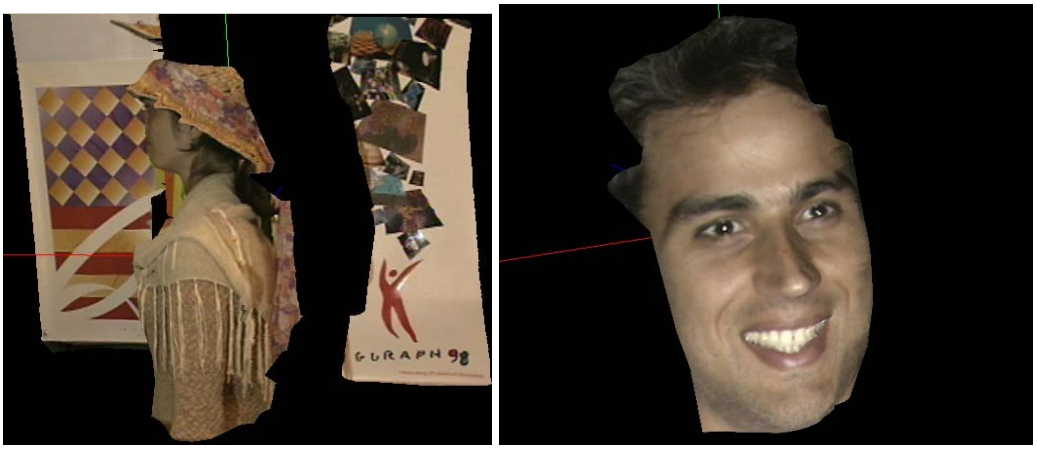

Figura 3.10: Demonstração da técnica de Vieira et al. (VIEIRA et al., 2004), extraído de (VIEIRA et al., 2005).

técnicas com boa velocidade de aquisição estão os alguns algoritmos de cálculos de disparidade a partir de estéreo, como por exemplo Yang et al. (YANG; POLLEFEYS, 2003) e Gong et al. (GONG; YANG, 2005), algumas aplicações que utilizam luz estruturada, como Vieira et al. (VIEIRA et al., 2004), e as câmeras 3D.

3.4.2 Portabilidade

Um fator importante na construção de um vídeo 3D diz respeito à possibilidade de se poder movimentar as câmeras (e outros aparelhos necessários para a captura) livremente no ambiente onde o vídeo está sendo gravado. Tendo em vista que as câmeras atuais podem ser carregadas para qualquer lugar com facilidade, deseja-se que o equipamento necessário para captura de um video 3D também seja portável dessa forma.

Dentre os métodos analisados, são portáveis os que calculam a disparidade a partir de estéreo (já que existem pares de câmeras fixas e pré-calibradas que podem ser facilmente movimentadas) além das câmeras 3D. Os métodos com luz estruturada pecam na portabilidade, já que precisam ser cuidadosamente calibrados e normalmente possuem uma antes de serem utilizados.

3.4.3 Qualidade dos Mapas de Disparidade Gerados

A qualidade dos mapas de disparidade gerados influencia diretamente na qualidade do vídeo 3D. Um mapa de baixa qualidade pode produzir artefatos na cena, ou falhas na composição dos objetos, o que prejudica sensivelmente a visualização.

Os métodos que apresentam melhor qualidade nesse aspecto são as câmeras 3D e as técnicas de Li Zhang et al. ((ZHANG et al., 2004), (ZHANG; CURLESS; SEITZ, 2002)) e de Song Zhang et al. ((ZHANG; HUANG, 2004)). Os métodos que calculam disparidade a partir de estéreo tendem a produzir ruído, por isso tem qualidade inferior. A técnica de Vieira et al. (VIEIRA et al., 2004), faz a reconstrução geométrica extraindo um reduzido número de vértices, apenas o necessário para construir uma malha parcial do modelo 3D (Figura 3.10).

3.4.4 Outras Restrições

Outras restrições também afetam diretamente a qualidade do vídeo. A técnica de Li Zhang et al. (ZHANG; CURLESS; SEITZ, 2002) , por exemplo, possui sérias restrições com relação a cor do objeto a ser filmado, que precisa necessariamente ser clara (ZHANG; CURLESS; SEITZ, 2002). A técnica de Zhang et al. (ZHANG et al., 2004) que utiliza estéreo e variação temporal da cena, embora suporte objetos dinâmicos, precisa ter seus mapas calculados com antecedência, já que utiliza informação de quadros futuros do vídeo e o cálculo de disparidade é um processo demorado.

Algumas técnicas possuem restrições com relação à utilização de objetos dinâmicos. As primeiras técnicas que utilizavam luz estruturada tinham esse problema, pela dificuldade de se obter a informação 3D a partir de uma ou poucas imagens (essas técnicas não foram analisadas no trabalho por não possuírem aplicações para vídeos 3D devido às suas restrições). Cálculos de mapas de disparidade a partir de estéreo utilizando *Graph Cuts* (HONG; CHEN, 2004) também não permitem a utilização de objetos dinâmicos devido a velocidade de obtenção de um mapa. Todavia, se os mapas forem pré-processados, estes podem ser utilizados para geração de vídeos 3D.

A Tabela 3.1 sumariza, a comparação entre os métodos de captura de informação tridimensional de cenas.

Tabela 3.1: Comparação entre os métodos de extração de informação referente à profundidade da cena.

Método	Tempo Real	Portável	Alta precisão, boa resolução	Permite objetos dinâmicos
Estéreo				
(VEKSLER, 2003)	Não	Sim	Não	Não
(YANG; POLLEFEYS, 2003)	Sim	Sim	Não	Sim
(YANG; POLLEFEYS; LI, 2004)	Sim	Sim	Não	Sim
(VEKSLER, 2005)	Sim	Sim	Não	Sim
(GONG; YANG, 2003)	Não	Sim	Não	Não
(GONG; YANG, 2005)	Sim	Sim	Não	Sim
(HONG; CHEN, 2004)	Não	Sim	Não	Não
Estéreo levando em consideração variações temporais				
(ZHANG et al., 2004)	Não	Não	Sim	Sim
Luz Estruturada				
(VIEIRA et al., 2004)	Sim	Não	Não	Sim
(ZHANG; CURLESS; SEITZ, 2002)	Não	Não	Sim	Sim
(ZHANG; HUANG, 2004)	Não	Não	Sim	Sim
Câmeras 3D				
(IDDAN; G., 2001)	Sim	Sim	Sim	Sim
(IIZUKA; KAWAKITA, 2002)	Sim	Sim	Sim	Sim

3.5 Discussão

Nesse capítulo, foram analisados e comparados diversos métodos para extrair a profundidade de uma cena. Foram colocados, frente a frente, técnicas baseadas em luz estruturada, estéreo e câmeras 3D.

A possibilidade de se calcular a profundidade através de um mapa de disparidades obtido a partir de um par de imagens estéreo é uma forma bastante simples dentre as analisadas, e pode ser facilmente aplicada em qualquer tipo de ambiente, pois não é dependente de hardware especializado (apenas duas câmeras devidamente calibradas são suficientes). Sua maior restrição é a qualidade dos mapas gerados, que tendem a apresentar ruído e precisão limitada.

Uma alternativa é a utilização de luz estruturada para obter a informação de profundidade da cena. Os estudos atuais demonstram que esse tipo de abordagem possui uma precisão bastante grande . Sua maior restrição é a dificuldade para mover o equipamento, que precisa estar sempre calibrado.

A utilização de câmeras 3D é uma boa alternativa para aqueles que dispõem de recursos para adquirir uma. Esse tipo de equipamento pode ser usado sem maiores problemas na geração do vídeo 3D e geram mapas de disparidade com precisão aceitável.

4 UMA ARQUITETURA PARA VISUALIZAÇÃO DE VÍDEO 3D EM TEMPO REAL

Este capítulo descreve uma arquitetura para captura e exibição de vídeos 3D em tempo real. Inicialmente é apresentada e analisada a arquitetura proposta. Em seguida, é feita uma descrição do cálculo do *3D Image Warping* (MCMILLAN, 1997) usado para a geração de novos pontos de vista da cena.

4.1 Arquitetura de Sistema para Captura e Visualização de Vídeo 3D

A arquitetura proposta nessa dissertação segue o esquema descrito na Figura 4.1. A partir da cena, são capturadas as informações de cor e profundidade para cada pixel de cada quadro do vídeo. Com a informação de profundidade é possível calcular a disparidade generalizada (MCMILLAN, 1997) para cada pixel da cena. O conceito de disparidade generalizada (MCMILLAN, 1997) foi criado por McMillan e é usado pela equação do *3D warping* (Equação 4.6) no cálculo de novos pontos de vista da cena. A Seção 4.1.2 detalha meios de se obter a disparidade generalizada (MCMILLAN, 1997). Em seguida, a geração de novos pontos de vista é realizada utilizando uma extensão do algoritmo de *warping 3d* (MCMILLAN, 1997) para vídeos. A este algoritmo, chamamos *3D video warping*. Quatro quadros de um exemplo de vídeo 3D produzido com essa abordagem podem ser vistos na Figura 4.2.

4.1.1 Processos de Captura do Vídeo 3D

Conforme é mostrado no bloco destacado do fluxograma exibido na Figura 4.1, a primeira etapa do processo consiste na extração de profundidade da cena, o que pode ser feito utilizando uma entre as diversas técnicas descritas no Capítulo 3. Tem-se disponível hoje para esse fim algoritmos de extração de disparidade a partir de imagens estereoscópicas, câmeras 3D e técnicas que utilizam luz estruturada.

Uma das formas mais acessíveis de se extrair a informação de profundidade de uma cena é o uso de técnicas que utilizam imagens estereoscópicas. Essas técnicas só requerem um par de câmeras calibradas para gerar o par estéreo usado nos algoritmos. Existem atualmente no mercado, câmeras de vídeo estéreo pré-calibradas (PGR, 2005) que podem ser usadas sem maiores problemas para captura de vídeos estéreos. Como foi visto no Capítulo 3, é possível extrair mapas de disparidade com essas técnicas em tempo real, todavia, a precisão dos resultados ainda é insatisfatória para aplicações que necessitam de vídeo 3D.

Técnicas que usam luz estruturada também podem ser utilizadas na extração de ma-

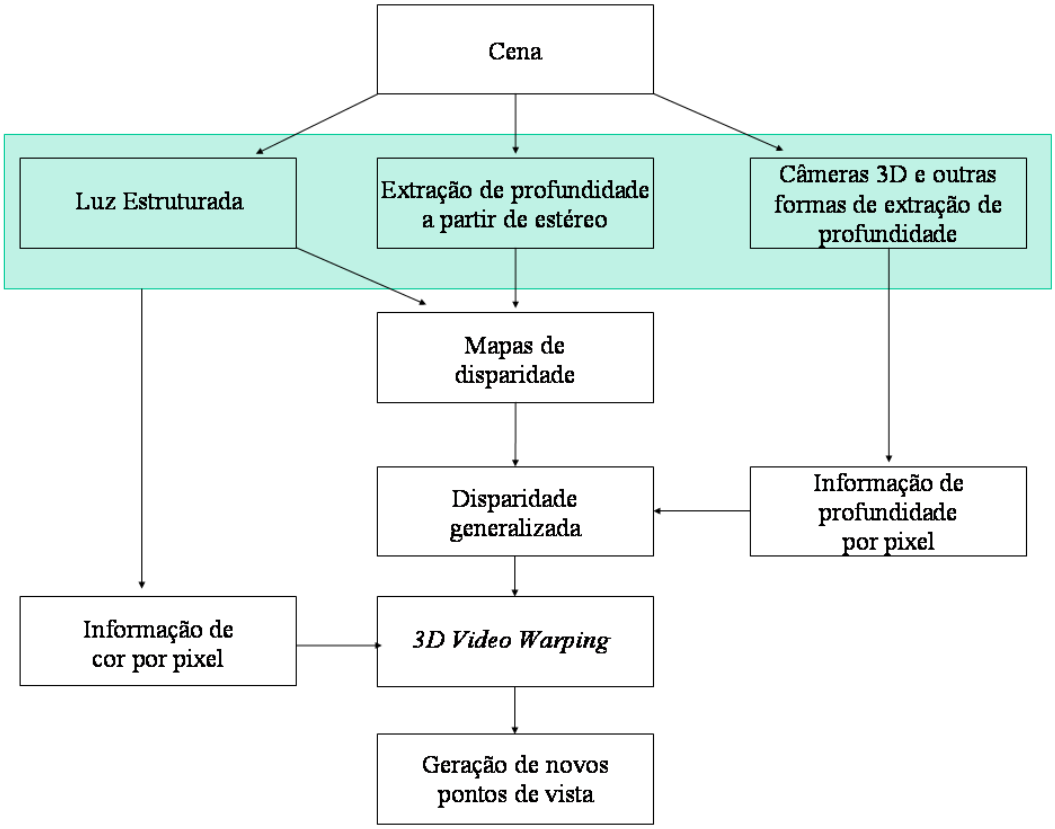

Figura 4.1: Fluxograma que descreve o funcionamento da arquitetura proposta. O bloco destacado mostra processos de captura das informações de cor e profundidade que podem ser usados. Com a informação de profundidade, extraída a partir de estéreo, de luz estruturada, de câmeras 3D ou de qualquer outra forma de extração de profundidade da cena, e a cor recebidas como entrada, é iniciado o processo de geração de novos pontos de vista da cena. Obtém-se, então, a disparidade generalizada (MCMILLAN, 1997) de cada pixel a partir das informações de profundidade ou mapas de disparidade e então, usando o algoritmo de *3D video warping*, pontos de vista alternativos da cena podem ser gerados.

pas de profundidade da cena. Esse tipo de técnica é, em geral, mais precisa que as outras abordagens para extração de profundidade da cena. Sua maior restrição é a dificuldade de deslocar o equipamento (que consiste de, no mínimo, uma câmera e um projetor de luz), já que este precisa estar sempre calibrado. Deve-se levar em consideração que não só a informação de profundidade é importante, mas também a informação de cor da cena. Logo, técnicas que usam luz estruturada que não capturam a cor da cena (ZHANG; CURLESS; SEITZ, 2002) não são apropriadas para a geração de video 3D.

A utilização de câmeras 3D é uma abordagem promissora. Os maiores problemas dessa abordagem dizem respeito a precisão e ao alcance das câmeras 3D (KAWAKITA et al., 2002) (IDDAN; G., 2001). Todavia, as imagens produzidas possuem alta resolução de captura. A *Axi-Vision camera*, por exemplo, gera quadros de vídeo de até 920.000 pixels (IIZUKA; KAWAKITA, 2002). Futuros projetos de câmeras 3D (KAWAKITA et al., 2002) (IDDAN; G., 2001) buscam aumentar a precisão do equipamento, além de reduzir o custo para a aquisição de um modelo. Com isso, os problemas relacionados a essa abordagem tendem a diminuir.

Figura 4.2: Quatro quadros de um vídeo 3D construído com a arquitetura aqui proposta. Tem-se acima os quadros vistos sob o ponto de vista da câmera, abaixo os quadros vistos de pontos de vista alternativos.

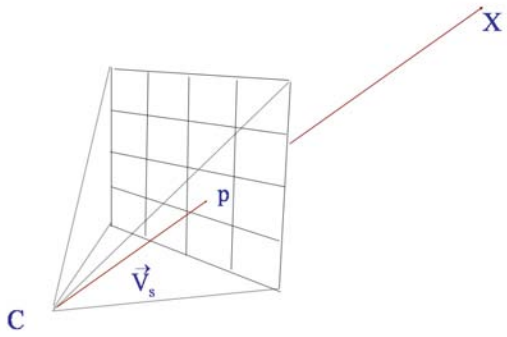

Figura 4.3: Câmera observado uma cena 3D: C - centro de projeção; X - ponto no espaço visualizado pela câmera; p - projeção de X no plano de imagem; $\vec{V}$ - vetor do centro de projeção *C* até *p*

4.1.2 Tratando a Informação de Profundidade da Cena

Como a arquitetura proposta permite a utilização de qualquer método de extração de profundidade, esta poderia ser representada de diversas maneiras. Antes de passar para o algoritmo de *3D video warping*, a informação de profundidade deve ser convertida em disparidade generalizada (MCMILLAN, 1997).

Considere uma câmera cujo centro de projeção esteja no ponto C, como mostra a Figura 4.3. A distância entre o ponto X visualizado pela câmera e C é dada pelo vetor $\vec{XC}$. Seja p a projeção do ponto X no plano da imagem. A distância entre p e C é dada pelo vetor $\vec{V}$. Considere agora um escalar t definido pela equação $t = \|\vec{XC}\|/\|\vec{V}\|$. A disparidade generalizada (MCMILLAN, 1997) δ referente à p é dada pela Equação 4.1.

$$\delta = \frac{1}{t} = \frac{\|\vec{V}\|}{\|\vec{XC}\|} \tag{4.1}$$

Seja Z_p a coordenada *Z* do ponto *X* 4.3 definida com relação ao sistema de referência da câmera. O valor da disparidade generalizada δ associada por pixel *p* (Figura 4.3) é dada pela Equação 4.2.

$$\delta = \frac{f}{Z_p} \tag{4.2}$$

O valor da disparidade generalizada (MCMILLAN, 1997) para mapas de disparidade obtida por meio do uso de técnicas estereoscópicas é obtido substituindo o valor de Z_p dado pela Equação 4.3 (TRUCCO; VERRI, 1998). Na Equação 4.3, f é a distância focal das câmeras utilizadas, b é a baseline (distância entre os centros de projeção das duas câmeras), e d é o valor da disparidade associado ao pixel em questão. Os resultados obtidos através de técnicas de luz estruturada também podem usar essa equação.

$$Z_p = \frac{fb}{d} \tag{4.3}$$

4.1.3 *3D Video Warping*

Uma etapa de suma importância na arquitetura proposta nessa dissertação, é a geração de novos pontos de vista da cena. Após a obtenção da disparidade generalizada (MCMILLAN, 1997), é usado um algoritmo de *3D image warping* (MCMILLAN, 1997), que rearranja os pixels de cada quadro do vídeo de forma a dar a impressão de se estar observando a cena de outro ponto de vista. Nessa dissertação, o algoritmo de *3D image warping* (MCMILLAN, 1997) foi estendido para tratar cada um dos quadros do vídeo, produzindo assim o que chamamos de *3D video warping*.

No processo de *warping 3D*, são necessários apenas a profundidade de cada pixel, distância focal, o campo de visão da câmera, posições e orientações inicial e desejada. A cena representada pela imagem pode então ser observada a partir de qualquer novo ponto de vista utilizando a Equação 4.6. Na Figura 4.4 tem-se um exemplo de warping 3D obtido a partir de um quadro de um vídeo sintético. À esquerda, tem-se a imagem original e à direita, a imagem observada de outro ponto de vista obtido a partir da imagem original. O que aparece em preto na imagem à direita são os pontos para os quais não se dispõe de informação de cor e profundidade por estes estarem inicialmente ocultos no quadro original.

Figura 4.4: Exemplo de warping de uma imagem, a esquerda a imagem original, à direita observada de um outro ponto de vista

Para entender o funcionamento do processo de *warping 3D* de uma imagem, considere uma câmera com centro de projeção *C* e campo de visão (*FOV*) definido pela matriz *M (3x3)*, de acordo com a Equação 4.4, que tem suas colunas composta pelos vetores $\vec{a}$, $\vec{b}$ *e* $\vec{c}$ de acordo com a Figura 4.5. Considere uma imagem cujos pixels (*p)* possuem, além da cor, um valor escalar (*t*) conforme definido na Seção 4.1.2. Considere agora um mesmo ponto *X* no espaço euclideano tridimencional sendo observado por duas câmeras simultâneamente como o ilustrado na Figura 4.6.

$$M = \begin{pmatrix} a_i & b_i & c_i \\ a_j & b_j & c_j \\ a_k & b_k & c_k \end{pmatrix} \tag{4.4}$$

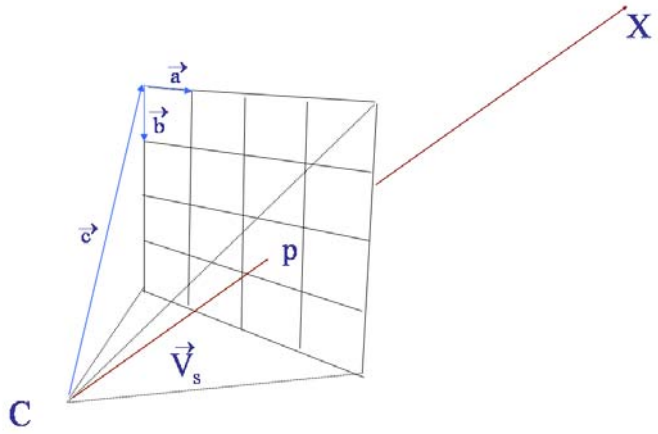

Figura 4.5: Especificação da câmera: C - centro de projeção; $\vec{a}$, $\vec{b}$ e $\vec{c}$ - vetores que definem a câmera; X - ponto no espaço visualizado pela câmera; p - pixel da imagem (onde o ponto X foi projetado); V - vetor do centro de projeção *C* até o pixel *p*; t = escalar descrito na seção 4.1.2.

O posicionamento no espaço do ponto *X* é especificado de acordo com a Equação 4.5, onde C_o é o centro de projeção da câmera origem, M_o é a matriz da câmera origem, p_o é o pixel que representa a projeção do ponto X na câmera origem, t_o é o valor escalar t associado ao pixel correspondente à projeção de X sobre o plano de imagem da câmera origem. Seja C_d o centro de projeção da câmera destino, M_d é a matriz da câmera destino, p_d é o pixel que representa a projeção do ponto X na câmera destino, t_d é o valor escalar *t* associado ao pixel correspondente na câmera destino.

$$X = C_o + t_o M_o p_o = C_d + t_d M_d p_d \tag{4.5}$$

A partir da Equação 4.5, pode-se isolar p_d (a projeção do ponto *X* na imagem destino) junto com seu valor de profundidade t_d. Utilizando-se de equivalência projetiva, pode-se eliminar t_d da equação. Divide-se os dois lados da equação por t_o, obtém-se a Equação 4.6 expressa em termos da disparidade generalizada δ. Ela nos dá as coordenadas do pixel p_t (em coordenadas homogêneas) que correspondem ao mapeamento do pixel original

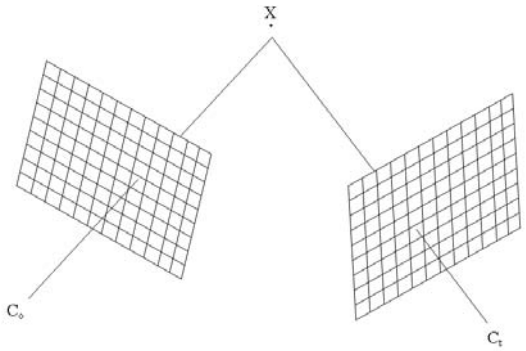

Figura 4.6: Ponto X visualizado por duas câmeras

para o plano de imagem destino. Essa técnica pode ser implementada de modo a fazer os cálculos de forma incremental (MCMILLAN, 1997), o que torna a operação mais otimizada.

$$p_d \doteq M_d^{-1}((C_o - C_d) + \delta M_o p_o) \tag{4.6}$$

Utilizando a técnica de *3D Image Warping*, pode-se observar uma imagem com informação de profundidade sob qualquer ponto de vista. Todavia, como existem áreas da imagem original que estão ocultas do ponto de vista da câmera original, essa informação não se encontra disponível e tais áreas aparecem como "buracos" nas imagens reconstruídas (Figura 4.4). Esta é uma limitação da técnica de *3D image warping* (MCMILLAN, 1997) e, consequentemente, de *3D video warping*. A ocorrência desses "buracos" pode ser evitada tratando-se cada pixel como o vértice de uma malha. A imagem destino seria então reconstruída renderizando-se os polígonos definidos por tal malha, uma vez que os vértices tenham sido mapeados utilizando a Equação 4.6. Essa abordagem de reconstrução pode, entretanto, levar a introdução de artefatos indesejáveis, resultantes da conexão inapropriada de superfícies disjuntas. Esses artefatos são comumente chamados de "skins" (MCMILLAN, 1997) (MARK; MCMILLAN; BISHOP, 1997).

4.1.3.1 Informações de Câmera

Leonard McMillan (MCMILLAN, 1997) utiliza, para o cálculo do *3D Image Warping*, a disparidade generalizada de cada pixel da imagem e as informações de câmera como parâmetros de entrada. Aqui será demonstrada uma forma de cálculo dos parâmetros de câmera baseada nas definições da matriz *M* descrita na equação vista na secção 4.1.3.

M é uma matriz 3x3 (Equação 4.4) composta pelos vetores $\vec{a}$, $\vec{b}$ e $\vec{c}$, linearmente independentes, que definem a base no espaço de câmera. Os comprimentos dos vetores $\vec{a}$ e $\vec{b}$ são, respectivamente, a largura e a altura dos pixels da imagem, enquanto $\vec{c}$ é o vetor que vai do centro de projeção da câmera até o canto superior esquerdo da imagem, como mostra a Figura 4.5.

Os valores de $\vec{a}$, $\vec{b}$ e $\vec{c}$ podem ser obtidos de maneira simples, baseando-se apenas nas dimensões do quadro em pixels e na distância focal da câmera. As Equações 4.7 descrevem a obtenção dos valores dos vetores, onde, *H* representa a altura em pixels da imagem, *W* a largura e *f* a distância focal da câmera. Dada uma posição e orientação para a câmera destino, obtém-se a matriz *M* correspondente e aplica-se na Equação 4.6.

$$\begin{array}{ccc} a_x = 1/W & a_y = 0 & a_z = 0 \\ b_x = 0 & b_y = -1/H & b_z = 0 \\ c_x = -W/2 & c_y = H/2 & c_z = f \end{array} \quad (4.7)$$

4.2 Resumo do Capítulo

Nesse capítulo foi apresentada uma arquitetura para captura e exibição de vídeos 3D em tempo real, detalhando as etapas necessárias para a sua produção. Inicialmente, foram analisados os possíveis processos de captura de informação de cor e profundidade da cena. Foram consideradas para esse fim o uso de câmeras 3D, luz estruturada e algoritmos de extração de profundidade a partir de imagens estereoscópicas.

Logo após, foi feita uma análise dos métodos para se extrair a disparidade generalizada (MCMILLAN, 1997) através da informação de profundidade resultante da etapa anterior da arquitetura. Foram apresentadas equações para conversão de mapas de disparidade e informação de profundidade em disparidade generalizada (MCMILLAN, 1997).

Em seguida foi descrito o processo de *3D Video Warping* a fim de permitir a geração de novos pontos de vista da cena a partir da seqüência de vídeo original.

5 ANÁLISE DOS RESULTADOS

Nesse capítulo é feita uma análise dos resultados obtidos com a implementação de um protótipo de sistema para a captura e visualização de vídeos 3D baseado na arquitetura descrita no Capítulo 4.

Utilizando o protótipo foram avaliados algoritmos para extração de profundidade a partir de estéreo em tempo real. Foram considerados tanto a qualidade dos mapas gerados, quanto a exatidão da reconstrução feita a partir dos mapas. As técnicas avaliadas foram Yang et al. (YANG; POLLEFEYS, 2003), Gong et. al (GONG; YANG, 2005) e Yang et al. (YANG; POLLEFEYS; LI, 2004). A técnica de Vieira et al. (VIEIRA et al., 2004) não foi avaliada por não possuir descrição suficiente disponível, além disso, assim como as demais técnica de extração de profundidade baseadas no uso de luz estruturada, esta requer um ambiente especial para a captura das informações de profundidade. Já a técnica de Veksler (VEKSLER, 2003) não produz resultados em tempo real.

Foram avaliados apenas algoritmos para extração de profundidade capazes de produzir resultados em tempo real. Isso justifica-se pois em grande parte dos vídeos, como eventos esportivos e transmissões ao vivo, a necessidade de informação é imediata, e o espectador não estaria disposto a esperar uma etapa de pré-processamento para a geração do vídeo.

5.1 Avaliações de Algoritmos de Extração de Profundidade

Nessa seção, é feita a comparação entre os resultados obtidos com o uso de diversas técnicas de extração de disparidade em tempo real. Diferentemente das demais formas de comparação de qualidade de mapas de disparidade presentes na literatura (SZELISKI; ZABIH, 1999) (SZELISKI, 1999) (SCHARSTEIN; SZELISKI, 2002b) (GONG; YANG, 2002), a abordagem adotada utiliza a informação exata da profundidade da cena para a comparação. A fim de se extrair com a precisão desejada a informação de profundidade da cena para a avaliação dos mapas, foi criado um ambiente sintético. Em um ambiente controlado é possível fazer uma análise bastante criteriosa da qualidade dos resultados das técnicas. As cenas construídas para a avaliação foram cuidadosamente criadas para abordar vários possíveis casos que podem ser encontrados no mundo real, tendo locais com poucos objetos (que geram mapas de disparidade mais homogêneo), com vários objetos (que servem para analisar principalmente a quantidade de ruído presente nas bordas dos objetos), utilizando objetos com pouca textura (analisando a capacidade do algoritmo no tratamento de ambiguidades) e com bastante textura (para analisar o desempenho quando há transições grandes de cor sem muita variação da profundidade).

Com o uso da arquitetura proposta nessa dissertação é possível avaliar a reconstrução de novos pontos de vista da cena gerados com a utilização dos mapas de disparidade resultante das técnicas que se deseja avaliar. A utilização do algoritmo de *warping 3D* de

McMillan (MCMILLAN, 1997) e a informação exata da profundidade da cena, garantem que os novos pontos de vista a serem gerados serão consistentes com a visualização da cena original. O uso de um ambiente sintético permite a obtenção de mapas de profundidade exatos a partir da informação armazenada no Z-buffer. A Equação 5.1 mostra como obter o valor da disparidade generalizada de um dado pixel. Nesta equação, Z_b é a informação armazenada no Z-buffer, n é o valor definido para plano de recorte *near* e f é o valor definido para plano de recorte *far*. O Apêndice A apresenta uma derivação para a Equação 5.1.

$$\delta \quad = \quad -Z_b + 1 + Z_b\,(n/f) \tag{5.1}$$

5.1.1 Avaliação de qualidade de mapas de disparidade

Alguns pesquisadores já buscaram formas de medir a qualidade de mapas de disparidade gerados a partir de estéreo. Szeliski et al. (SZELISKI; ZABIH, 1999) realizou uma comparação pixel a pixel entre mapas de disparidade gerados a partir de estéreo de cenas reais e uma disparidade de referência da cena. Essa disparidade utilizada como referência é calculada manualmente pixel a pixel para o par de imagens desejado. Apesar disso, os mapas de disparidade usados como referência contém imprecisões. Essa técnica não considera a velocidade dos algoritmos e nem a capacidade de reconstrução da cena para um novo ponto de vista, que são aspectos fundamentais para a captura e visualização, respectivamente, de vídeos 3D em tempo real.

Scharstein et al. (SCHARSTEIN; SZELISKI, 2002b) procuraram melhorar a métrica descrita em (SZELISKI; ZABIH, 1999). Os pesquisadores consideram a suavidade do mapa gerado e, para melhor avaliar o desempenho dos algoritmos, levaram em consideração o desempenho dos algoritmos em áreas com e sem texturas. Todavia, também não consideram a velocidade dos algoritmos e nem a capacidade de reconstrução da cena para um novo ponto de vista.

Gong et al. (GONG; YANG, 2002) propuseram uma maneira de comparar métodos de extração de disparidade. Essa técnica se baseia na premissa de que a disparidade dos pixels é praticamente invariante para regiões de cores homogêneas. O mapa de disparidade é, então, avaliado por sua suavidade nas regiões de cores homogêneas.

Cabe ressaltar que nenhuma das análises anteriores se valeu informação de profundiade exata. A abordagem aqui proposta, a fim de se obter uma análise criteriosa da qualidade dos mapas gerados, se utiliza de mapas de disparidade exatos obtidos por meio do uso de cenas sintéticas, utilizando técnicas de computação gráfica.

5.1.2 Avaliação de qualidade baseada em predição de movimento

Já foram pesquisadas formas de se comparar métodos de extração de profundidade a partir de estéreo que comparam a reconstrução de novos pontos de vista da cena. Todavia, a abordagem presente na literatura diferencia-se sensivelmente da apresentada nesse trabalho.

Szeliski (SZELISKI, 1999) propôs uma métrica para avaliar métodos de predição de movimento. Szeliski utilizou um par estéreo e mais uma câmera. O par serve para a extração da disparidade e a terceita câmera para servir como padrão para comparar a reconstrução de um novo ponto de vista da cena.

A técnica de comparação adotada nessa dissertação diferencia-se sensivelmente da adotada por Szeliski (SZELISKI, 1999). A abordagem aqui apresentada permite a comparação para qualquer ponto de vista desejado, não só para uma única câmera. Isso é obtido comparando imagens de referência da cena sintética obtidas a partir de pontos de vista arbitrários com os resultados obtidos ao realizar o *warping 3D* (utilizando os mapas de disparidade obtidos pelos vários algoritmos) para estes pontos de vista. Nesse caso, as diferenças entre as imagens transformadas e as imagens de referência refletem as diferenças nos mapas de disparidade

5.1.3 Avaliando o Desempenho dos Algoritmos de Extração de Profundidade em Tempo Real

De modo a avaliar os vários algoritmos utilizando a cena sintética utilizou-se a seguinte configuração: uma câmera estéreo virtual com 1 unidade de distância focal e 6 unidades de baseline (Figura 5.1). Para a avaliação, foram usados pares de imagens com poucos objetos, com muitos objetos e também pequenas seqüências de vídeos. As comparações foram feitas entre os mapas de disparidade (pixel a pixel) e também com a reconstrução da cena com uma câmera virtual convencional, a partir de um ponto de vista alternativo levemente deslocado, com uma translação de 4,625 unidades para a direita e rotação de 4,5 graus em torno do eixo *Y* no sentido anti-horário.

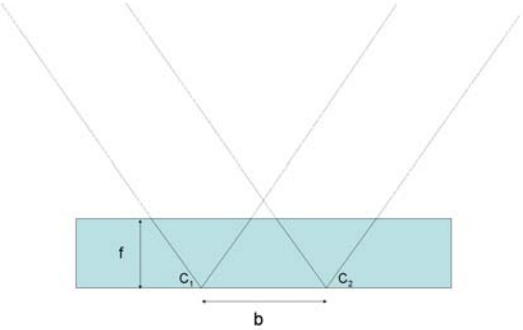

Figura 5.1: Representação esquemática de uma câmera estéreo virtual. A distância *b*, entre os centros de projeções C_1 e C_2, representa a baseline, e f sua distância focal

A Figura 5.2 mostra dois quadros de um vídeo 3D usados para a avaliação dos algoritmos. Foram utilizadas imagens com vários objetos, para avaliar o desempenho dos algoritmos com imagens contendo altas freqüências, e com poucos objetos, para avaliar o desempenho em cenas mais homogêneas e suaves. Em ambos os casos, as paredes foram recobertas com textura para reduzir a ambiguidade durante o processamento dessas imagens pelos algoritmos de estéreo.

A Figura 5.3 mostra os resultados da reconstrução da cena com poucos objetos feita com os algoritmos avaliados. A Figura 5.4 mostra os mapas de disparidades gerados pelos algoritmos. Nota-se ao analisar a Figura 5.3 que a técnica de Yang et al. (YANG; POLLEFEYS; LI, 2004) obteve visualmente o melhor resultado (Figuras 5.3 (b-e)). Percebe-se que, quanto mais níveis de mipmap são usados, mais a quantidade de ruído na reconstrução é reduzida. A técnica de Gong et al. (GONG; YANG, 2005) obteve visualmente um resultado mediano (Figura 5.3 a), que foi bastante prejudicado pela quantidade de ruídos na reconstrução da imagem. A técnica de Yang et al. (YANG; POLLEFEYS, 2003) obteve, visualmente, quando poucos níveis de mipmap foram usados (Figuras 5.3 (f-g)),

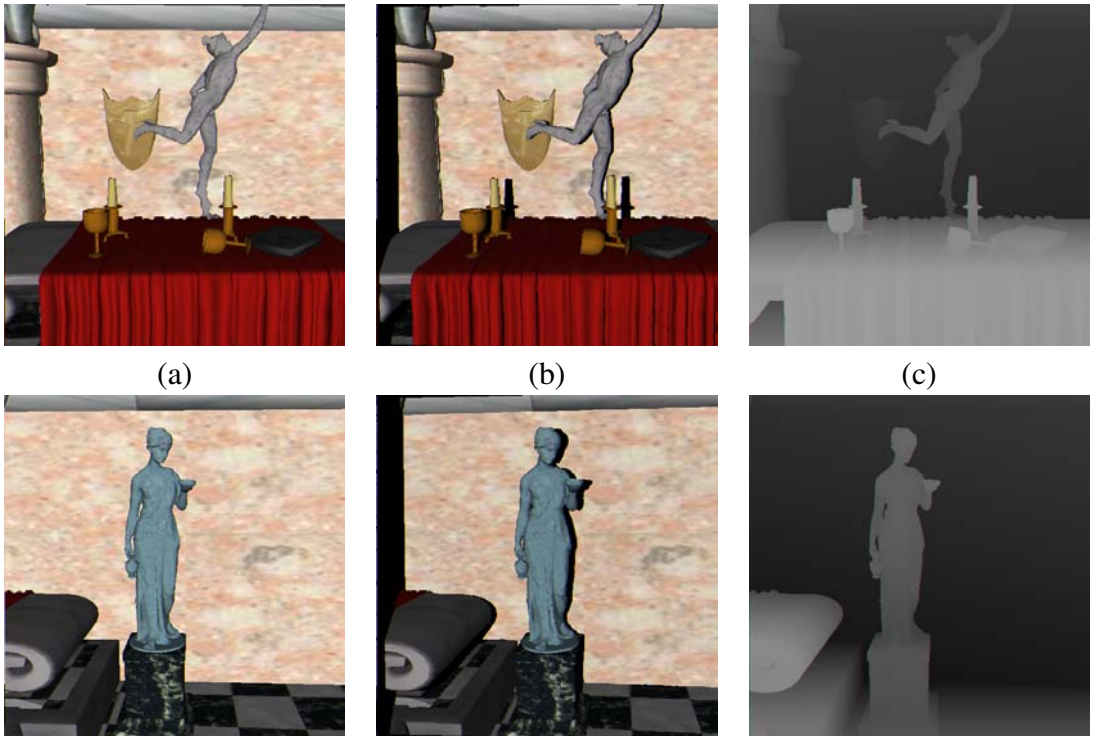

Figura 5.2: Dois quadros de vídeos 3D usados na avaliação dos algoritmos de extração de profundidade a partir de estéreo em tempo real. Acima a cena com vários objetos, abaixo a cena com poucos objetos. As imagens (a) e (d) mostram os pontos de vista originais da cena. (b) e (e) mostram pontos de vista alternativos obtidos pelo processo de *3D video warping* descrito no Capítulo 4. (c) e (f) mostram os mapas de disparidade generalizada extraídos a partir do Z-buffer da cena.

resultados de qualidade muito baixa devido ao excesso de ruído. Já com o uso de vários níveis de mipmap (Figuras 5.3 (j-l)) o mapa de disparidades fica super-suavizado, o que resulta em reconstruções bastante imprecisas nas bordas (Figuras 5.3 (j-l)). O uso de um número médio de mipmaps para essa técnica (Figuras 5.3 (h-i)) gerou um resultado de qualidade visual realtivamente baixa, devido tanto ao ruído quanto a setores super-suavizados da imagem.

Analisando os mapas de disparidade da Figura 5.4 em correspondência direta com os elementos da Figura 5.3 nos leva a perceber que a quantidade de ruído no mapa de disparidade afeta diretamente o resultado da reconstrução. Os mapas mais ruidosos (Figuras 5.4 (a, f, g)) tiveram a percepção tridimensional da cena prejudicada devido a alta quantidade de pixels reconstruídos erroneamente. Todavia, fica evidente também que mapas muito suaves também podem influenciar negativamente no resultado visual da reconstrução (Figuras 5.4 (j-l)). Essa imprecisão afeta principalmente as bordas do objeto (Figuras 5.3 (j-l)).

A Figura 5.5 mostra os resultados da reconstrução da cena com vários objetos a partir de um novo ponto de vista feita com os algoritmos avaliados. A Figura 5.6 mostra os mapas de disparidades gerados pelos algoritmos. Com relação às reconstruções, a técnica de Yang et al. (YANG; POLLEFEYS; LI, 2004) (Figuras 5.5 (b-e)) foi, novamente, a que produziu resultados menos ruidosos dentre as abordagens analisadas. Nota-se que, com o

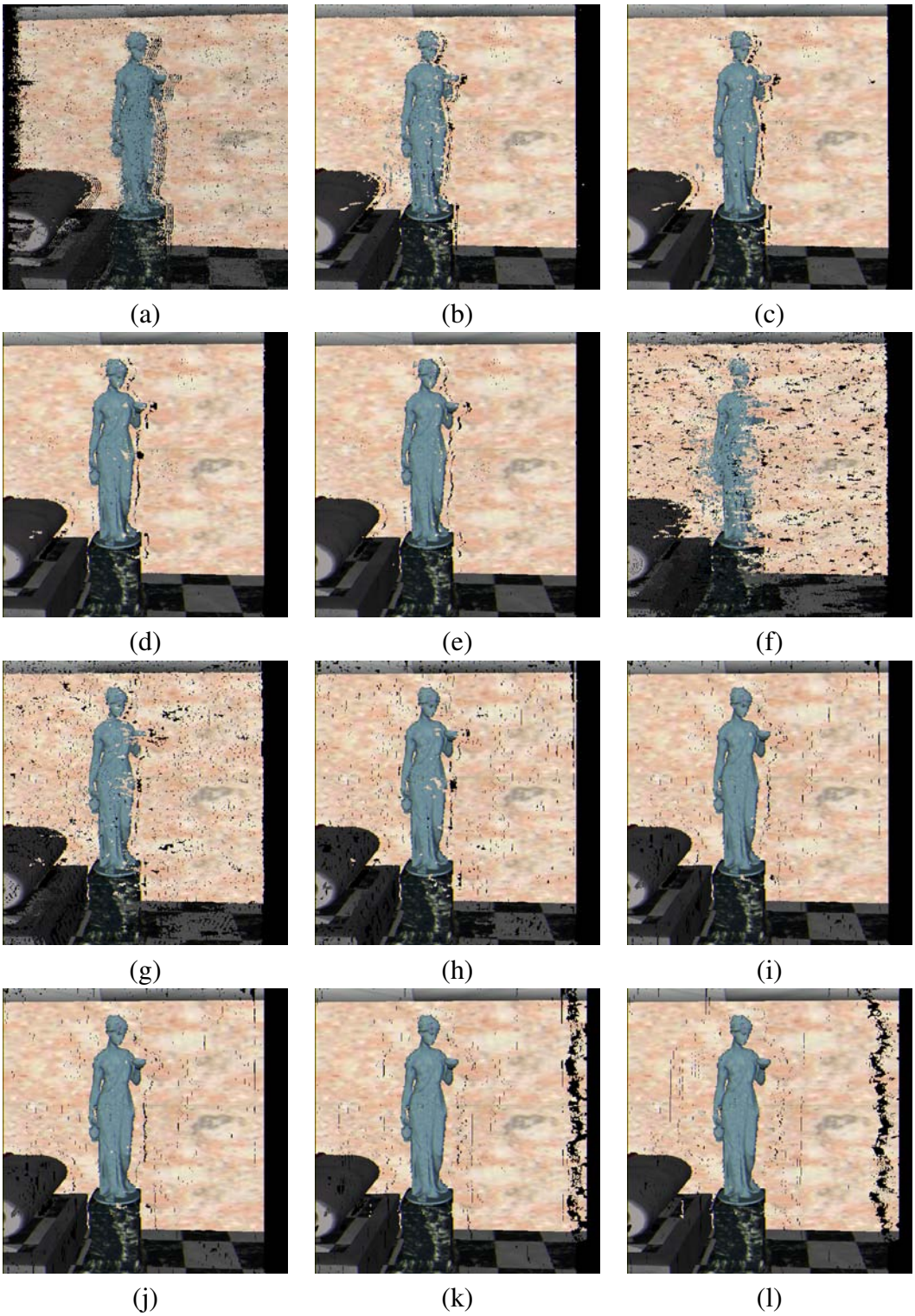

Figura 5.3: Pontos de vista alternativos calculados com mapas de disparidade obtidos usando algoritmos de extração de profundidade a partir de estéreo para uma cena com poucos objetos. (a) mostra a reconstrução usando uma implementação do algoritmo de Gong et al. (GONG; YANG, 2005) com apenas um passo para suavização do mapa, (b-e) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) com 1, 2, 3 e 4 níveis de mipmap, respectivamente, (f-l) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS, 2003) com 1, 2, 3, 4, 5, 6 e 7 níveis de mipmap respectivamente.

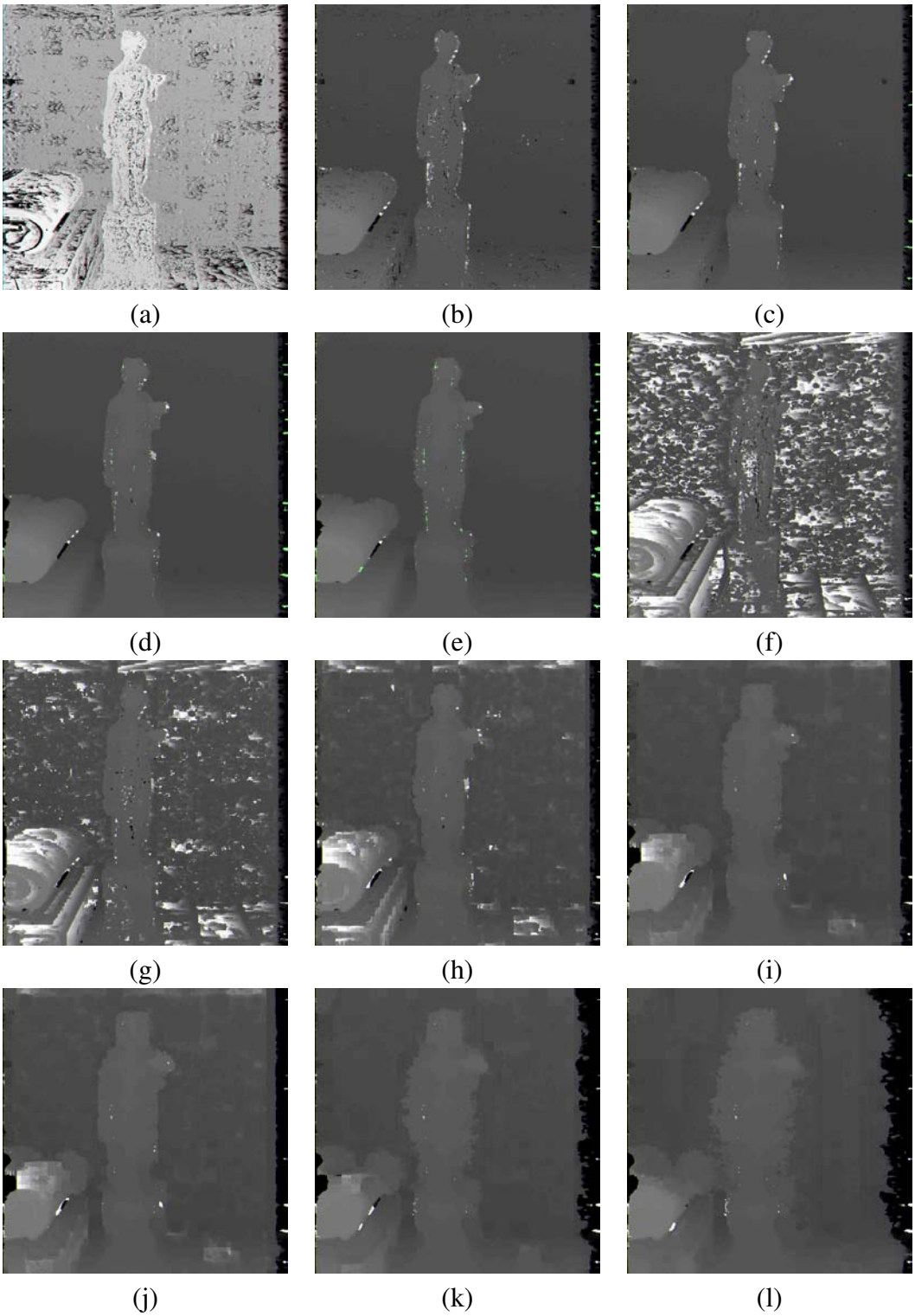

Figura 5.4: Mapas de disparidade extraídos a partir de algoritmos de extração de profundidade a partir de estéreo para uma cena com poucos objetos. (a) mostra a reconstrução usando uma implementação do algoritmo de Gong et al. (GONG; YANG, 2005) com apenas um passo para suavização do mapa, (b-e) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) com 1, 2, 3 e 4 níveis de mipmap respectivamente, (f-l) mostra a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS, 2003) com 1, 2, 3, 4, 5, 6 e 7 níveis de mipmap respectivamente

aumento do número de níveis de mipmap usados, o ruído diminui. Todavia, ocorre também uma perda de precisão nas bordas, bastante perceptível na estátua ao fundo quando o uso de apenas um nível de mipmap é confrontada com o uso de quatro níveis de mipmap (Figuras 5.5 (b, e)). A abordagem de Gong et al. (GONG; YANG, 2005) obteve resultados bastante ruidosos, o que prejudicou a percepção de pequenos objetos (Figura 5.5 (a)). A técnica de Yang et al. (YANG; POLLEFEYS, 2003), quando usados poucos níveis de mipmap (Figuras 5.6 (f-g)) também obteve resultados ruidosos em suas recontruções. Todavia, quando utilizados muitos níveis de mipmap, o ruído diminui, mas as bordas dos objetos ficam deformadas (Figuras 5.5 (j-l)). Visualmente, os melhores resultados foram obtidos com três e quatro níveis de mipmap usados, apesar de haver uma quantidade considerável de ruído e das bordas dos objetos não estarem precisas.

Analisando os resultados das reconstruções (Figura 5.5) em conjunto com os mapas de disparidade (Figuras 5.6), percebe-se que a quantidade de ruído nos mapas gerados pelas técnicas de Gong et al. (GONG; YANG, 2005) e de Yang et al. (YANG; POLLEFEYS, 2003) com poucos níveis de mipmap (Figuras 5.6 (a, f, g)) é maior na região da mesa (Figura 5.2 a), que é onde se concentram a maioria dos objetos. A abordagem de Yang et al. (YANG; POLLEFEYS, 2003) com poucos níveis de mipmap também obteve problemas nas paredes (Figuras 5.6 (e-g)). Percebe-se também que os objetos ficam deformados na abordagem de Yang. et al. (YANG; POLLEFEYS, 2003) quando vários níveis de mipmap são usados (Figuras 5.6 (j-l)). A abordagem de Yang et al. (YANG; POLLEFEYS; LI, 2004), apesar de ser a menos ruidosa, obteve problemas principalmente nas bordas das estátuas (Figuras 5.6 (b-e)).

As Tabelas 5.1.3 e 5.2 sumarizam a comparação entre os resultados mostrados nas Figuras 5.3, 5.4, 5.5 e 5.6. Foram analisados a taxa de quadros por segundo, a qualidade dos mapas de disparidade e a qualidade da reconstrução de novos pontos de vista. A qualidade do mapa de disparidade foi calculada pixel a pixel, fazendo a média das somas do quadrado das diferenças (SSD_{av}), como mostra a Equação 5.2, onde p_1 é a disparidade do pixel calculada com o algoritmo de extração de profundidade, p_z é a disparidade do pixel extraída do Z-buffer da cena, e np é o número de pixels em um dos mapas de disparidade. Conforme SSD_{av} aumenta, a qualidade do mapa diminui. O resultado das reconstruções também foi comparado pixel a pixel, comparando cada componente de cor. São considerados *pixels errôneos* aqueles que apresentarem diferença maior que 0,09 em uma das componentes de cor. Na Tabela 5.3, há a média dos resultados obtidos (calculados dessa mesma forma) para um pequeno vídeo de 25 quadros.

$$SSD_{av} = \sum \left(p_1 - p_z\right)^2 / np \tag{5.2}$$

Analisando as tabelas resultantes das comparações feitas, observou-se que o nível de imprecisão das técnicas de extração de profunidadde analisadas é bastante alto. Percebe-se que, apesar da quantidade de ruído, a abordagem de Gong. et al. foi a que obteve a reconstrução mais próxima da reconstrução feita com a informação exata da profundidade da cena. Nota-se, pelo resultado do SSD_{av} dessa técnica, que os pixels que tiveram disparidades incorretas associadas tiveram alta diferença (Tabela 5.1.3) para a cena com muitos objetos e, para a cena com poucos objetos (Tabela 5.2), uma diferença bem menor que a observada por outras técnicas.

A técnica de Yang. et al. (YANG; POLLEFEYS; LI, 2004), apesar de obter resultados visuais superiores às demais (linhas 2 até 5 das Tabelas 5.1.3, 5.2 e 5.3), apresentou altos

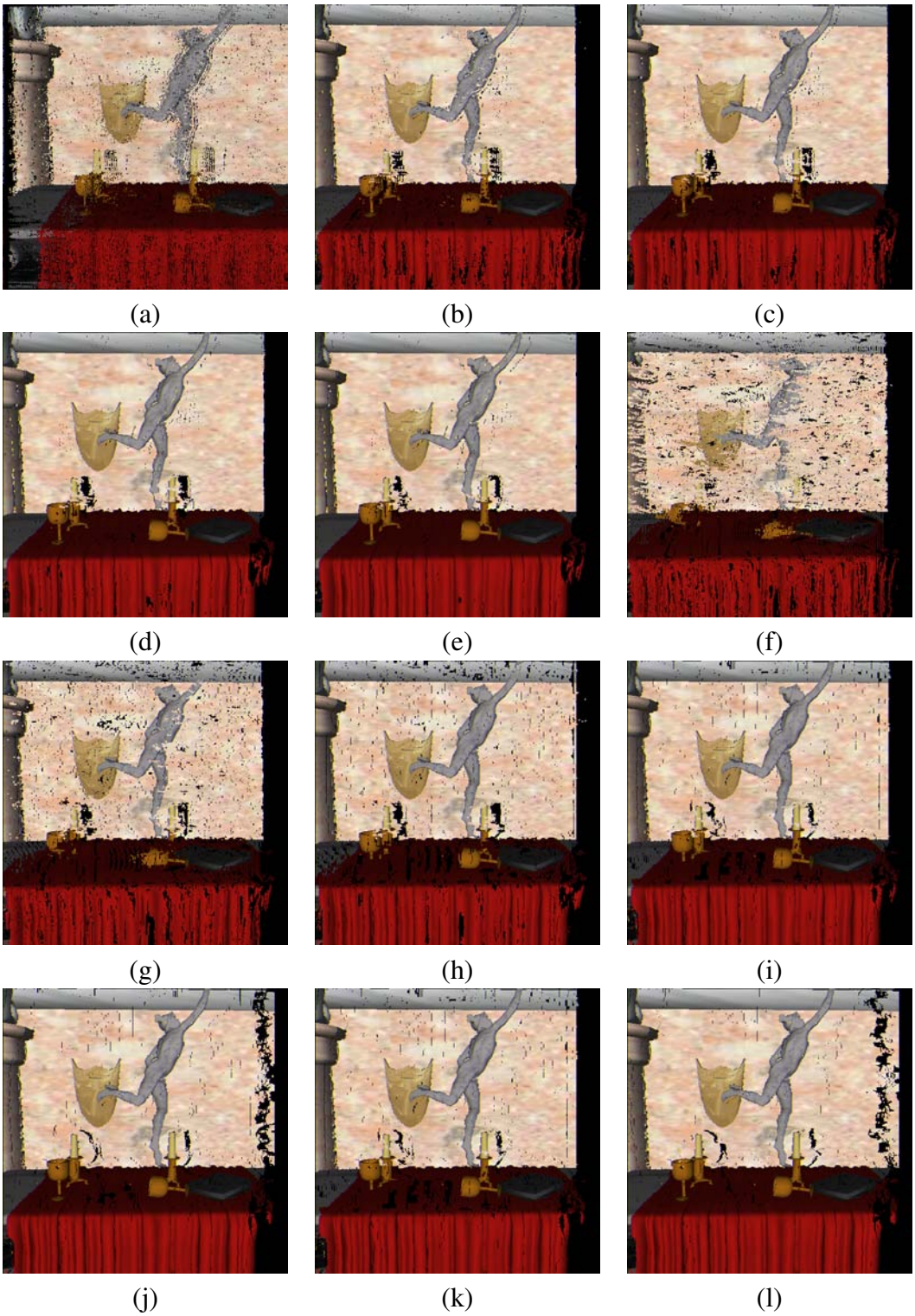

Figura 5.5: Pontos de vista alternativos calculados com mapas de disparidade obtidos usando algoritmos de extração de profundidade a partir de estéreo para uma cena com vários objetos. (a) mostra a reconstrução usando uma implementação do algoritmo de Gong et al. (GONG; YANG, 2005) com apenas um passo para suavização do mapa, (b-e) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) com 1, 2, 3 e 4 níveis de mipmap respectivamente, (f-l) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS, 2003) com 1, 2, 3, 4, 5, 6 e 7 níveis de mipmap respectivamente

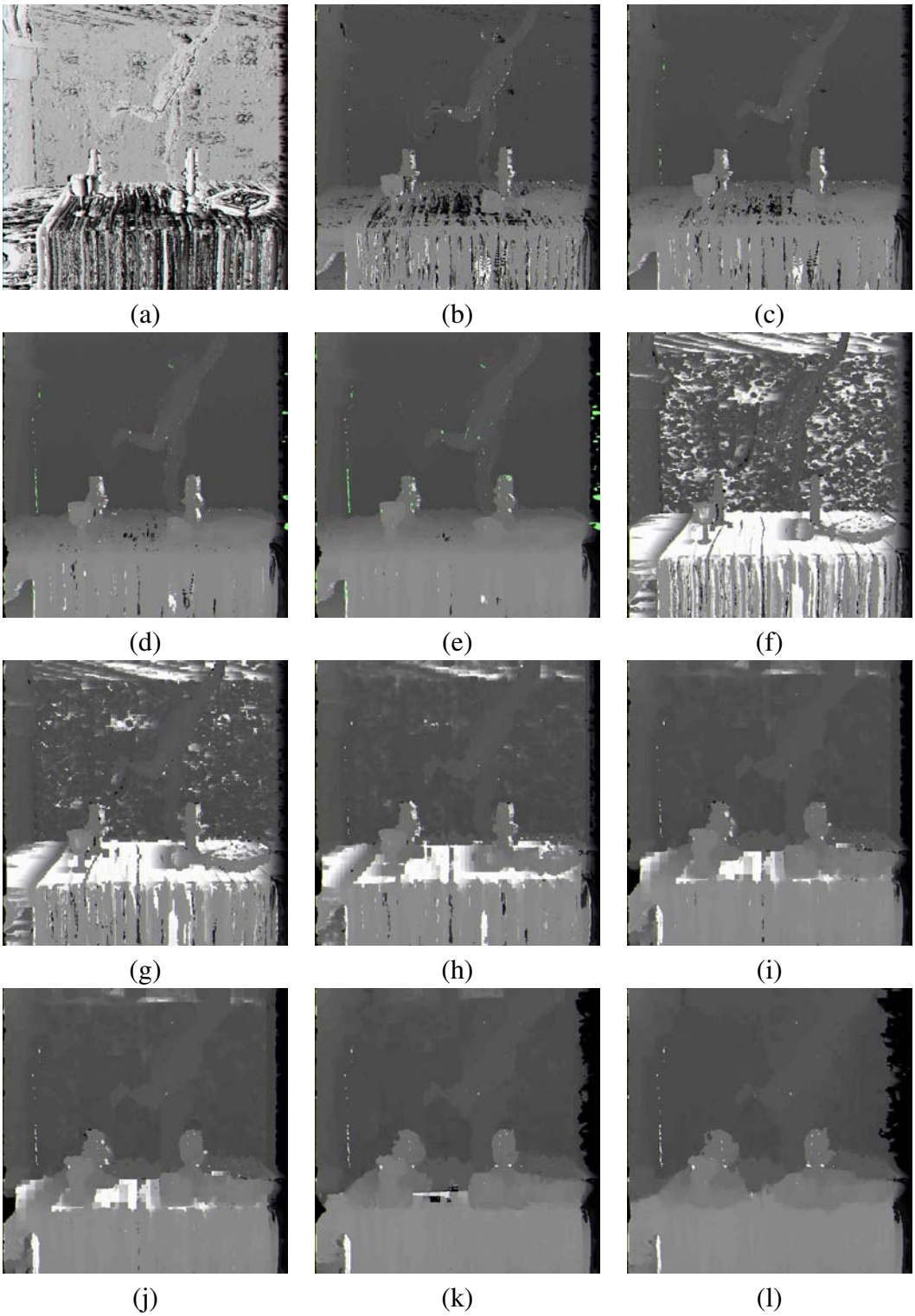

Figura 5.6: Mapas de disparidade extraídos a partir de algoritmos de extração de profundidade a partir de estéreo para uma cena com vários objetos. (a) mostra a reconstrução usando uma implementação do algoritmo de Gong et al. (GONG; YANG, 2005) com apenas um passo para suavização do mapa, (b-e) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) com 1, 2, 3 e 4 níveis de mipmap respectivamente, (f-l) mostram a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS, 2003) com 1, 2, 3, 4, 5, 6 e 7 níveis de mipmap respectivamente

Tabela 5.1: Taxa de quadros por segundo, qualidade do mapa (SSD_{av}) e porcentagem de pixels reconstruídos erroneamente para cenas com muitos objetos. Na tabela, (1) refere-se ao algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) e (2) refere-se ao algoritmo de Yang et al. (YANG; POLLEFEYS, 2003).

Algoritmo	Quadros por segundo	SSD_{av}	pixels errôneos
Gong et al.(GONG; YANG, 2005)	15,6	12.323,68	35,40%
(1) 4 mipmap	25,7	859,57	58,92%
(1) 3 mipmap	27,8	930,70	59,08 %
(1) 2 mipmap	30,2	1.095,07	59,48%
(1) 1 mipmap	32,4	1.449,65	59,71 %
(2) 7 mipmap	12,9	1.027,45	58,65 %
(2) 6 mipmap	13,4	1.074,35	59,20 %
(2) 5 mipmap	13,4	1.290,17	60,21 %
(2) 4 mipmap	14,5	1.283,96	60,22 %
(2) 3 mipmap	15,1	1.887,90	61,57 %
(2) 2 mipmap	15,7	3.069,53	62,24 %
(2) 1 mipmap	15,1	5.435,61	62,33%

Tabela 5.2: Taxa de quadros por segundo, qualidade do mapa (SSD_{av}) e porcentagem de pixels reconstruídos erroneamente para cena com poucos objetos. Na tabela, (1) refere-se ao algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) e (2) refere-se ao algoritmo de Yang et al. (YANG; POLLEFEYS, 2003).

Algoritmo	Quadros por segundo	SSD_{av}	pixels errôneos
Gong et al.(GONG; YANG, 2005)	15,6	14.690,03	24,90 %
(1) 4 mipmap	25,7	977,46	64,15 %
(1) 3 mipmap	27,8	973,43	64,23%
(1) 2 mipmap	30,2	977,95	64,44%
(1) 1 mipmap	32,4	1.016,41	64,76%
(2) 7 mipmap	12,9	1.184,15	63,72%
(2) 6 mipmap	13,4	1.182,70	64,00%
(2) 5 mipmap	13,4	1.268,51	64,75%
(2) 4 mipmap	14,5	1.248,28	64,75%
(2) 3 mipmap	15,1	1.602,19	65,81%
(2) 2 mipmap	15,7	2.919,87	67,71%
(2) 1 mipmap	15,1	6.493,87	69,74%

Tabela 5.3: Taxa de quadros por segundo, qualidade do mapa (SSD_{av}) e porcentagem de pixels reconstruídos erroneamente a partir de um video com 25 quadros. Na tabela, (1) refere-se ao algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) e (2) refere-se ao algoritmo de Yang et al. (YANG; POLLEFEYS, 2003).

Algoritmo	Quadros por segundo	SSD_{av}	pixels errôneos
Gong et al.(GONG; YANG, 2005)	15,6	1979,52	52,64%
(1) 4 mipmap	25,7	6244,7	62,95%
(1) 3 mipmap	27,8	6099,5	63,28%
(1) 2 mipmap	30,2	6114,5	63,36%
(1) 1 mipmap	32,4	6282,6	63,44%
(2) 7 mipmap	12,9	6820,1	64,02%
(2) 6 mipmap	13,4	6573,2	64,55%
(2) 5 mipmap	13,4	6236,5	65,13%
(2) 4 mipmap	14,5	6009,2	65,29%
(2) 3 mipmap	15,1	5709,1	65,84%
(2) 2 mipmap	15,7	5446,3	62,47%
(2) 1 mipmap	15,1	4972,5	59,40%

índices de erros. Isso ocorre porque os mapas de disparidade gerados por essa técnica são bastante suaves (Figuras 5.6 (j-l)), todavia, não são precisos. Isso fica evidente principalmente ao comparar visualmente os resultados das reconstruções com as técnicas de Yang et al. (YANG; POLLEFEYS; LI, 2004) e as reconstruções feitas a partir da informação de profundidade exata (Figuras 5.5 (b-e) e 5.2 (b, e)). Na cena com vários objetos, por exemplo, a estátua que aparece mais à direita na reconstrução com a informação de profundidade exata não se mostra nem no campo de visão da reconstrução feita através da técnica de Yang. et al. (YANG; POLLEFEYS; LI, 2004).

A Figura 5.8 mostra os mapas de disparidade extraídos com a técnica de Yang et al. (YANG; POLLEFEYS; LI, 2004) para um quadro de um vídeo de cena real. Esse vídeo foi obtito utilizando uma câmera estéreo pré-calibrada (*bumblebee* (PGR, 2005), Figura 5.7). Quatro quadros desse vídeo podem ser visualizados na Figura 5.9. Essa câmera pode ser utilizada, em conjunto com a arquitetura proposta, para a captura de vídeos estéreo. Com um equipamento como esse em conjunto com um algoritmo de extração de disparidade a partir de pares de imagens estéreo, é possível produzir um vídeo 3D natural.

Figura 5.7: Câmera estéreo pré-calibrada modelo Bumblebee produzida pela PointGrey Research (PGR, 2005)

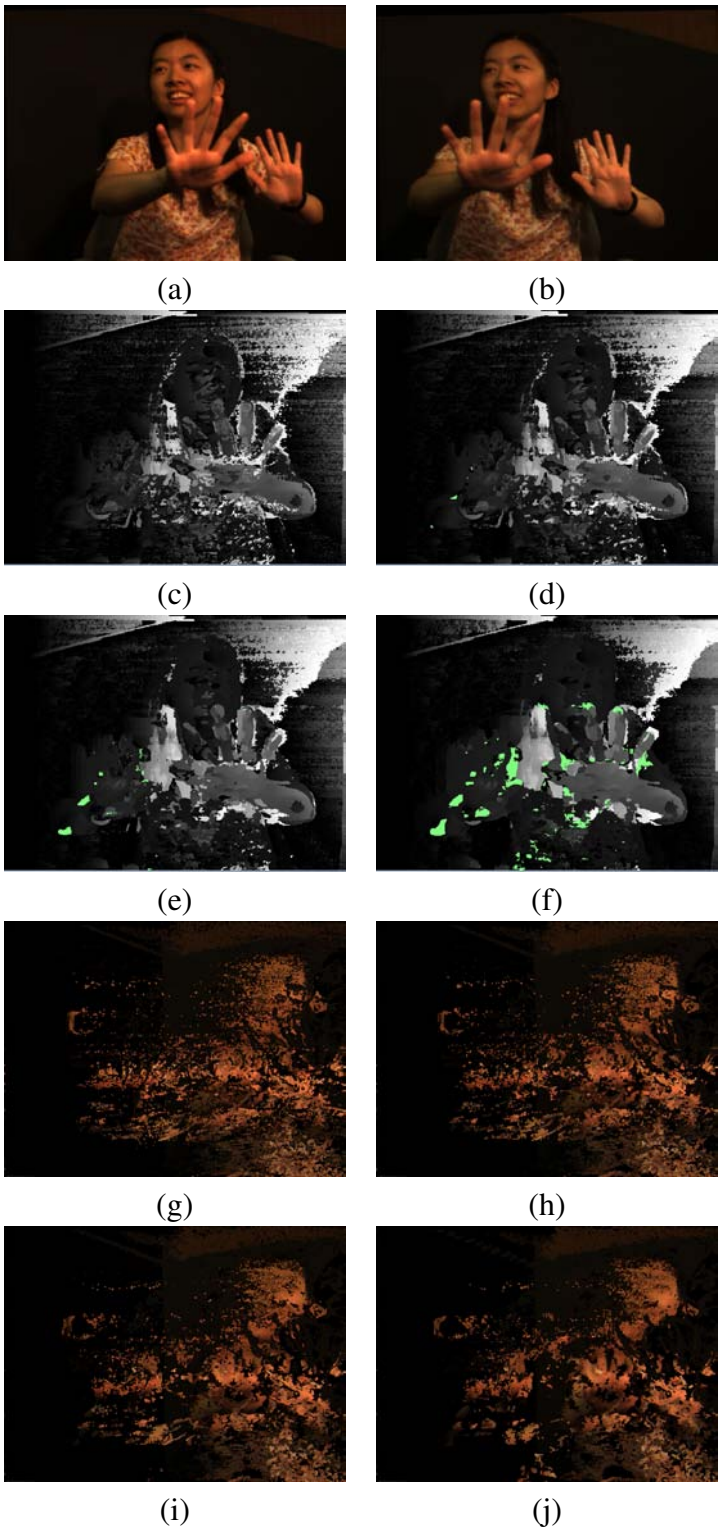

Figura 5.8: Mapas de disparidade extraídos a partir de algoritmos de extração de profundidade a partir de estéreo e reconstrução para uma cena Real. (a-b) mostra o par de imagens, (c-f) mostra os mapas de disparidade usando o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) com 1, 2, 3 e 4 níveis de mipmap usados respectivamente, (g-j) mostra a reconstrução usando o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004) com 1, 2, 3 e 4 níveis de mipmap usados respectivamente

Figura 5.9: Quatro quadros de um vídeo estereoscópico capturados com uma câmera mostrada na Figura 5.7. Cortesia de Ruygang Yang.

5.2 Discussão

Fica evidente ao se analisar as Tabelas 5.1.3, 5.2 e 5.3, que os algoritmos de extração de disparidade a partir de estéreo em tempo real deixam a desejar em qualidade. Os mapas de disparidade resultantes são ruidosos (Figuras 5.4 (a, f, g) e 5.6 (a, f, g)) ou "super-suavizados" (Figuras 5.4 (j-l) e 5.6 (j-l)), o que prejudica bastante a precisão do processo de reconstrução da cena sob um novo ponto de vista (Figuras 5.3 e 5.5).

O algoritmo de Gong et al. (GONG; YANG, 2005) foi o que teve maior índice de acerto em seu mapa de disparidade. Todavia, o mapa de disparidade resultante ainda foi bastante ruidoso e, por isso, a percepção tridimensional da cena fica sensivelmente prejudicada.

Por outro lado, o algoritmo de Yang et al. (YANG; POLLEFEYS; LI, 2004), quando utilizando quatro níveis de *mipmap*, obteve um resultado sem muitos ruídos e com uma taxa de quadros por segundo bastante alta. Apesar de seu resultado obter um número alto de pixels remapeados em lugares errados, a percepção 3D da cena não ficou totalmente prejudicada.

Conclui-se, então, que os atuais algoritmos de extração de disparidade a partir de imagens estereoscópicas em tempo real ainda não são adequados para a produção de vídeos 3D.

6 CONCLUSÃO

Esse trabalho apresentou uma arquitetura para captura e exibição de vídeos 3D que pode ser usada tanto para o entretenimento quanto para avaliar a qualidade de métodos de extração de profundidade de cenas reais.

No Capítulo 2 foram discutidas diversas formas de se produzir vídeos 3D. Podemos classificar essas abordagens em quatro grandes grupos: Vídeos com Profundidade, Sistemas Multi-Câmeras, Sistemas com Modelagem Completa dos Objetos e Sistemas Imersivos. A arquitetura proposta nessa dissertação encaixa-se no grupo dos "Vídeos com Profundidade" e se diferencia das demais abordagens por não utilizar qualquer estágio de pré-processamento, não utilizar em nenhum momento a geometria da cena, e por objetivar a geração de novos pontos de vista da cena.

No Capítulo 3 foram descritas técnicas de extração de profundidade de cenas naturais. As abordagens analisadas foram: a extração de profundidade a partir de estéreo, a utilização de luz estruturada e o uso de câmeras 3D. Ao final do capítulo, foram apresentadas as vantagens e desvantagens de cada uma das abordagens analisadas.

O Capítulo 4 apresentou a arquitetura para captura e exibição de vídeos 3D em tempo real. A arquitetura foi detalhada passo a passo e o conceito de *3D video warping* foi apresentado para a geração de novos pontos de vista da cena.

No Capítulo 5 foi feita uma análise de métodos para extração de disparidade em tempo real comparando a qualidade dos mapas gerados, a capacidade de reconstrução de novos pontos de vista a partir destes mapas, e a velocidade dos algoritmos.

Tendo em vista as análises feitas no decorrer do trabalho, conclui-se que tão logo haja uma forma de se extrair com precisão a informação de profundidade de uma cena, será possível, sem maiores dificuldades, produzir vídeos 3D utilizando técnicas de rendering baseados em imagens. Para que se possa transmitir vídeos 3D "ao vivo", é necessário que a informação de profundidade seja obtida em tempo real. Dessa forma, até mesmo eventos esportivos poderiam ser transmitidos, dando a possibilidade de, por exemplo, o telespectador escolher o ângulo de visualização de um lance polêmico em um jogo de futebol. Infelizmente, os resultados dos métodos existentes para extração de profundidade em tempo real ainda não possuem qualidade suficiente para que se produzam vídeos 3D com precisão satisfatória. Todavia, no caso de a extração da informação de profundidade da cena não ser feita em tempo real, o mapa de disparidades poderia ser calculado com maior precisão.

O avanço constante no desenvolvimento das câmeras 3D torna a idéia de se produzir vídeos 3D bastante promissora. Quem sabe, em não mais que dez anos, as câmeras atuais possuam precisão milimétrica e um alcance maior (de 20 a 30 metros). Utilizando aparelhos de *GPS* de alta precisão para obter a posição e orientação de várias câmeras 3D usadas para capturar uma determinada cena, seria possível utilizar uma composição de

imagens com profundidade capturadas para gerar os novos pontos de vista da cena sem as restrições atuais da presente técnica (*i.e.*, a ocorrência de regiões não amostradas que aparecem como buracos nas imagens finais (Figura 4.4)).

Além de servir para produzir vídeos 3D para entretenimento, a arquitetura proposta nessa dissertação mostrou-se uma eficiente forma de se comparar resultados de métodos de extração de profundidade. Com ela, é possível analisar a reconstrução de novos pontos de vista utilizando o resultado de diversos métodos de extração de profundidade.

Como foi visto na Seção 2.1, Fehn et al. (FEHN, 2003) propuseram um meio de incorporar informação de profundidade em vídeos MPEG-2. Com isso, além de se manter a compatibilidade entre vídeos com profundidade e os vídeos tradicionais (que contém apenas informação de cor), poucas alterações precisam ser feitas nos aparelhos atuais para que se possa processar a informação de profundidade. Uma vez recebida a informação de cor e o mapa de disparidade da cena, basta que os aparelhos receptores de sinais de televisão *(set top boxes)* implementem o algoritmo de *3D Video Warping*. Como essas funções são computacionalmente simples e independem da complexidade geométrica da cena, tal implementação é perfeitamente viável.

Futuramente, pretende-se integrar a arquitetura com um processo de compressão de vídeos para que os vídeos 3D possam ser transmitidos mais eficientemente. É necessário criar uma forma de sincronizar a transmissão de imagens e informações de profundidade, diferente da que foi descrita por Fehn et al. (FEHN, 2003). Outro ponto evidenciado com a realização desse trabalho é a necessidade de se desenvolver novos algoritmos para o cálculo de disparidade em tempo real capazes de produzir mapas mais precisos do que os obtidos atualmente.

REFERÊNCIAS

BAYAKOVSKI, Y.; LEVKOVICH-MASLYUK, L.; IGNATENKO, A.; KONUSHIN, A.; TIMASOV, D.; ZHIRKOV, A.; HAN, M.; PARK, I. K. Depth image-based representations for static and animated 3D objects. In: ICIP (3), 2002. **Anais...** [S.l.: s.n.], 2002. p.25–28.

BOYER, K. L.; KAK, A. C. Color-Encoded Structured Light for Rapid Active Ranging. In: IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 1987. **Anais...** [S.l.: s.n.], 1987. v.9, p.14–28.

CROW, F. Summed-area tables for texture mapping. In: SIGGRAPH, 1984. **Proceedings...** [S.l.: s.n.], 1984. p.207 – 212.

FEHN, C. A 3D-TV System Based On Video Plus Depth Information. In: ASILOMAR CONF. ON SIGNALS, SYSTEMS, AND COMPUTERS, 37., 2003, Pacific Grove, CA, USA. **Proceedings...** [S.l.: s.n.], 2003. p.1529–1533.

FEHN, C.; COOKE, E.; SCHREER, O.; KAUFF, P. 3D Analysis and Image-Based Rendering for Immersive TV Applications. In: INTERNATIONAL CONF. ON AUGMENTED VIRTUAL ENVIRONMENTS AND THREE-DIMENSIONAL IMAGING, 2001, Mykonos, Greece. **Proceedings...** [S.l.: s.n.], 2001. p.192–195.

FEHN, C.; KAUFF, P. Interactive Virtual View Video (IVVV) – The Bridge Between 3D-TV and Immersive TV. In: SPIE THREE-DIMENSIONAL TV, VIDEO AND DISPLAY I, 2002, Boston, MA, USA. **Proceedings...** [S.l.: s.n.], 2002. p.14–25.

FEHN, C.; KAUFF, P.; SCHREER, O.; SCHäFER, R. Interactive Virtual View Video for Immersive TV Applications. In: INTERNATIONAL BROADCAST CONF., 2001, Amsterdam, The Netherlands. **Proceedings...** [S.l.: s.n.], 2001. p.53–62.

GONG, M.; YANG, Y. Fast Stereo Matching Using Reliability-Based Dynamic Programming and Consistency Contraints. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2003, Nice, France. **Anais...** [S.l.: s.n.], 2003. v.1, p.610–617.

GONG, M.; YANG, Y.-H. Genetic-Based Stereo Algorithm and Disparity Map Evaluation. **International Journal of Computer Vision**, [S.l.], v.47, n.1-3, p.63–77, 2002.

GONG, M.; YANG, Y.-H. Near Real-time Reliable Stereo Matching Using Programmable Graphics Hardware. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2005, San Diego, United States. **Anais...** [S.l.: s.n.], 2005. p.924–931.

GROSS, M.; WURMLIN, S.; NAEF, M.; LAMBORAY, E.; SPAGNO, C.; KUNZ, A.; KOLLER-MEIER, E.; SVOBODA, T.; GOOL, L. V.; LANG, S.; STREHLKE, K.; MOERE, A. V.; STAADT, O. Blue-c: a spatially immersive display and 3d video portal for telepresence. **ACM Trans. Graph.**, [S.l.], v.22, n.3, p.819–827, 2003.

HARMAN, P. Home Based 3D Entertainment - An Overview. In: ICIP, 2000. **Anais...** [S.l.: s.n.], 2000. p.1–4.

HILLS, A. R. **Eye of the World**: john logie baird and television. Disponível em: <http://www.arts.uwaterloo.ca/FINE/juhde/hills961.htm>. Acesso em: nov 2004.

HONG, L.; CHEN, G. Segment-Based Stereo Matching Using Graph Cuts. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2004, Washington,USA. **Anais...** [S.l.: s.n.], 2004. v.1, p.74–81.

HORIMURA, T. **Comunicação de Massa**. Disponível em: <http://www.br.emb-japan.go.jp/portugues/cultura/downloads/comunicacao.htm>. Acesso em: set 2005.

IDDAN, G. J.; G., Y. 3D Imaging in the Studio (and Elsewhere ...). In: SPIE 4298: VIDEOMETRICS AND OPTICAL METHODS FOR 3D SHAPE MEASUREMENTS, 2001. **Proceedings...** [S.l.: s.n.], 2001. p.48–55.

IIZUKA, K.; KAWAKITA, M. Unconventional Imaging. In: OPTICS & PHOTONICS NEWS, 2002. **Anais...** [S.l.: s.n.], 2002. p.61.

ISO/IEC JTC 1/SC 29/WG 11 Joint Video Specification (ITU-T Rec. H.264 - ISO/IEC 14496-10 AVC). [S.l.]: JVT Document E146d34, 2002.

ISO/IEC JTC1/SC29/WG11 N4415: pdam of iso/iec 14496-1 / amd4. [S.l.]: JVT Document E146d34, 2001.

KAWAKITA, M.; KURITA, T.; KIKUCHI, H.; YAMANOUCHI, Y.; INOUE, S.; ; IIZUKA, K. HIGH-DEFINITION THREE-DIMENSION CAMERA: hdtv version of an axi-vision camera. In: NHK LABORATORIES, 2002. **Anais...** [S.l.: s.n.], 2002. p.479.

LEE, H. I. **Internet Transmission of Real-time MPEG Video**. 1998. BA Thesis — Monografia Mount Holyoke College.

MARK, W. R.; MCMILLAN, L.; BISHOP, G. Post-rendering 3D warping. In: SI3D, 1997. **Anais...** [S.l.: s.n.], 1997. p.7–16.

MATUSIK, W.; BUEHLER, C.; RASKAR, R.; GORTLER, S. J.; MCMILLAN, L. Image-Based Visual Hulls. **Proc. ACM SIGGRAPH 2000**, [S.l.], p.369–374, 2000.

MATUSIK, W.; PFISTER, H. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. **Proc. of ACM SIGGRAPH 2004**, [S.l.], v.23, n.3, p.814–824, 2004.

MCMILLAN, L. **An Image-based Approach to Three-Dimensional Computer Graphics**. 1997. PhD Dissertation — University of North Carolina at Chapel Hill.

MITCHELL, J. L.; PENNEBAKER, W. B.; FOGG, D. J. L. Chad e. **MPEG Video Compression Standard**. [S.l.]: Chapman & Hall, 1997.

MVFX. **Matrix Visual Effects**. Disponível em: <http://wwwmvfx.com> Acessado em nov 2004.

NEIDER, J.; DAVIS, T.; WOO, M. **OpenGL Programming Guide**. [S.l.]: Addison-Wesley, 1996.

PGR. **Bumblebee Technical Specifications**. Disponível em: <http://www.ptgrey.com/products/bumblebee/bumblebee.PDF>. Acesso em: jun 2005.

PRICE, M.; THOMAS, G. 3D Virtual Production and Delivery Using MPEG-4. **IBC'00 Conf. Publication**, [S.l.], p.616 – 621, 2000.

RITTERMANN, M.; SCHULDT, M. 3D Television Production Based on MPEG-4 Principles. **WSCG**, [S.l.], p.2–8, 2003.

SCHARSTEIN, D.; SZELISKI, R. **Middlebury College Stereo Vision Research Page**. "Disponível em: <http://cat.middlebury.edu/stereo/>. Acesso em: maio 2005.

SCHARSTEIN, D.; SZELISKI, R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. **International Journal of Computer Vision**, [S.l.], v.47, n.1-3, p.7–42, 2002.

SCHIRMACHER, H.; MING, L.; SEIDEL, H.-P. On-the-Fly Processing of Generalized Lumigraphs. **j-CGF**, [S.l.], v.20, n.3, setembro 2001.

SHADE, J.; GORTLER, S.; HE, L.; SZELISKI, R. Layered depth images. In: SIGGRAPH, 1998. **Proceedings...** [S.l.: s.n.], 1998. p.231–242.

SZELISKI, R. Prediction Error as a Quality Metric for Motion and Stereo. In: ICCV, 1999. **Anais...** [S.l.: s.n.], 1999. p.781–788.

SZELISKI, R.; ZABIH, R. An Experimental Comparison of Stereo Algorithms. In: WORKSHOP ON VISION ALGORITHMS, 1999. **Anais...** [S.l.: s.n.], 1999. p.1–19.

TRUCCO, E.; VERRI, A. **Introductory Techniques for 3-D Computer Vision**. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

VEKSLER, O. Fast Variable Window for Stereo Correspondence using Integral Images. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2003, Madison, United States. **Anais...** [S.l.: s.n.], 2003. v.1, p.565–570.

VEKSLER, O. Stereo Correspondence by Dynamic Programming on a Tree. In: IEEE CVPR, 2005. **Anais...** [S.l.: s.n.], 2005. p.384–390.

VIEIRA, M. B.; VELHO, L.; SÁ, A.; CARVALHO, P. Real-Time 3D Video. In: VISUAL PROC. SIGGRAPH, 2004. **Anais...** [S.l.: s.n.], 2004.

VIEIRA, M. B.; VELHO, L.; SÁ, A.; CARVALHO, P. **Video 4D Project**. Disponível em: <http://w3.impa.br/˜ mbvieira/video4d/>. Acesso em: nov 2005.

WELLE, D. **O Lazer**. Disponível em: <http://www.dw-world.de/dw/article/0,1564,1062641,00.html>. Acesso em: set, 2005.

WILLIAMS, L. Pyramidal Parametrics. In: ACM SIGGRAPH 1983, 1983, Detroit, Michigan, United States. **Proceedings...** [S.l.: s.n.], 1983. p.1 – 11.

WüRMLIN, S.; LAMBORAY, E.; STAADT, O. G.; GROSS, M. H. 3D Video Recorder. In: PG '02: PROC. PACIFIC CONF. ON COMPUTER GRAPHICS AND APPLICATIONS, 2002. **Anais...** IEEE Computer Society, 2002. p.325.

YANG, R.; POLLEFEYS, M. Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware. In: CVPR (1), 2003. **Anais...** [S.l.: s.n.], 2003. p.211–220.

YANG, R.; POLLEFEYS, M.; LI, S. Improved Real-Time Stereo on Commodity Graphics Hardware. In: WORKSHOP ON REAL TIME 3D SENSORS AND THEIR USE, 2004. **Anais...** [S.l.: s.n.], 2004.

YANG, R.; WELCH, G.; BISHOP, G. Real-Time Consensus-Based Scene Reconstruction Using Commodity Graphics Hardware. **Comput. Graph. Forum**, [S.l.], v.22, n.2, p.207–216, 2003.

ZHANG, L.; CURLESS, B.; SEITZ, S. M. Rapid Shape Acquisition Using Color Structured Light and Multi-pass Dynamic Programming. In: THE 1ST IEEE INTERNATIONAL SYMPOSIUM ON 3D DATA PROCESSING, VISUALIZATION, AND TRANSMISSION, 2002. **Anais...** [S.l.: s.n.], 2002. p.24–36.

ZHANG, L.; SNAVELY, N.; CURLESS, B.; SEITZ, S. M. Spacetime Faces: high-resolution capture for modeling and animation. In: ACM ANNUAL CONFERENCE ON COMPUTER GRAPHICS, 2004. **Anais...** [S.l.: s.n.], 2004. p.548–558.

ZHANG, S.; HUANG, P. High-resolution, Real-time 3D shape acquisition. In: IEEE COMPUTER VISION AND PATTERN RECOGNITION WORKSHOP (CVPRW'04), 2004. **Anais...** [S.l.: s.n.], 2004. p.28.

APÊNDICE A OBTENDO DISPARIDADE GENERALIZADA A PARTIR DE VALORES ARMAZENADOS NO Z-BUFFER

O cálculo do *3D Video Warping* faz uso da disparidade generalizada (MCMILLAN, 1997), que é dada pela Equação A.1, onde f é a distância focal da câmera, Z_c é o valor da coordenada Z do ponto X no sistema de referência da câmera (Figura 4.3).

$$\delta = f/Z_c \tag{A.1}$$

Para que seja possível usar o *Z-buffer* de uma cena sintética na equação do *3D Warping* (Equação 4.6), é necessário calcular a disparidade generalizada do pixel a partir do valor correspondente armazenado no Z-buffer. Considere a expressão que calcula o valor armazenado no Z-buffer (Z_b para um dado pixel (Equação A.2)), quando utilizando OpenGL (NEIDER; DAVIS; WOO, 1996). Nesta equação, Z_c é o valor da coordenada Z do ponto em 3D expressa no sistema de referência da câmera, n e f são os valores dos planos de recorte *near* e *far* respectivamente.

$$Z_b = 0.5 + 0.5\left(\frac{f+n}{f-n} + \frac{1}{Z_c}\frac{2fn}{f-n}\right) \tag{A.2}$$

Isolando Z_c na Equação A.2 e substituindo esse valor na Equação A.1, é possível obter a disparidade generalizada para cada ponto visível da cena. As Equações A.3 e A.4 mostram este processo passo a passo.

$$
\begin{aligned}
Z_b &= 0.5 + 0.5\left(\frac{f+n}{f-n} + \frac{1}{Z_c}\frac{2fn}{f-n}\right) \\
Z_c &= \frac{(2fn)}{[2\,(Z_b - 0.5)\,(f-n) - (f-n)]} \\
Z_c &= \frac{(2fn)}{2\,[Z_b\,(f-n) - f]} \\
Z_c &= \frac{(fn)}{[Z_b\,(f-n) - f]}
\end{aligned} \tag{A.3}
$$

Substituindo o valor de Z_c encontrado em A.3 multiplicado por -1 (pois o OpenGL utiliza sistema de coordenadas "mão direita") na Equação A.1, tem-se:

$$
\begin{aligned}
\delta &= \frac{n}{-\frac{(fn)}{[Z_b(f-n)-f]}} \\
\delta &= \frac{Z_b\,(f-n) - f}{-f} \\
\delta &= -Z_b + 1 + Z_b\,(n/f)
\end{aligned} \tag{A.4}
$$