LUCAS PADOVANI TRIAS

Janela 3D: uma Ferramenta de Telecomunicação Visual Sensível ao
Ponto de Vista do Usuário

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Engenharia

2009
São Paulo

LUCAS PADOVANI TRIAS

Janela 3D: uma Ferramenta de Telecomunicação Visual Sensível ao
Ponto de Vista do Usuário

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Engenharia

Área de Concentração:
Engenharia Elétrica

Orientador:
Prof. Livre-Docente Romero Tori

2009
São Paulo

DEDICATÓRIA

*Dedico este trabalho a meus pais,
colegas, professores e a todos que
me incentivaram e apoiaram na busca
do conhecimento desde os mais tenros
momentos.*

AGRADECIMENTOS

Quero agradecer a todos aqueles que de alguma forma colaboraram para o desenvolvimento deste trabalho, de maneira direta ou indireta. Agradeço aos meus colegas de laboratório pelas incontáveis conversas que muitas vezes terminavam com o esclarecimento de idéias a respeito deste projeto ou, ao menos, serviam para divertir e tornar menos cansativo o seu desenvolvimento. Sou grato também a todos os funcionários da EPUSP que, muitas vezes sem saber, trabalharam para propiciar melhores condições para a pesquisa e o pesquisador e a meu orientador.

Agradeço aos amigos que por muitas vezes me apoiaram ao longo desta jornada seja com uma participação direta, ou apenas me ouvindo quando eu precisava falar. Sei que deixo de agradecer ainda a uma infinidade de pessoas que tornaram possível esta realização, cada qual à sua maneira, e me desculpo por não relacionar uma a uma neste texto, mas saibam que sou consciente da gratidão que tenho a vocês.

RESUMO

Sistemas de teleconferência e telepresença são ferramentas de comunicação cada vez mais comuns. Partindo da existência de um canal de comunicação de alta capacidade, busca-se permitir visualização tridimensional realista, sensível ao ponto de vista do usuário e que mantenha a estrutura física da cena sem conhecimento prévio de sua estrutura, por meio de câmeras estéreo. A partir de pares de imagens temporalmente coerentes são sintetizadas visões intermediárias da cena alvo, de modo que um usuário rastreado tenha a ilusão de estar vendo a cena real através de uma janela. O sistema implementado baseia-se em dispositivos que provejam melhor qualidade visual no componente de geração de imagens intermediárias e procura maximizar a facilidade de uso na parte de visualização, empregando dispositivos domésticos e dispensando procedimentos complexos de calibração.

Palavras-chave: telepresença, teleconferência, geometria epipolar, visão estéreo, telecomunicação imersiva

ABSTRACT

Teleconference and telepresence systems are increasingly common. Based on the existence of a high capacity communication channel, it is aimed to provide user's point of view sensible realistic 3D visualization without physical distortion or any prior knowledge about the structure of the scene, using stereo cameras. Pairs of temporally coherent images are used to generate intermediary view of the target scene so that a tracked user will have the illusion of posing in front of the real scene. The implemented system is based on high visual quality components on the side of views generation and tries to maximize ease of use in the visualization part by using commodity components and being free of complex calibration procedures.

keywords: telepresence, teleconference, epipolar geometry, stereo vision, immersive telecommunication

LISTA DE FIGURAS

1.1	Conceito da Janela 3D	5
2.1	Modelo conceitual de câmera de orifício	7
2.2	Modelo esquemático descritivo dos conceitos da geometria epipolar envolvidos em uma montagem genérica de câmeras, adaptada de Heyden e Pollefeys (5) .	11
2.3	<i>Continuum</i> da realidade misturada, adaptada de Milgram et al. (8)	14
3.1	Exemplo de distorção causada pelo processo de <i>image morphing</i> . Extraída de Seitz (4)	16
3.2	Comparação entre <i>Image Morphing</i> e <i>View Morphing</i> , adaptada de Seitz (4) . .	16
3.3	Exemplo de diferentes quantidades de <i>pixels</i> para representar os mesmos objetos em visões distintas e de monotonicidade, Seitz (4)	20
3.4	Formulação matemática da restrição de monotonicidade	21
3.5	Configuração singular de câmeras, adaptada de Seitz (4)	23
3.6	Mapeamento cilíndrico, adaptada de Timm (11)	30
3.7	Abordagens de múltipla resolução, adaptada de Timm (11)	31
3.8	Aproximação de elementos reais	33
4.1	Visão geral da arquitetura do sistema	37
4.2	Conceito de rastreamento do usuário	39
5.1	Resultados da prova de conceito com caixa	44
5.2	Resultados da prova de conceito com pilha de toalhas	44
5.3	Resultados da prova de conceito com planta	45
5.4	Efeito da distribuição dos pontos correspondentes	45
5.5	Efeito da concentração dos pontos correspondentes	46
5.6	Resultados obtidos com 21 pontos correspondentes	47
5.7	Comparação da geração de visões	49
5.8	Amostra de resultado obtido com três câmeras abrangendo 30°	49
5.9	Amostra de resultado obtido com doze câmeras abrangendo 55°	50
5.10	Amostra de resultado obtido com três câmeras abrangendo 30°	50
5.11	Renderização da cena vista do ponto 0.45	51

5.12	Diferença de cor entre as câmeras	53
5.13	Efeito da distorção radiométrica	54
5.14	Marcador colorido posicionado ao redor da lente de uma câmera	56
5.15	Exemplos de marcadores fiduciais, ARToolkit (38)	58
5.16	Exemplos de características Haar-like, adaptada de Viola e Jones (48)	60
5.17	Variação da detecção de Z Haar-like	61
5.18	Comparação do tempo gasto com e sem filtragem passa-baixas	62
5.19	Comparação da variação da detecção em X(5.19a), Y(5.19b) e Z(5.19c) com e sem filtro passa-baixas	62
5.20	Comparação da variação da detecção em X(5.20a), Y(5.20b) e Z(5.20c) entre as diferentes soluções testadas	63
5.21	Comparação do tempo gasto por diferente métodos de detecção	64
6.1	Montagem do Sistema de Visualização	66
6.2	Câmera usada no Sistema de Geração	66
6.3	Amostras para calibração de câmeras	67
6.4	Amostras para calibração de câmeras	67
6.5	Diagrama de Classes - parte 1	70
6.6	Diagrama de Classes - parte 2	71
6.7	Sequência de tarefas de um ciclo de atualização do Sistema	71
7.1	Efeito da interpolação de imagens com tons diferentes para objetos iguais	75
7.2	Efeito da interpolação de imagens com tons diferentes para objetos iguais, com correção da distorção radiométrica	75
7.3	Resultados de diferentes procedimentos de adequação das imagens à técnica empregada	76

LISTA DE TABELAS

5.1	Comparação de sistemas de rastreamento	63
-----	--	----

LISTA DE ABREVIATURAS E SIGLAS

- 2.5D: duas dimensões e meia
- 3D: Tridimensional
- AMD: *Advanced Micro Devices*
- AVTC: Arquitetura para Teleconferência Baseada na Integração de Vídeo Avatar Estereoscópico em Ambiente Virtual Tridimensional
- AVMix: Sistema de Vídeo-Avatar com Detecção de Colisão para Realidade Aumentada e Jogos
- CCD: *Charge-Coupled Device*
- CMOS: *Complementary Metal-Oxide-Semiconductor*
- DDR2: *Double Data Rate 2*
- FAPESP: Fundação de Amparo à Pesquisa do Estado de São Paulo
- GB: *Gigabyte*
- LCD: *Liquid Crystal Display*
- MIP-MAP: *Multum In Parvo Map*
- NCP: *Normalized Cylindrical Projection*
- RAM: *Random Access Memory*
- RIP-MAP: *Rectangular MIP-MAP*
- TIDIA: Tecnologia da Informação para o Desenvolvimento da Internet Avançada
- TIDIA-Ae: Tecnologia da Informação para o Desenvolvimento da Internet Avançada - Aprendizado Eletrônico
- UML: *Unified Modeling Language*
- XMIP-MAP: *Extended MIP-MAP*
- XMIP-NCP: *Extended MIP-MAP - NCP*

SUMÁRIO

1	Introdução	4
1.1	Objetivos	5
1.2	Organização do Texto	6
2	Técnicas e Conceitos Essenciais	7
2.1	Modelo de Câmera de Orifício	7
2.1.1	Definições	8
2.2	Visão Estéreo e Geometria Epipolar	8
2.2.1	Definições	9
2.3	Interpolação de Visões	11
2.4	Rastreamento	12
2.4.1	Marcadores Coloridos	12
2.4.2	Marcadores Fiduciais	13
2.4.3	Classificadores de Características Haar-like	13
2.5	Mistura do Mundo Real com o Virtual	14
2.6	Considerações Finais	14
3	Trabalhos Relacionados	15
3.1	Image-Based Transformation of Viewpoint and Scene Appearance	15
3.1.1	Image Morphing e View Morphing	15
3.1.2	O Algoritmo	17
3.1.3	Visões Paralelas	17
3.1.4	Visões Não Paralelas com Configuração Conhecida	18
3.1.5	Visões Quaisquer	23
3.1.6	Relação com o Projeto	24
3.2	AVTC	24
3.2.1	Aparência Visual do Avatar	24
3.2.2	Movimentação	25
3.2.3	Interação	25
3.2.4	Comunicação Sonora	26

3.2.5	Relação com o Projeto	26
3.3	AVMix	26
3.3.1	Mapa de Profundidade	27
3.3.2	Detecção de Colisão Espacial	28
3.3.3	Relação com o Projeto	28
3.4	Real Time View Morphing of Video Streams	29
3.4.1	Otimizações	29
3.4.2	Relação com o Projeto	32
3.5	Microfacet Billboarding	32
3.5.1	Relação com o Projeto	33
3.6	Considerações Finais	33
4	O Sistema Janela 3D	35
4.1	Conceito	35
4.2	Requisitos	36
4.3	Arquitetura	37
4.3.1	Subsistema de Visualização	38
4.3.2	Subsistema de Geração	39
4.3.3	Subsistema de Comunicação	40
4.4	Extensões e Integração	41
4.5	Considerações Finais	41
5	Experimentos Práticos	42
5.1	Estudo de Viabilidade do <i>View Morphing</i>	42
5.1.1	Proposta	42
5.1.2	Justificativa	43
5.1.3	Base Teórica	43
5.1.4	Resultados	43
5.1.5	Discussão	46
5.2	Uso de Maior Quantidade de Câmeras	47
5.2.1	Proposta	47
5.2.2	Justificativa	47
5.2.3	Base Teórica	47
5.2.4	Resultados	48
5.2.5	Discussão	51
5.3	Interpolação de Visões em Vídeo	51
5.3.1	Proposta	51
5.3.2	Justificativa	52
5.3.3	Base Teórica	52

5.3.4	Resultados	52
5.3.5	Discussão	53
5.4	Rastreamento de Cabeça	54
5.4.1	Marcadores Coloridos	54
5.4.2	Marcadores Fiduciais	56
5.4.3	Classificadores Características Haar-like	58
5.5	Considerações Finais	62
6	Implementação	65
6.1	Infra-Estrutura Física	65
6.1.1	Sistema de Visualização	65
6.1.2	Sistema de Geração	65
6.2	Software Relativo à Infra-Estrutura Física	66
6.2.1	Retificação de Imagens	66
6.2.2	Calibração de Convergência	67
6.2.3	Correção de Cor	68
6.2.4	Obtenção das Matrizes Fundamentais	69
6.2.5	Sistema de Comunicação	69
6.3	Artefatos da Engenharia de <i>Software</i>	69
6.3.1	Diagrama de Classes	70
6.3.2	Diagrama de Sequência	71
6.4	Considerações Finais	72
7	Resultados	73
7.1	Macro Objetivos	73
7.2	Sistema Visualização	73
7.3	Sistema Geração	74
7.4	Sistema Comunicação	77
7.5	Desenvolvimento do Protótipo	77
7.6	Considerações Finais	77
8	Conclusão e Trabalhos Futuros	78
8.1	Trabalhos Futuros	79
8.2	Agradecimentos	79

1 Introdução

É cada vez maior a busca por sensação de imersão em aplicações interativas. Gradualmente sistemas baseados em comunicação impessoal passam a permitir o uso de avatares ou a apresentação de vídeos, como, por exemplo, programas usados para trocar mensagens instantâneas. Ambientes tridimensionais, recursos de som espacial, inserção de avatares personalizáveis do usuário em comunidades virtuais, é crescente a busca por permitir que as pessoas se sintam inseridas em ambientes remotos, virtuais ou não.

Os tradicionais sistemas de teleconferência baseados em televisor e câmera, vêm sendo substituídos por instalações que permitem aos participantes de uma reunião olhar diretamente para seus colegas que estão em um ambiente remoto como se estivessem todos na mesma sala. No cinema são empregados cada vez mais canais de áudio e técnicas especiais de processamento que fazem com que os espectadores tenham a ilusão de estar dentro da cena, ainda que a estejam vendo uma projeção. Telas cada vez maiores, óculos especiais que permitem ilusão de profundidade nas imagens e efeitos de câmera como as tomadas *bullet time* (em que se pode circundar a cena congelada) popularizadas no filme *The Matrix* também estão se tornando usuais com o passar do tempo.

As empresas cada vez mais integram soluções tecnológicas a seus processos. O uso de cavernas digitais e outros ambientes imersivos no projeto de estruturas mecânicas, como carros, vem sendo adotado como meio de projeto colaborativo e interativo criando possibilidades que os projetistas sequer poderiam imaginar em um ambiente tradicional, e permitindo que desde cedo o projeto seja testado, ainda que virtualmente, com seres humanos.

Neste trabalho descreve-se a concepção e a implementação da Janela 3D, um sistema que permite a comunicação visual realista de usuários em diferentes ambientes. Como alternativa às atuais ferramentas de telepresença, a Janela 3D permite que o usuário tenha visualização tridimensional da cena remota dentro de determinados limites. Ou seja, quando o observador se move a imagem remota mostrada em sua tela se adapta à sua nova posição dando a ilusão de que este observa a cena através de uma janela. Da mesma maneira que no filme *The Matrix* é possível se circundar a cena que é exibida na tela. Entretanto, na Janela 3D isso acontece em tempo-real, é o observador é quem define qual o ângulo que deseja observar e quem ele enxerga na tela é uma pessoa de seu convívio, tornando a comunicação muito mais natural e imersiva. A Figura 1.1 ilustra este efeito.

O desenvolvimento desta pesquisa é associado ao Programa TIDIA-Ae (Tecnologia da Informação para o Desenvolvimento da Internet Avançada - Aprendizado Eletrônico), um dos

Figura 1.1 – Conceito da Janela 3D

componentes do Programa TIDIA(Tecnologia da Informação para o Desenvolvimento da Internet Avançada) da FAPESP(Fundação de Amparo à Pesquisa do Estado de São Paulo), que visa incentivar a pesquisa científico-tecnológica em projetos que envolvam redes de alta velocidade. O Programa é composto por três projetos: TIDIA-Ae, destinado a pesquisa de tecnologias para educação a distância; Incubadora Virtual de Conteúdos, que dispõe um ambiente via *Web* para criação colaborativa de conteúdos; e o KyaTera, que é focado no desenvolvimento experimental de tecnologias de internet avançada. Mais especificamente este projeto enquadra-se dentro do projeto Vídeo Avatar, de responsabilidade do Interlab (1), dentro do TIDIA-Ae.

Considerando a participação deste trabalho num projeto maior, cujo desenvolvimento de tecnologias e infra-estrutura básica de comunicação de alta velocidade já está atribuído a um grupo de pesquisa, define-se que pesquisas deste tipo encontram-se fora do escopo da Janela 3D e não serão abordadas nesta dissertação. Serão, sim, abordadas apenas as questões relevantes à comunicação em nível de aplicação dos componentes do sistema.

1.1 Objetivos

Este trabalho busca desenvolver um sistema que permita telecomunicação visual com qualidade gráfica em nível de televisão e coerência de projeção em relação a mudanças no ponto de vista do usuário. Pontualmente definidos, os objetivos deste projeto são:

- Desenvolver um sistema de rastreamento de usuário que seja de uso fácil, permita movimentação livre dentro de determinados limites e minimize o uso de equipamentos acoplados ao observador;
- Criar um sistema que gere imagens sintéticas realistas a partir de duas ou mais imagens obtidas em vídeos de uma cena real, em taxas interativas;
- Empregar protocolos de rede conhecidos de maneira a atender às necessidades de comunicação entre os sistemas citados;

- Desenvolver um protótipo que permita provar o conceito e realizar testes que subsidiem futuros projetos nesta linha.

1.2 Organização do Texto

O Capítulo 2 apresenta uma revisão conceitual de técnicas e conceitos importantes para a compreensão do projeto. Os principais trabalhos relacionados são discutidos no Capítulo 3. Uma formalização da especificação do sistema desenvolvido é feita no Capítulo 4. Os experimentos realizados e a implementação do protótipo são apresentados nos Capítulos 5 e 6. Os resultados obtidos são mostrados no Capítulo 7 e a conclusão do trabalho se dá no Capítulo 8.

2 Técnicas e Conceitos Essenciais

Nesta Seção são revisados conceitos cuja clareza é de vital importância para a compreensão do escopo e do funcionamento da Janela 3D. Para facilitar o entendimento eles são descritos individualmente e a relação entre eles e com o projeto é apresentada ao longo do texto.

2.1 Modelo de Câmera de Orifício

O modelo de câmera de orifício é um modelo matemático que procura reproduzir o comportamento da luz quando esta é captada por uma câmera real que emprega um pequeno orifício para direcioná-la a um anteparo fotossensível. Ele mapeia pontos da cena real tridimensional para coordenadas bidimensionais do plano retinal da câmera, através de uma transformação projetiva. Os raios de luz passam pelo centro de projeção da câmera (o orifício) e atingem o anteparo, onde fica o filme. A Figura 2.1 mostra as relações geométricas deste modelo. A seguir são dadas algumas definições de conceitos relevantes deste modelo que serão usadas ao longo deste trabalho, baseadas em Faugeras (2).

Figura 2.1 – Modelo conceitual de câmera de orifício

2.1.1 Definições

Plano de Projeção

Também chamado de *plano retinal*, é o plano onde ocorre a formação da imagem nas câmeras. No caso das câmeras digitais é o sensor de imagem, usualmente *CCD* (*Charge-Coupled Device*) ou *CMOS* (*Complementary Metal-Oxide-Semiconductor*) e em câmeras fotográficas analógicas, o anteparo fotossensível. Na Figura 2.1 o plano de projeção da câmera é denotado pela letra γ .

Centro Óptico

Também chamado de *centro de projeção* da câmera, é o ponto onde os raios de luz passam antes de atingir o plano retinal. Em uma câmera de orifício real o centro óptico é o próprio orifício. Na Figura 2.1 o centro óptico é denotado pela letra C .

Eixo Óptico

É a reta que passa pelo *centro óptico* C e é perpendicular ao *plano retinal* γ .

Distância Focal

É a distância entre o *plano retinal* γ e o *centro óptico* C , medida paralelamente ao *eixo óptico*. Na Figura 2.1 é denotada pela letra f .

2.2 Visão Estéreo e Geometria Epipolar

A visão estereo é aquela que se baseia em duas visadas de uma mesma cena, como os olhos humanos ou um par de câmeras. A partir de duas imagens é possível reconstruir a posição tridimensional de um ponto qualquer da cena com uma montagem em que se saiba as orientações e posições relativas das câmeras. Para isso uma das etapas consiste em identificar correspondências entre os elementos das duas imagens, ou seja, identificar a posição em que um elemento real se encontra nas duas imagens. Esta busca também é empregada, com

finalidade distinta, no processo de *view morphing*, descrito na Seção 3.1, e tem grande custo computacional. A grande quantidade de informação a ser processada faz necessária a adoção de uma geometria especial chamada geometria epipolar, para modelar o problema e torná-lo mais manejável.

A geometria epipolar é definida por Hartley e Zisserman (3) como a geometria projetiva intrínseca a duas visões, independente da estrutura da cena, e relacionada apenas aos parâmetros internos(ou intrínsecos) e externos(ou extrínsecos) das câmeras. Em termos práticos, ela constitui uma ferramenta matemática poderosa que, através de seus modelos e relações, permite analisar características da cena-alvo. A compreensão destes é de vital importância para o entendimento do processo de geração de visões intermediárias sintéticas.

2.2.1 Definições

Visão ou Câmera

É o elemento que faz o papel de observador da cena. São definidas por Seitz (4) como o par ordenado $V = \langle \mathcal{I}, \mathbf{\Pi} \rangle$, onde $\mathcal{I}$ é uma imagem e $\mathbf{\Pi}$ é a matriz de projeção perspectiva 3×4 associada a ela, dada por $\mathbf{\Pi} = [\mathbf{H} | -\mathbf{HC}]^1$, onde $\mathbf{H}$ é a matriz 3×3 que especifica a transformação linear do plano de imagem da câmera em relação à origem do sistema de coordenadas do mundo e $\mathbf{C}$ especifica o centro óptico da câmera. No contexto da geometria epipolar há, ao menos, duas câmeras em uma dada estrutura de visualização.

Linha Base

Em um par de câmeras, é a reta que une seus *centros ópticos*. Na Figura 2.2 é a linha $\overline{C_1 C_2}$.

Epipolo

O epipólo $E_{e,d}$ ² é definido como projeção do centro óptico da câmera e na imagem d por Heyden e Pollefeys (5). Uma definição alternativa é a de que os epipólos são os pontos de

¹A notação $A|B$ denota a concatenação das matrizes A e B , ou seja, a disposição da matriz B ao lado da matriz A formando uma nova matriz cuja quantidade de colunas é igual à soma da quantidade de colunas de A e B

²A notação $E_{a,b}$ significa “epipólo de a em b ”, a projeção do centro óptico da câmera a no plano de retinal da câmera b

intersecção entre a linha base e os planos de retinais. Na Figura 2.2 o epipólo esquerdo é chamado $e_{2,1}$ e o direito de $e_{1,2}$.

Plano Epipolar

É o plano definido pelos dois epipólos e um ponto qualquer da cena real. Na Figura 2.2 $e_{1,2}$, $e_{2,1}$ e M definem o plano epipolar Π . Tomando-se infinitos pontos reais define-se um pincel de planos epipolares, que têm em comum a linha base.

Linha Epipolar

É a linha formada pela intersecção de um plano epipolar com o plano de projeção de uma das câmeras. Toda linha epipolar de uma das câmeras tem sua equivalente no plano de imagem da outra câmera. Em decorrência da existência de pincéis de planos epipolares, existem também pincéis de linhas epipolares que têm em comum o epipólo. Na Figura 2.2 são as linhas $\overline{m_1 e_{1,2}}$ e $\overline{m_2 e_{2,1}}$.

Matriz Fundamental

É a matriz que encapsula a relação geométrica entre duas câmeras não calibradas, como definido por a Hartley e Zisserman (3). Sejam x_0 e x_1 as projeções de um ponto X nos planos retinais de duas câmeras. A matriz fundamental F é aquela que satisfaz a relação:

$$x_0^T F x_1 = 0.$$

Restrição Epipolar

A restrição epipolar garante que se um ponto é visível para as duas câmeras, as projeções deste nos dois planos retinais estarão nas linhas epipolares correspondentes. Ela se baseia no fato de as projeções do ponto real serem coplanares, caindo, assim, nas intersecções do plano epipolar com os planos retinais.

Na Figura 2.2 as projeções m_1 e m_2 do ponto M aparecem na linhas epipolares $\overline{m_1 e_{1,2}}$ e $\overline{m_2 e_{2,1}}$. A restrição epipolar permite reduzir a dimensão do problema de encontrar as projeções correspondentes nas imagens das duas câmeras. Sem ela seria necessário varrer a imagem

Figura 2.2 – Modelo esquemático descritivo dos conceitos da geometria epipolar envolvidos em uma montagem genérica de câmeras, adaptada de Heyden e Pollefeys (5)

completa aplicando uma métrica de comparação de semelhança a conjuntos de *pixels* para identificar o que mais se aproximasse ao que é buscado. Dada a restrição epipolar, esta busca se reduz à linha epipolar correspondente à do *pixel* cuja projeção semelhante é procurada.

2.3 Interpolação de Visões

A interpolação de visões (*view morphing*) foi introduzida por Seitz e Dyer (6) no trabalho em que é proposta uma técnica criar visões interpoladas de uma cena a partir de duas visões-base, mantendo a coerência espacial. Ou seja, a partir de duas imagens de uma mesma cena, é possível sintetizar imagens intermediárias sem introduzir distorções que comprometam a estrutura física da cena.

É importante diferenciar a interpolação de imagens e a interpolação de visões. Enquanto na primeira considera-se apenas a imagem como fonte de dados para interpolação, a segunda baseia-se na imagem e na matriz de transformação da câmera que a capturou. Desta maneira é possível interpolar também valores de posicionamento da câmera, dando a ilusão de que esta se move sobre a linha base.

Para obter este efeito em montagens quaisquer de câmeras (salvo configurações singulares descritas em 3.1.4) emprega-se um algoritmo de três fases. A primeira faz o alinhamento das linhas epipolares correspondentes baseada na relação geométrica entre as câmeras. A fase intermediária encontra as correspondências e usa os valores de posição obtidos para realizar uma operação de *image morphing* comum entre as imagens com linhas epipolares alinhadas. Finalmente, a terceira etapa projeta o resultado da interpolação de imagens intermediária

para que este tenha a aparência visual de uma imagem capturada por uma câmera real na orientação obtida na interpolação intermediária. Na Seção 3.1 a técnica de *view morphing* será apresentada em uma formulação mais formal, e com mais detalhes.

2.4 Rastreamento

Para atingir os objetivos descritos na Seção 1.1 é necessário empregar um sistema que permita conhecer a posição da cabeça do usuário em relação à sua plataforma de visualização. Um sistema deste tipo tem a finalidade de encontrar a posição da cabeça do usuário dentro de um segmento conhecido do espaço.

Existem diversas alternativas para rastreamento e elas diferem principalmente em precisão, custo, facilidade de obtenção, sensibilidade e necessidade de acoplar dispositivos ao elemento rastreado, no caso, o usuário. Rastreadores magnéticos como Flock of Birds (7) são precisos, possuem 6 graus de liberdade, são fáceis de se obter e seu preço não é proibitivo. Entretanto são sensíveis a interferências e limitados em área de atuação, sendo que modelos mais robustos atingem preços bem mais altos que os mais simples. O principal aspecto negativo deste tipo de ferramenta é a necessidade de acoplamento de um dispositivo ao usuário, o que se revela desconfortável, pouco prático e anti-higiênico.

Assim como os magnéticos, os rastreadores de ultra-som necessitam que um elemento seja acoplado ao usuário e são sensíveis a interferências. Além disso também têm área de atuação limitada e são sensíveis a eco. Seu uso é quase sempre limitado a pesquisas e eles são difíceis de se obter.

Técnicas baseadas em visão computacional, por outro lado, mostram-se adequadas para a aplicação pois não têm custo elevado e podem ser independentes de acoplamento de dispositivos ao usuário. Também são simples de implementar e precisas, ficando sua restrição a cargo da sensibilidade à iluminação do ambiente, oclusão e sombra. Dada a simplicidade da restrição, a visão computacional é a opção mais adequada para este sistema. A seguir são descritas teoricamente algumas técnicas levantadas para aplicação na Janela 3D e os resultados obtidos em experimentação são apresentados no Capítulo 5.

2.4.1 Marcadores Coloridos

O rastreamento de marcadores coloridos tem custo baixo e é fácil de ser instanciado em ambientes diversos. Há diversas formas de implementar a detecção de cores específicas e no geral elas são de baixo custo computacional. Os principais fatores contrários à utilização deste tipo de ferramenta são a grande sensibilidade à iluminação do ambiente, a necessidade de

calibração para cada cenário de uso, que também não pode ter nenhum elemento da mesma cor que as rastreadas e a obrigação de acoplar um marcador ao usuário. Na Figura 5.14 é mostrado um exemplo de marcador colorido. A Seção 5.4.1 explica melhor os algoritmos envolvidos e os resultados obtidos em testes.

2.4.2 *Marcadores Fiduciais*

Marcadores fiduciais são ferramentas de rastreamento que atendem, com excesso, os requisitos levantados para o sistema de rastreamento. Eles consistem em desenhos pré-determinados e conhecidos pelo sistema de rastreamento que, quando detectados, permitem calcular sua posição e orientação no espaço. Apresentam pouca sensibilidade à iluminação devido o grande contraste entre o branco e o preto puros, usados para desenhar o marcador em uma superfície qualquer. A necessidade de um desenho específico torna este tipo de marcador um pouco mais difícil de reproduzir que os coloridos, sem, contudo, inviabilizar seu uso. Por trabalhar com a detecção de uma forma geométrica conhecida eles conseguem maior estabilidade que os marcadores coloridos, sem que seja necessário um processo especial de otimização. A Seção 5.4.2 dá uma visão mais aprofundada dos marcadores fiduciais e comenta resultados obtidos em testes. Um fator contrário a este tipo de ferramenta é que também apresentam o contratempo de obrigarem o usuário a acoplar elementos ao corpo. A Figura 5.15 mostra exemplos de marcadores fiduciais.

2.4.3 *Classificadores de Características Haar-like*

Características *Haar-like* são variações de contraste na imagem, como exemplificado na Figura 5.16. Quando empregadas em conjunto, estas características permitem reconhecer elementos mais complexos, como linhas ou contornos de objetos. O reconhecimento de conjuntos destes objetos torna possível identificar, por exemplo, o rosto do usuário.

A principal vantagem deste tipo de técnica frente às outras apontadas é não depender de um elemento que deve ser acoplado ao usuário. A sensibilidade à iluminação também é menor, já que alterações globais não alteram o contraste local, a não ser em casos extremos. Além disso não é necessário criar nenhum tipo de dispositivo para interagir com o sistema. A Seção 5.4.3 aborda este sistema com mais detalhes e apresenta os resultados obtidos em seções de testes.

2.5 Mistura do Mundo Real com o Virtual

A mistura de ambientes reais e virtuais tem sido objeto de interesse de diversos autores. Grande quantidade de resultados poderia ser descrita nesta Seção, porém ela se dedica especialmente a situar este trabalho dentro deste campo de pesquisa analisando quais as contribuições possíveis.

A Figura 2.3 apresenta uma taxonomia proposta por Milgram et al. (8) para classificar sistemas em que elementos reais e sintéticos são colocados em um ambiente comum. Nos extremos estão a realidade pura e a realidade virtual sem nenhum tipo de elemento real. Entre elas, a realidade aumentada e a virtualidade aumentada. Em ambas elementos reais e virtuais são colocados no mesmo ambiente, porém na primeira há predominância de elementos reais e na segunda de virtuais. Muitas vezes é difícil distinguir qual é a classificação de um sistema ou se ele pode ser enquadrado em mais de uma, assim usa-se a denominação de *sistema de realidade misturada* para definir este tipo de aplicação de maneira geral.

Figura 2.3 – *Continuum* da realidade misturada, adaptada de Milgram et al. (8)

A Janela 3D pode se tornar tanto um sistema de realidade aumentada quanto de virtualidade aumentada. Se a renderização de um avatar for inserida em um ambiente virtual como em Siscoutto e Tori (9) ou Life (10) será o caso de uma aplicação de virtualidade aumentada, já se as visões interpoladas forem enriquecidas com elementos virtuais será um caso de realidade aumentada baseada em imagens do mundo real sintetizadas.

2.6 Considerações Finais

Foram apresentados conceitos das diversas áreas do conhecimento envolvidas no desenvolvimento da Janela 3D. O principal objetivo desta Seção é familiarizar o leitor com as principais teorias e técnicas que serão empregadas ao longo do texto. Uma visão mais aprofundada de cada uma delas pode ser obtida nas referências citadas.

3 Trabalhos Relacionados

Nesta Seção são apresentados trabalhos da área de telecomunicação imersiva, telepresença ou que, de alguma maneira, tenham relação direta com as técnicas e conhecimentos apresentados nesta dissertação.

A Seção 3.1 fala sobre o trabalho *Image-Based Transformation of Viewpoint and Scene Appearance*, de Seitz (4), que é a base para o desenvolvimento deste. Em 3.2 é discutido um trabalho anterior também desenvolvido no Interlab (1), cujo objetivo era criar um vídeo-avator estereoscópico para aplicações de vídeo conferência. A Seção 3.3 introduz um sistema de vídeo avator com controle de colisão 3D(tridimensional) baseado em equipamentos de baixo custo. Na Seção 3.4 é analisado um trabalho de Timm (11), cujo objetivo é gerar visões sintéticas ao longo da linha base de um sistema de duas câmeras, porém aplicado especificamente a rostos de pessoas. Uma técnica baseada no uso de grande quantidade de pequenos *billboards* pra aproximar a superfície de um objeto é apresentada na Seção 3.5.

3.1 Image-Based Transformation of Viewpoint and Scene Appearance

O termo *view morphing* foi empregado pela primeira vez em Seitz e Dyer (6), entretanto será analisada a tese de doutorado Seitz (4) do mesmo autor, publicada no ano seguinte. A opção pela tese se baseia no fato de a mesma ter sido base para o artigo em que se cunhou o termo *view morphing* e conter toda a informação deste, além de estudos mais aprofundados dos temas relacionados à técnica.

3.1.1 Image Morphing e View Morphing

Embora os nomes se pareçam *image morphing* e *view morphing* são técnicas diferentes para atingir resultados relacionados, mas distintos. Ambas se dedicam a obter transições naturais entre elementos, porém há uma diferença conceitual entre eles.

As técnicas de *image morphing* interpolam bidimensionalmente cores e formas a partir de duas imagens, criando uma transição suave entre elas como em Beier e Neely (12). Não há, no entanto, preocupação com possíveis distorções ou com a geração de imagens intermediárias que sejam fisicamente inconsistentes. A Figura 3.1 exemplifica este tipo de distorção quando

Figura 3.1 – Exemplo de distorção causada pelo processo de *image morphing*. Extraída de Seitz (4)

mostra o relógio curvado durante a transição.

Já o *view morphing* não se baseia em duas imagens, mas em duas visões, como sugere o nome. Interpolair visões é um processo mais abrangente e complexo do que interpolar imagens e consiste em criar valores intermediários para a orientação e posição da câmera, além de reprojeter imagens. Intuitivamente o efeito visual de tais interpolações é o mesmo de capturar uma tomada de vídeo em que a câmera se movimenta entre dois pontos posicionados nas visões inicial e final. Ao contrário de no *image morphing* a estrutura física da cena retratada é mantida nas imagens intermediárias geradas. A Figura 3.2a mostra um exemplo de imagens intermediárias geradas usando *image morphing*. Nota-se como ocorre distorção na geometria do objeto, chegando ao ponto de haver desconexão entre os componentes do brinquedo. Já em 3.2b as imagens intermediárias não introduzem distorção física no objeto, e a rotação é visualmente consistente.

(a) *Image Morphing*

(b) *View Morphing*

Figura 3.2 – Comparação entre *Image Morphing* e *View Morphing*, adaptada de Seitz (4)

3.1.2 O Algoritmo

A proposta de Seitz é um algoritmo desenvolvido evolutivamente. Partindo de uma situação de captura em que a formulação matemática do problema é simples são adicionadas restrições e pressupostos que permitem flexibilizar a estrutura da cena chegando a condições livres. Nas Seções seguintes serão expostas as condições de captura e os prós e contras destas situações.

3.1.3 Visões Paralelas

Esta é a situação que requer menos robustez do modelo matemático das transições, sendo uma interpolação linear suficiente para gerar imagens intermediárias fisicamente consistentes. Supondo que uma câmera com distância focal f_0 seja posicionada na origem do sistema de coordenadas do mundo e capture uma imagem $\mathcal{I}_0$, sua matriz de projeção $\mathbf{\Pi}_0$ é definida por Seitz (4) como:

$$\mathbf{\Pi}_0 = \begin{bmatrix} f_0 & 0 & 0 & 0 \\ 0 & f_0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Se esta câmera for movida em direção paralela ao seu plano de imagem para $(C_x, C_y, 0)$ e tiver sua distância focal alterada para f_1 consegue-se uma nova imagem $\mathcal{I}_1$ cuja matriz de projeção é:

$$\mathbf{\Pi}_1 = \begin{bmatrix} f_1 & 0 & 0 & -f_1 C_x \\ 0 & f_1 & 0 & -f_1 C_y \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

O efeito visual desta operação é uma movimentação da câmera na linha $\overline{\mathbf{C}_0\mathbf{C}_1}$, que liga o centro óptico da câmera nas posições $(0, 0, 0)$ e $(C_x, C_y, 0)$, em conjunto com uma alteração em seu *zoom*. Para gerar as visões intermediárias de maneira natural é preciso interpolar as imagens-base e as matrizes de projeção. Sejam $\mathbf{p}_0$ e $\mathbf{p}_1$ projeções de um ponto $[X \ Y \ Z \ 1]^T$ da cena real nas imagens $\mathcal{I}_0$ e $\mathcal{I}_1$, respectivamente. A interpolação linear deles é dada por:

$$\begin{aligned} (1-s)\mathbf{p}_0 + s\mathbf{p}_1 &= (1-s)\frac{1}{Z}\mathbf{\Pi}_0\mathbf{P} + s\frac{1}{Z}\mathbf{\Pi}_1\mathbf{P} = \\ &= \frac{1}{Z}\mathbf{\Pi}_s\mathbf{P}. \end{aligned}$$

Onde:

$$\begin{aligned}
\mathbf{\Pi}_s &= (1-s)\mathbf{\Pi}_0 + s\mathbf{\Pi}_1 \\
&= \begin{bmatrix} f_s & 0 & 0 & 0 & -f_s\alpha_s C_X \\ 0 & f_s & 0 & 0 & -f_s\alpha_s C_Y \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \\
f_s &= (1-s)f_0 + sf_1 \\
\alpha_s &= \frac{sf_1}{(1-s)f_0 + sf_1} \\
s &\in (0, 1)
\end{aligned}$$

e α é um fator de correção empregado para uniformizar a transição, já que esta tende a acontecer mais lentamente na medida em que a distância focal aumenta. Desta maneira o resultado de cada interpolação é uma imagem com matriz de projeção $\mathbf{\Pi}_s$ e centro óptico

$$\mathbf{C}_s = (\alpha_s C_X, \alpha_s C_Y, 0). \quad (3.1)$$

3.1.4 Visões Não Paralelas com Configuração Conhecida

Nesta abordagem a restrição de paralelismo entre as câmeras é removida. Para isso é empregada reprojeção de imagens, que permite que visões distintas sejam transformadas em visões paralelas e desta maneira é possível empregar a interpolação linear como no caso anterior. O algoritmo é dividido agora em três etapas, havendo uma projeção inicial, a interpolação propriamente dita e uma projeção final.

Projeção Inicial

Nesta etapa as imagens oriundas de câmeras em posições não paralelas são transformadas para que se tornem coplanares. Duas visões que tenham um mesmo centro óptico são relacionadas por uma transformação projetiva planar Seitz (4). Dadas duas imagens $\mathcal{I}$ e $\hat{\mathcal{I}}$ com matrizes de projeção $\mathbf{\Pi} = [\mathbf{H} | -\mathbf{HC}]$ e $\hat{\mathbf{\Pi}} = [\hat{\mathbf{H}} | -\hat{\mathbf{H}}\mathbf{C}]$, as projeções $\mathbf{p} \in \mathcal{I}$ e $\hat{\mathbf{p}} \in \hat{\mathcal{I}}$ de um ponto $\mathbf{P}$ da cena real se relacionam por:

$$\begin{aligned}
\hat{\mathbf{H}}\mathbf{H}^{-1}\mathbf{p} &= \hat{\mathbf{H}}\mathbf{H}^{-1}\mathbf{I}\mathbf{P} \\
&= \hat{\mathbf{I}}\mathbf{P} \\
&= \hat{\mathbf{p}}
\end{aligned}$$

onde $\hat{\mathbf{H}}\mathbf{H}^{-1}$ é a matriz 3×3 que representa a transformação do plano de imagem $\mathcal{I}$ no plano $\hat{\mathcal{I}}$.

Escolhas convenientes tornam o cálculo das projeções mais simples. Posicionando as câmeras de modo que fiquem no eixo x do sistema de coordenadas do mundo resta apenas definir as posições nos eixos y e z . É comum que o eixo y seja colocado na direção do produto vetorial das duas normais dos planos de imagem. Diferentes escolhas para a posição no eixo z representam diferentes planos, assim elas deve ser iguais para ambas as câmeras.

Aplicando-se a transformação $\mathbf{H}_0^{-1}$ à imagem $\mathcal{I}_0$ e $\mathbf{H}_1^{-1}$ à imagem $\mathcal{I}_1$ obtêm-se as projeções iniciais $\hat{\mathcal{I}}_0$ e $\hat{\mathcal{I}}_1$ sobre o mesmo plano, necessárias para a próxima etapa do algoritmo, o *image morphing*.

Image Morphing

A operação de *image morphing* baseia-se em duas imagens e interpola cores e posições de *pixels* entre elas, gerando transições suaves. Ela não garante a conservação da estrutura tridimensional da cena retratada nas imagens nas imagens geradas, porém sua aplicação é válida na medida em que o passo anterior, a pré-projeção, garante que as imagens base sejam coplanares e que suas linhas epipolares estejam alinhadas.

A transição é especificada por duas imagens $\mathcal{I}_0$ e $\mathcal{I}_1$ e mapeamentos $C_{01} : \mathcal{I}_0 \Rightarrow \mathcal{I}_1$ e $C_{10} : \mathcal{I}_1 \Rightarrow \mathcal{I}_0$ que especificam a correspondência entre os pontos das duas imagens. A necessidade de dois mapas se deve ao fato de que a correspondência pode não ser um-para-um. Em termos práticos, alguns pares de pontos com correspondências conhecidas são usados para estimar as demais correspondências por métodos como descritos em Beier e Neely (12), Lee et al. (13) e Wolberg (14). A partir dos mapas de correspondências é possível definir funções de transição para as duas imagens, baseadas em interpolação linear. Dados dois pontos $\mathbf{p}_0 \in \mathcal{I}_0$ e $\mathbf{p}_1 \in \mathcal{I}_1$, e $s \in [0, 1]$ as funções de interpolação $C_{0s}(\mathbf{p}_0)$ e $C_{1s}(\mathbf{p}_1)$ são da forma, segundo definido por Seitz (4):

$$\begin{aligned}
C_{0s}(\mathbf{p}_0) &= (1 - s)\mathbf{p}_0 + sC_{01}(\mathbf{p}_0) \\
C_{1s}(\mathbf{p}_1) &= (1 - s)C_{10}(\mathbf{p}_1) + s\mathbf{p}_1.
\end{aligned}$$

Existem diversas técnicas para gerar transições de imagens, que usualmente se distinguem pela maneira como geram os mapas de correspondência. Em Seitz (4) o termo *image morphing* é usado em referência às técnicas baseadas em operações pontuais lineares, como as empregadas nos já citados Beier e Neely (12), Lee et al. (13) e Wolberg (14). Esta decisão se justifica na medida em que a operação de pré-projeção pode causar alterações na imagem que tornem difícil identificar estruturas específicas na imagem enquanto a busca de semelhanças entre pontos é facilitada sobremaneira pelo alinhamento das linhas epipolares.

Busca por Correspondências

Apesar de alinhadas, as linhas epipolares de diferentes visões apresentam quantidades diferentes de *pixels* para representar os mesmos objetos. A Figura 3.3 mostra este efeito nas imagens I_0 , I_s e I_1 . É possível notar como a mesma parte de um objeto tem comprimento diferente quando projetada sobre os diferentes planos de imagem, por exemplo, a projeção dos extremos do objeto S_2 . A mesma Figura ilustra também um pressuposto da *monotonicidade*, que permite tornar o problema de busca de correspondências nas linhas epipolares mais simples. Adotando uma definição informal, a monotonicidade garante que a ordem dos elementos em uma cena é mantida, ou seja, se um elemento aparece à esquerda de um outro elemento na câmera direita ele aparecerá à esquerda nas visões intermediárias e na câmera direita também. Isto permite que a busca de correspondências possa ser feita apenas em trechos das linhas epipolares e seguindo uma orientação, o que diminui o número de operações executadas.

Figura 3.3 – Exemplo de diferentes quantidades de *pixels* para representar os mesmos objetos em visões distintas e de monotonicidade, Seitz (4)

De maneira formal, a restrição da monotonicidade pode ser definida como ilustrado pela Figura 3.4. Sejam P e Q pontos quaisquer em um plano epipolar de uma dada montagem de câmeras. Sejam Θ_0 e Θ_1 os ângulos entre as linhas que ligam P e Q aos centros ópticos C_0

e C_1 . Para que a restrição de monotonicidade seja verdadeira $\Theta_0\Theta_1 > 0$. Se esta condição for satisfeita em C_0 e C_1 , também será em qualquer $C_s \in \overline{C_0C_1}$.

Figura 3.4 – Dados dois pontos P e Q no mesmo plano epipolar que determinam ângulos Θ_0 e Θ_1 com os centros ópticos C_0 e C_1 das câmeras, a monotonicidade existe se $\Theta_0\Theta_1 > 0$, Seitz (4)

Para estabelecer a relação entre as representações dos objetos é necessário identificar *pixels* correspondentes ao longo das linhas epipolares no par estéreo. Com isso é possível obter as posições que devem ser usadas como bases para interpolação da imagem. Usualmente busca-se por correspondências em zonas de transição da imagem, como trechos em que a representação de um objeto termina e começa a de outro. Além de requerer menos buscas e comparações, esta abordagem é eficiente na medida em que nunca ocorrerá correspondência total em linhas epipolares de imagens de visões distintas.

A busca baseia-se na comparação de uma vizinhança de um *pixel* candidato a correspondente do *pixel* atual para definir qual é o que mais se assemelha a este. Para isso é necessário empregar uma métrica de similaridade que permita estabelecer ordenação entre os componentes da vizinhança analisada. Por não haver necessidade de valores absolutos representativos, mas apenas coerência entre os valores obtidos para cada componente da vizinhança, e para reduzir o custo computacional é usada a somatória dos módulos das diferenças entre os *pixels*. Seja $\mathcal{N}$ o conjunto dos pontos $\mathbf{p}_t$ que compõem uma vizinhança do ponto base $\mathbf{p}_b$. A similaridade de $\mathbf{p}_t$ com um ponto base $\mathbf{p}_b$ é dada por:

$$S(\mathbf{p}_b, \mathcal{N}) = \sum_t |\mathbf{p}_b - \mathbf{p}_t|.$$

A distância euclidiana poderia ser empregada da mesma forma, porém o custo computacional de uma operação de extração de raiz quadrada não compensa o benefício obtido. Um outro artifício usado para tornar o problema menos custoso é quantizar os tons de cinza. Seitz determinou empiricamente que 30 intervalos são suficientes para gerar resultados visualmente convincentes com nível de processamento aceitável.

Partindo-se das imagens base pré-projetadas e aplicando-se a interpolação consegue-se uma imagem intermediária que será coplanar às duas imagens base pré-projetadas. Para chegar ao

resultado final é necessário fazer uma projeção desta imagem, usando uma matriz obtida pela das duas matrizes base.

Pós-Projeção

Nesta etapa a imagem obtida pelo *image morphing* é transformada para que deixe de ser coplanar às duas imagens base pré-projetadas e tenha a orientação esperada para uma câmera real. A pós-projeção completa o processo de transição de acordo com o conceito de visão (imagem aliada à matriz de projeção da câmera que a gerou) adotado neste trabalho, gerando valores intermediários para a matriz de projeção das visões intermediárias. Recapitulando, a matriz de projeção de uma visão intermediária é da forma $\Pi_s = [\mathbf{H}_s | -\mathbf{H}_s \mathbf{C}_s]$. Como a $\mathbf{C}_s$ é dado pela Equação 3.1, a tarefa resume-se a obter $\mathbf{H}_s$, $s \in (0, 1)$. Esta transformação pode ser obtida a partir da rotação dos planos de imagem ao redor de um eixo. Sejam $\vec{\mathbf{N}}_0$ e $\vec{\mathbf{N}}_1$ os vetores unitários normais a um plano de imagem, o eixo $\mathbf{D}$ e o ângulo θ de rotação são dados por:

$$\begin{aligned}\mathbf{D} &= \mathbf{N}_0 \times \mathbf{N}_1 \\ \theta &= \cos^{-1}(\mathbf{N}_0 \cdot \mathbf{N}_1).\end{aligned}$$

Em termos práticos a pós-projeção pode ser obtida de maneira automática a partir das matrizes de pré-projeção ou manualmente. A obtenção automática é vantajosa quando há restrições de tempo na execução do programa, apesar de introduzir erro inerente ao cálculo. Já a obtenção manual gera resultados de maior qualidade, porém demanda interação do usuário. O principal fator de aumento de qualidade é a possibilidade de que a o cálculo da matriz de pós-projeção seja feito de maneira iterativa, submetendo os resultados intermediários à avaliação humana para que se atinja um resultado final satisfatório. Outra grande vantagem da obtenção manual é que, por ser baseada no apontamento de pontos correspondentes, é possível criar transições entre imagens diferentes de objetos semelhantes, por exemplo, rostos de duas pessoas.

Configurações Singulares

Apesar da robustez da técnica de pré e pós-projeção existem configurações de câmeras que não podem ser tornadas paralelas. Situações em que o centro óptico de uma câmera esteja dentro do campo de visão da outra e vice-versa são configurações singulares e não podem ser feitas paralelas através de reprojeção. Um exemplo é mostrado na Figura 3.5. Nesta

montagem o alinhamento seria impossível pois a câmera A enxerga o centro óptico da câmera B.

Figura 3.5 – Configuração singular de câmeras, adaptada de Seitz (4)

3.1.5 Visões Quaisquer

Para estender a técnica de três etapas para quaisquer cenas, é necessário eliminar a restrição de que a configuração das câmeras deve ser conhecida. Ou seja, deve ser possível computar as matrizes de pré-projeção sem que se conheça explicitamente as matrizes Π_0 e Π_1 de projeção das câmeras.

Supondo que uma cena capturada por uma câmera cuja matriz de transformação é Π sofre uma transformação projetiva T , a projeção de um ponto P da cena será da forma $\tilde{P} = \Pi(TP)$. Dada a associatividade da multiplicação de matrizes pode-se reescrever a equação da forma:

$$\begin{aligned}\tilde{P} &= (\Pi T)P \\ &= \tilde{\Pi}P\end{aligned}$$

onde $\tilde{\Pi}$ é a matriz de projeção da câmera que também engloba a transformação projetiva da cena. Isto permite que, a partir de um processo de calibração sobre as imagens base, se obtenha implicitamente a matriz da transformação projetiva sofrida pelas câmeras, o que permite que elas seja posicionadas livremente. Indo mais além, é possível usar inclusive imagens base de cenas distintas. Para adequar as imagens das câmeras ao algoritmo de três passos proposto é preciso obter transformações projetivas que façam com que pontos correspondentes da imagem apareçam nas mesmas linhas epipolares. Diversas técnicas podem ser empregadas para isso, sendo que Seitz calcula estas matrizes a partir de oito ou mais pares pontos correspondentes indicados manualmente. Uma vez obtidas, é possível realizar o *image morphing* intermediário.

O último passo para obter transições com coerência física usando o algoritmo de três etapas é fazer a pós-projeção, para levar a imagem resultante do *view morphing* ao plano resultante da interpolação dos planos das visões base. No caso ideal seria necessário obter uma matriz de pós projeção H_s , $s \in (0, 1)$, o que se revela proibitivo pelo tempo gasto. Seitz propõe um

método alternativo que se baseia na especificação de correspondências entre as imagens base $\mathcal{I}_0$ e $\mathcal{I}_1$ e a interpolação $\mathcal{I}_{0.5}$. O envoltório retangular de uma região das imagens é apontado manualmente nas imagens base e também na imagem intermediária, e assim consegue-se saber com exatidão a matriz de pós-projeção para $s = 0.5$. Após isso é feita uma interpolação entre as posições dos vértices apontados em $\mathcal{I}_0$, $\mathcal{I}_{0.5}$ e $\mathcal{I}_1$ e as matrizes de pós-projeção para todo s são estimadas.

3.1.6 *Relação com o Projeto*

Seitz (4) introduziu a técnica de síntese de visões a partir de visões quaisquer. Outros pesquisadores, entre eles Chen e Williams (15), Laveau e Faugeras (16), Tomasi e Kanade (17), Ott et al. (18), Poggio e Brunelli (19), já conseguiam gerar visões intermediárias a partir de imagens base porém dependendo de configurações conhecidas e câmeras calibradas. Embora seja possível atingir resultados semelhantes com estes métodos, no contexto deste projeto é desejável que seja possível montar o sistema que gera imagens independente de processos de calibração complexos e que este seja flexível a movimentações (ainda que exigindo nova calibração) e imparalelismo das câmeras.

3.2 AVTC

Na AVTC (Arquitetura para Teleconferência Baseada na Integração de Vídeo Avatar Esteoscópico em Ambiente Virtual Tridimensional), Siscoutto e Tori (9) delineiam um sistema para teleconferência contemplando capacidade de movimentação, realismo visual da imagem transmitida, comunicação sonora, virtualidade aumentada por vídeo-avatar e capacidade de interação de maneira flexível, permitindo o uso de diversos tipos de equipamentos. São analisados fatores humanos da aplicação das tecnologias sugeridas e é descrita a implementação de uma prova de conceito do sistema. A seguir são brevemente analisados os principais aspectos considerados pelos autores do trabalho.

3.2.1 *Aparência Visual do Avatar*

O fator interativo cujos resultados foram mais elaborados foi a comunicação visual. Numa infra-estrutura com duas câmeras, uma pessoa tem sua imagem capturada e inserida em um ambiente virtual tridimensional. Para isso é empregada a técnica de *chroma keying*, o que restringe a captura a um ambiente preparado, com ao menos uma superfície de cor conhecida

usada como fundo e condições especiais de iluminação.

O par de câmeras permite que haja uma sensação maior de profundidade em relação ao avatar inserido na cena virtual já que cada uma faz o papel de um olho da pessoa que visualizará a cena. A imagem segmentada do usuário é aplicada como textura a *billboards*¹ e os *pixels* de fundo são considerados transparentes. Apresentando a imagem da câmera esquerda ao olho esquerdo do usuário e a direita ao olho direito consegue-se criar uma ilusão maior de profundidade no avatar. Na implementação feita, o controle da apresentação para cada olho era feito através de óculos obturadores sincronizados com a renderização.

3.2.2 Movimentação

A arquitetura proposta prevê a existência de uma interface de movimentação para o usuário. Esta deve permitir que ele mova o avatar em todas as direções sem que sua posição seja alterada em relação ao ambiente em que ele se encontra e mantendo a maior verossimilhança possível entre o movimento executado pelo avatar e pelo usuário. Desta maneira é possível que o avatar caminhe no ambiente virtual sem que o usuário e o sistema de captura tenham que se locomover no ambiente real. Também é prevista a funcionalidade de controle para a interface de movimentação, por meio de botões e ações determinadas o usuário pode movimentar o avatar e a câmera ou o ambiente virtuais.

Inicialmente a idéia era criar um tipo de esteira com sensores de movimentação que permitisse detectar a direção e o sentido da movimentação do usuário em dois eixos e estimar a movimentação diagonal a partir deles. A seleção do tipo de manipulação da cena virtual que se deseja seria selecionada através de botões. Para a prova de conceito implementada a interface de movimentação consiste em uma superfície com uma *trackball* controlada com os pés e um *mouse* sem fio para faz o papel dos botões da esteira.

3.2.3 Interação

A interação do avatar com o ambiente virtual existe, porém foi implementada de maneira a prover apenas coerência espacial, sem preocupação com manipulação ou capacidade de detectar colisões em pequena escala. No sistema de detecção de colisões o corpo do avatar é aproximado por um sólido envoltório e cada elemento virtual da cena também tem sua geometria aproximada por um sólido envoltório. Assim, o avatar pode mover-se livremente enquanto o seu sólido envoltório não invadir o espaço do sólido envoltório de algum elemento

¹*Billboard* é uma técnica para apresentação de imagens bidimensionais em cenas tridimensionais. Elas são usadas como texturas em planos que estão sempre paralelos ao plano de imagem da câmera. Maiores informações podem ser obtidas em Watt (20)

da cena. Esta verificação impede que ele atravesse os objetos virtuais.

Apesar de simples este mecanismo de interação entre o avatar e a cena é de grande valia para manter a sensação de imersão das pessoas que assistem à inserção do avatar na cena virtual. O investimento de tempo para obter grande qualidade gráfica e maior impressão de profundidade seria injustificável sem um mecanismo que impedisse incoerências grosseiras no ambiente de realidade misturada.

3.2.4 Comunicação Sonora

É feita através da captura dos sons do ambiente remoto por um microfone. A informação de som capturada é preparada para transmissão e é enviada junto com os dados de controle e outros tipos de dados. O cliente é responsável por executar este som de acordo com a chegada das informações.

3.2.5 Relação com o Projeto

Esse foi o trabalho que iniciou a linha de pesquisa de telecomunicação imersiva dentro do Interlab (1), e é uma proposta conceitual de arquitetura para teleconferência. Também foi empregada uma solução simples que dá a ilusão de tridimensionalidade sem considerar o ponto de vista do espectador, porém que mostrou na prática a melhoria na sensação de imersão dada por uma apresentação visual mais coerente com a realidade.

3.3 AVMix

Nakamura e Tori (21) propõem no AVMix(Sistema de Vídeo-Avatar com Detecção de Colisão para Realidade Aumentada e Jogos) um sistema para inserção de vídeo avatar interativo em um ambiente virtual. Diferentemente de em Siscoutto e Tori (9), o volume do avatar não é aproximado por um sólido uniforme, mas constrói-se uma nuvem de pequenos sólidos que permite fazer a detecção de colisão no espaço 3D com maior precisão e detalhamento. Além disso, baixo custo, uso de produtos de prateleira e a necessidade de apenas um computador doméstico para execução são alguns requisitos atendidos pelo projeto. A principal contribuição dada pelo trabalho é a integração de técnicas e algoritmos aplicada a um sistema de baixo custo, focado em aplicações de educação e entretenimento.

As principais técnicas empregadas no projeto são mapas de disparidade para a estimativa de uma mapa de profundidade da cena real e nuvem de pontos para detecção de colisão

em ambiente 3D virtual. Aplicadas em conjunto, estas técnicas permitem que haja registro entre a cena real e a virtual, possibilitando que um usuário interaja com elementos sintéticos livremente, apenas movimentando-se no espaço.

3.3.1 Mapa de Profundidade

Dada uma cena real C projetada no plano de imagem I de uma câmera, de largura l e altura a . Denomina-se mapa de profundidade de C a matriz M de mesmas dimensões que I , cujo cada elemento M_{ij} contém a menor distância do elemento real amostrado no *pixel* I_{ij} ao plano de imagem I .

A obtenção do mapa de profundidade através do mapa de disparidade da cena é uma técnica baseada no sistema de visão humano. São empregadas duas câmeras para obter imagens de pontos de vista diferentes de uma mesma cena. Tais imagens apresentam um pequeno deslocamento dos objetos e esta paralaxe é usada para calcular a profundidade do objeto no ambiente. A situação ideal para imageamento estéreo é quando há deslocamento horizontal nas câmeras, porém elas estão sobre a mesma linha e seus eixos ópticos são paralelos, de maneira a minimizar a reamostragem de *pixels* nas linhas epipolares das imagens paralelas. Além disso também é necessário que haja superposição dos campos de visão da câmera. Desta maneira, conhecendo-se os parâmetros intrínsecos das câmeras, por semelhança de triângulos é possível obter a disparidade dos objetos da cena. Maiores informações sobre o cálculo das disparidades podem ser obtidas em Gonzalez e Woods (22).

Outras técnicas que podem ser empregadas para estimar mapas de profundidade se baseiam em diferenças de foco e em diferenças de iluminação. Subbarao e Surya (23) introduzem uma técnica de obtenção de profundidade a partir de duas imagens usando apenas a diferença de foco entre elas, não exigindo que as câmeras sejam focadas em nenhum ponto específico da cena. Zhang e Nayar (24) empregam um projetor como ferramenta de iluminação da cena e conseguem obter informação de profundidade através de uma câmera e de um modelo de desfocagem do projetor.

Liao et al. (25) desenvolvem uma técnica para extrair a profundidade dos elementos de uma cena baseada em imagens capturadas de um mesmo ponto de vista e sob condições de iluminação cuja variação é conhecida, por exemplo, uma lâmpada que se mova por um caminho pré-determinado. De acordo com a variação da intensidade da luz refletida pelos objetos é possível estimar a profundidade destes. Baseados apenas nas restrições de que as fontes de luz da cena devem estar distantes dos objetos e devem ser isotrópicas, ou seja, possuem as mesmas propriedades em todas as direções, Basri et al. (26) apresentaram uma técnica que permite recuperar informação de profundidade através da otimização de um problema de harmônicas esféricas de baixa dimensão.

3.3.2 Detecção de Colisão Espacial

A detecção de colisões tridimensional é feita determinando a pertinência ou não de um elemento ao volume do espaço delimitado por outro, ou seja, calculando se um objeto contém pelo menos um ponto de um segundo objeto. A representação das geometrias de colisão envolvidas proposta é distinta para os sólidos da cena virtual e para o avatar. Enquanto os corpos virtuais podem ser representados por esferas ou cubos, o avatar é representado por uma *octree* de cubos alinhados com o sistema de coordenadas das câmeras.

A *octree* quantiza um volume do espaço real em que a detecção de colisão será realizada. Assim, a sua resolução espacial depende da dimensão do volume a ser quantizado e de sua profundidade máxima, fator que pode causar lentidão no processamento, inviabilizando o uso da aplicação em taxas interativas. O menor deslocamento perceptível (resolução) R numa determinada configuração é dado pela expressão

$$R = \frac{L}{2^{N-1}} \quad (3.2)$$

onde N é a profundidade da *octree* e L é a dimensão de uma aresta da partição cúbica do espaço real em que a detecção de colisão ocorre. Uma vez montada a *octree*, a detecção de colisão é feita, testando-se recursivamente a intersecção entre dois volumes em cada nível, enquanto houver intersecções. Quando é atingida um nó terminal da árvore diz-se que houve colisão.

3.3.3 Relação com o Projeto

O relacionamento entre o trabalho descrito nesta dissertação e Nakamura e Tori (21) pode ser analisado tanto pela semelhança das técnicas empregadas quanto pela proximidade e possibilidade de integração dos sistemas resultantes. Nakamura e Tori (21) fazem uso intensivo dos modelos da geometria epipolar para obter a profundidade real de cada *pixel* da cena observada pelas câmeras. Neste trabalho o mesmo tipo de técnicas da geometria projetiva é empregado para retificar as imagens capturadas pelas câmeras e conseguir a manutenção da estrutura 3D da cena real nas imagens sintetizadas.

Além disso, ambos trabalhos integram a linha de pesquisa de telecomunicação imersiva do Interlab, cujo primeiro trabalho foi Siscoutto e Tori (9) e fazem parte do Projeto Video-Avatar do Interlab no TIDIA-Ae. Dentro de um sistema completo de telepresença, as condições semelhantes de captura e as técnicas parecidas usadas no AVMix e na Janela 3D permitem seu uso em conjunto para obter uma apresentação visual realista da cena alvo com consistência em relação à movimentação do usuário em relação à superfície de visualização e ao mesmo

tempo prover interação tridimensional entre objetos virtuais e um avatar.

3.4 Real Time View Morphing of Video Streams

Nesse trabalho é criada uma aplicação sobre as bases teóricas introduzidas por Seitz (4). Reduzindo o escopo da aplicação de *view morphing* à cabeça e ombros de seres humanos, Timm (11) se baseia em pressupostos acerca do objeto de interesse para conseguir taxas de atualização que permitem operar sobre vídeos em tempo real em computadores de uso doméstico. O conhecimento prévio do objeto de interesse torna possível uma série de simplificações e otimizações. Dessa maneira Timm consegue executar *view morphing* de vídeos em tempo real, enquanto reporta que em Seitz (4) cada renderização demorava alguns segundos.

3.4.1 Otimizações

Remoção de Fundo

A primeira otimização advinda dos pressupostos relacionados é a possibilidade de remoção do fundo da imagem, que permite trabalhar com uma imagem de dimensões menores e com a mesma quantidade de informação que a imagem original no contexto do problema a ser resolvido. Além disso a remoção de fundo também pode ser otimizada. Da maneira geral o usuário tende a ocupar grande parte da área da imagem e sua silhueta ocupa a parte inferior, podendo ou não chegar ao topo. A técnica de *chroma keying* foi implementada de maneira a aproveitar essa heurística. Os *pixels* são varridos de baixo para cima, inicialmente da esquerda para a direita, até que se encontre uma sequência de valores que não são fundo. A partir daí começa uma varredura da direita para esquerda na mesma linha da imagem, novamente até que se encontre *pixels* que não são fundo. Este processo é repetido em cada linha da imagem até que se encontre uma linha composta apenas por valores de fundo ou até que se chegue no topo da imagem. Desta maneira o número de *pixels* que são varridos é reduzido drasticamente já que os valores de dentro da silhueta do usuário não lidos e a busca é interrompida assim que se chega à altura máxima da cabeça. Conhecer aproximadamente a forma do usuário também permite que o sistema de remoção de fundo seja mais tolerante a falhas desconsiderando pequenos falsos-positivos de fundo, ou grandes discontinuidades na silhueta.

Normalização Cilíndrica

Também devido ao conhecimento da forma aproximada da cabeça é possível pré-processar as imagens para que a busca por equivalências no par estéreo seja mais eficiente. Timm afirma que cada linha da imagem da cabeça é uma projeção de uma secção transversal desta e que sua representação mais adequada seria quando projetada sobre um cilindro e não sobre um plano. Assim as imagens passam por uma normalização e por um mapeamento cilíndrico, o que resulta em imagens que contém apenas o usuário e permitem que a busca seja mais eficiente em regiões como orelhas e laterais do rosto. A Figura 3.6 mostra o efeito do mapeamento cilíndrico. Em 3.6a é mostrada uma ilustração da imagem capturada pela câmera e 3.6b mostra o resultado esperado para o mapeamento.

XMIP-NCP

XMIP-NCP (*Extended Multum In Parvo - Normalized Cylindrical Projection*) é uma técnica empregada para tornar o processo de busca de correspondências ao longo das linhas epipolares mais eficiente através de duas bases, a normalização cilíndrica e o MIP-MAP (*Multum In Parvo Map*), Williams (27).

Este é uma técnica muito usada na área de computação gráfica para tornar o uso de texturas mais ágil. A idéia é que se tenha uma imagem original da textura e uma série de imagens reduzidas dela, que são empregadas conforme a necessidade, por exemplo, a resolução da textura pode diminuir de acordo com o tamanho do objeto ou com a distância do visualizador. Originalmente as imagens eram escaladas por potências de 2, assim, supondo uma textura original 256x256 as sub-texturas teriam tamanhos 128x128, 64x64, ..., 1x1, o que aproveita otimizações das bibliotecas gráficas como OpenGL (28) e DirectX (29), mas isso nem sempre é adequado do ponto de vista estético. Para dar melhor suporte a elementos de formato não quadrado foram inventados os RIP-MAPs (*Rectangular Multum In Parvo Map*),

Figura 3.6 – Mapeamento cilíndrico, adaptada de Timm (11)

Heckbert (30), que são MIP-MAPs retangulares. Timm cria MIP-MAPs retangulares a partir das imagens de normalização cilíndrica e usa estes conjuntos como base para uma busca eficiente de correspondências no par estéreo. Na implementação para a tese, esta estrutura era empregada apenas para executar buscas na sub-imagem de tamanho mais adequado, porém existe a possibilidade de fazer uma busca em diversos níveis de detalhamento, caso a melhora nos resultados obtidos justifique seu custo computacional. A Figura 3.7 mostra exemplos dos diversos tipos de otimização de texturas usadas por Timm para melhorar a busca por correspondências. Em 3.7a é mostrado um exemplo de MIP-MAP simples. 3.7b exemplifica o MIP-MAP retangular, e 3.7c mostra o uso de MIP-MAPs sobre a normalização cilíndrica da imagem original.

Perturbações de pontos chave

A busca por correspondências no par epipolar também é feita de maneira diferente por Seitz (4) e Timm (11). Enquanto no primeiro trabalho é feita uma busca completa, o segundo se aproveita da relativa uniformidade dos objetos de interesse para buscar a posição de pontos-chave e, a partir deles, estimar a alteração de uma linha epipolar para a linha equivalente da segunda imagem. São usadas as bordas das orelhas e os contornos do nariz como pontos-chave.

A vantagem desta abordagem é que o número de comparações necessárias é menor, e é possível pré-calcular os deslocamentos(em *pixels*) necessários na memória. Como a busca de correspondências nas linhas epipolares é de vital importância para o funcionamento da técnica de *view morphing* e é o principal gargalo do sistema como um todo, esta otimização aliada ao XMIP-NPC são os responsáveis pela possibilidade de execução do *morphing* em taxas interativas.

Figura 3.7 – Abordagens de múltipla resolução, adaptada de Timm (11)

3.4.2 Relação com o Projeto

Esse trabalho tem grande relação com o projeto. Em Timm (11) foi implementada uma versão do que se espera para o sistema que gera imagens neste trabalho, a parte da Janela 3D responsável por capturar os vídeos e gerar a imagem intermediária de acordo com a posição do usuário que é informada pelo sistema de visualização.

Houve um grande avanço em relação à possibilidade de execução do *view morphing* em tempo real, quando consideradas algumas restrições em relação ao ambiente e ao objeto de interesse. De qualquer maneira as principais otimizações desenvolvidas por Timm são conceitualmente válidas em diversos domínios e podem ter implementações equivalentes em diferentes aplicações. Além disso foi feito um estudo prático a respeito de dispositivos de captura que serve para validar hipóteses levantadas durante o desenvolvimento deste projeto e diminuem o tempo gasto para obter resultados já conhecidos.

3.5 Microfacet Billboarding

Neste trabalho Yamazaki et al. (31) propõem uma técnica para representação virtual realista de cenas reais que procura equilibrar o uso de informação geométrica e gráfica. A idéia principal é empregar em conjunto outras técnicas amplamente divulgadas para apresentar resultados de boa qualidade renderizados em taxas interativas. Para isso é empregado conhecimento prévio a respeito da geometria e texturas da cena adquiridos em uma primeira fase, para posteriormente executar a renderização de um modelo menos complexo.

A primeira fase da técnica não é executada em tempo real. Ela é responsável por obter conhecimento a respeito da geometria da cena, o que pode acontecer por meio de modelagem, escaneamento 3D, reconstrução ou outros métodos, e de suas características visuais, através de câmeras. Uma vez que se conhece a estrutura da cena e se possui um conjunto de imagens usadas como textura é gerado um modelo 3D de *voxels*, cujas cores são alteradas de acordo com o ponto de vista renderizado e com as texturas.

Modelos tridimensionais de subdivisão espacial, representados por matrizes de *voxels*, são computacionalmente pesados, o que compromete o desempenho de aplicações que as utilizem como meio de visualização. Para contornar esta característica a segunda fase da técnica, responsável pela renderização do modelo adquirido na fase anterior, aproxima cada *voxel* visível por um pequeno plano, cuja coloração é alterada dinamicamente de acordo com o ponto de vista, para que o resultado seja consistente com as texturas adquiridas da cena real. Buscando obter maior uniformidade na imagem renderizada e coerência com o mundo real, estes planos são tratados como *billboards*, esgotando a possibilidade de que um observador se mova a câmera virtual o suficiente para parecer estar vendo uma tela de cinema, e não

um modelo tridimensional, ou consiga enxergar a separação entre os planos. A Figura 3.8 ilustra a diferença das possíveis abordagens. Na Figura 3.8a é apresentada uma secção da aproximação de um objeto real, representado pela linha senoidal, por *voxels*, representados pelos segmentos de reta. Segmentos de reta também são empregados na Figura 3.8b para representar a aproximação do mesmo objeto por planos. É possível notar como uma tomada lateral revelaria a distância entre estes planos, além de gerar uma imagem incoerente. A solução adotada para este problema é representada na Figura 3.8c, onde são ilustrados os *billboards* e é possível notar que o observador tem a impressão de estar frente a uma superfície ininterrupta.

Figura 3.8 – Diferentes abordagens para aproximação de elementos reais, adaptada de Yamazaki et al. (31)

3.5.1 Relação com o Projeto

A técnica apresentada por Yamazaki et al. (31) é, assim como o presente projeto, focada em gerar imagens realistas usando o menor processamento possível. Sua aplicação neste projeto é restringida por sua primeira fase, uma vez que é indesejável que seja necessário conhecer a estrutura da cena, ou passar por processos longos de calibração.

Uma adaptação possível é fazer uso de alguma ferramenta que consiga gerar mapas de profundidade em taxas interativas em conjunto com vídeo capturado em tempo real. Desta maneira não seria possível gerar um modelo 3D, mas uma representação 2.5D cuja visibilidade é limitada pelo campo de visão dos dispositivos de captura de vídeo. Esta é uma técnica prevista para uso no projeto Vídeo Avatar do Interlab (1).

3.6 Considerações Finais

Foram apresentados trabalhos que têm relacionamentos diversos com a Janela 3D. Siscoutto e Tori (9) abordam sistemas de telepresença segundo uma visão mais abrangente e mostram em uma aplicação real os benefícios de uma apresentação visual mais realista para

o avatar. Nakamura e Tori (21) apresentam um sistema que permite interação natural entre um avatar e o ambiente virtual em que este se insere. Além da similaridade da aplicação, os requisitos de *hardware* e as técnicas empregadas são bem próximas das da Janela 3D o que reforça a possibilidade de integração entre os sistemas. Seitz (4) cunhou o termo *view morphing* e desenvolveu toda a base teórica para a interpolação de visões. Timm (11) desenvolveu diversas otimizações para um domínio restrito que se assemelha em muito ao da Janela 3D, além de técnicas que podem ser adaptadas a novos sistemas de *view morphing*.

4 O Sistema Janela 3D

Esta Seção apresenta a formalização do Sistema Janela 3D. Nela são apresentados sua arquitetura e seus componentes, assim como são descritos os relacionamentos entre eles e seus fluxos de informações. A colaboração dos conceitos revisados no Capítulo 2 para a concepção do Sistema como um todo também é explorada aqui, de maneira segmentada, sob uma visão sistêmica. A composição do produto final é analisada de forma a facilitar a compreensão de seu funcionamento e permitir clareza conceitual na participação de cada componente para sua obtenção.

A Seção 4.1 discorre a respeito do conceito da Janela 3D. Os requisitos do Sistema são apresentados pontualmente na Seção 4.2. Em 4.3 é apresentada a arquitetura proposta e seus componentes são mais bem explicados nas Seções 4.3.1, 4.3.3 e 4.3.2. Na Seção 4.5 são feitas as considerações finais deste Capítulo.

4.1 Conceito

A Janela 3D procura estender os atuais sistemas de videoconferência provendo maior realismo e coerência dimensional na imagem apresentada ao observador. Esta melhoria é atingida permitindo que a imagem remota mostrada ao usuário não seja sempre a mesma, mas se adapte à sua posição, em oposição aos sistemas tradicionais. Desta maneira o observador tem uma sensação de imersão maior, uma vez que a resposta do vídeo remoto à sua movimentação é coerente com o que aconteceria se ele observasse uma cena real através de uma janela. Para prover esta funcionalidade aos usuários da Janela 3D são necessárias mudanças na concepção dos sistemas de video-conferência tradicionais.

O lado do visualizador deixa de ser apenas uma plataforma de observação, e passa a ser uma interface de entrada e saída do sistema. A funcionalidade principal, de apresentação de imagens, é mantida, porém necessita-se conhecer a posição do observador em relação à tela e assim agrega-se um sistema de rastreamento a esta extremidade do sistema. Esta composição permite que o usuário consiga expressar seu desejo de que a imagem que lhe é apresentada seja alterada sem que seja necessário executar alguma tarefa especial. A naturalidade da interação é extremamente desejável e o rastreamento provê transparência, uma vez que permite fechar um ciclo: o sistema só deve se atualizar quando o observador se move e o observador só espera que o sistema se atualize em decorrência de uma movimentação sua. Considerando o aspecto

ergonômico, é de grande valia a capacidade de um sistema de rastreamento ser independente de elementos que devam ser acoplados à pessoa.

Já a parte responsável por gerar as imagens para o observador passa a operar em três dimensões, ao passo que em sistemas comuns uma imagem bidimensional da câmera já é adequada, além de também receber dados, quando um sistema tradicional apenas os envia. O aumento de dimensões se dá para que haja a possibilidade de mudança do ponto de vista de acordo com a posição do observador que é recebida, pois seria impossível manter a coerência com o ambiente real tridimensional a partir da imagem plana de uma câmera.

Consolidando as diferenças entre sistemas tradicionais de videoconferência e a Janela 3D destaca-se que esta é sensível ao ponto de vista do observador. Em decorrência desta característica aspectos particulares do funcionamento de subsistemas necessitam de alterações para fornecerem a base técnica requerida.

4.2 Requisitos

Os requisitos de sistema da Janela 3D são definidos a partir dos objetivos apresentados na Seção 1.1. Muito embora o todo do sistema seja considerado nos objetivos, os requisitos de sistema definidos são apresentados de maneira segmentada, para cada subsistema, com o intuito de facilitar a sua localização no processo de desenvolvimento. Nas Seções seguintes cada subsistema será introduzido e serão explanados os materiais e métodos empregados para atender aos requisitos.

Pontualmente definidos, os requisitos definidos para o Sistema de Visualização são:

1. Deve exigir o mínimo possível de equipamentos agregados ao usuário, preferencialmente nenhum;
2. O usuário deve ter movimentação livre dentro do espaço que consegue ser rastreado;
3. Deve prover três graus de liberdade: translação em X , Y e Z ;
4. Maximizar o uso de componentes de prateleira, que possam ser adquiridos com facilidade pelos usuários;
5. Facilidade de uso, implicando na independência de procedimentos complexos de calibração.

Já o Sistema de Geração possui os seguintes requisitos:

1. Deve ser possível gerar imagens sintéticas com a câmera virtual em qualquer ponto da linha que liga as câmeras reais, a partir dos dados de posição do usuário recebidos do Sistema Visualização;

2. As imagens geradas devem transmitir ao usuário sensação de realismo, ou seja, devem ser fiéis à cena verdadeira;
3. O sistema de geração deve fazer criar novas imagens considerando, ao menos, o deslocamento do usuário na direção horizontal. Se possível, também na vertical.

Em vista das necessidades discutidas na Seção 4.1 e dos requisitos aqui definidos é proposta para o sistema a arquitetura descrita na Seção 4.3.

4.3 Arquitetura

Nesta Seção é apresentada uma proposta de arquitetura para desenvolvimento do Sistema Janela 3D. São levados em consideração aspectos de distribuição geográfica dos componentes, sua colaboração para a composição do todo, coesão dos recursos utilizados e o acoplamento dos subsistemas.

A partir da descrição dos objetivos, assim como analisando o conceito apresentado para o Sistema Janela 3D pode-se notar uma segmentação semântica em sua composição. É clara uma divisão inicial em duas partes, uma responsável por exibir as imagens e detectar a posição do observador e outra à qual compete gerar as imagens espacialmente coerentes. Tal divisão permite uma utilização racional dos recursos e permite isolar gargalos e fazer análises de cada componente de maneira individual. Mantendo a simplicidade da arquitetura é introduzido um componente de comunicação, responsável por controlar os fluxos de dados do sistema e consolidar seu funcionamento.

Desta maneira é proposta uma arquitetura composta por três subsistemas de funções e escopos bem definidos. O primeiro é denominado Visualização e é responsável por apresentar as imagens ao observador e rastrear sua posição; ao subsistema de Geração compete gerar as imagens espacialmente coerentes; os fluxos de dados entre os dois subsistemas são controlados por um terceiro denominado Comunicação. A Figura 4.1 ilustra as relações e o escopo de cada subsistema.

Figura 4.1 – Visão geral da arquitetura do sistema

Nas Seções seguintes serão apresentadas as características básicas de cada subsistema tratando-os como componentes coesos e individuais.

4.3.1 *Subsistema de Visualização*

O subsistema Visualização tem três atribuições:

Exibição: apresentar as imagens geradas ao observador recebidas do subsistema Comunicação;

Rastreamento: rastrear a posição do observador;

Interfaceamento: enviar ao subsistema Comunicação a posição do observador.

A exibição é feita de maneira convencional, em uma plataforma de visualização como monitor, projetor, etc. O interfaceamento deve seguir o protocolo da camada de aplicação definido pelo subsistema Comunicação no momento em que envia a ele a posição atual do observador e quando recebe as imagens para exibição.

A atribuição mais relevante, e a que definitivamente o distingue dos sistemas de exibição tradicionais, é o rastreamento. Neste âmbito, o termo “rastrear” tem o significado de conhecer a posição de um objeto de interesse (no caso, o observador) em cada ciclo de atualização do sistema. Como posto na Seção 2.4 existem diversos tipos de solução para este problema, cada qual com seus pontos fortes e fracos.

Analizando os requisitos, especialmente a necessidade de apenas três graus de liberdade e a preferência por não acoplar dispositivos ao observador, pode-se eleger um sistema baseado em visão computacional como solução de custo baixo, fácil de implementar e que atende na medida exata os requisitos principais. Além disso, dentre as opções apresentadas, ainda há a possibilidade de usar rastreamento sem marcadores.

Adotando-se como base a visão computacional o funcionamento do sistema de rastreamento é ilustrado na Figura 4.2. Uma câmera é capaz de observar um determinado segmento do espaço, representado pelo tronco de pirâmide. Nesta condição, rastrear o observador significa conhecer a posição (x, y) de sua cabeça no espaço de coordenadas bidimensional do plano da imagem, denominado α na Figura, e a posição de α no sistema de coordenadas do mundo, z . Para facilitar a instanciação deste ambiente e diminuir as restrições de movimentação a posição em Z do plano de imagem pode ser definida um fator escala em relação a um valor inicial. Obtendo-se os valores da posição da cabeça do observador no sistema de coordenadas da imagem e a posição do plano de imagem no sistema de coordenadas do mundo consegue-se um ponto (x, y, z) o que permite atender o requisito de três graus de liberdade.

Figura 4.2 – Conceito de rastreamento do usuário

4.3.2 Subsistema de Geração

Existem diversas técnicas que podem ser empregadas para obter imagens espacialmente coerentes a partir de um ponto de vista conhecido. A mais difundida consiste em criar um modelo 3D da cena, aplicar texturas fotorrealistas e apresentar renderizações desta cena virtual. Apesar de eficiente, esta técnica requer conhecimento prévio sobre o ambiente, não é flexível e demanda procedimentos complexos de calibração e ajuste para que a imagem do elemento de interesse seja corretamente inserida no ambiente virtual. Além disso questões como diferença na iluminação da cena real e da cena virtual podem se tornar grandes fontes de inconsistência.

Técnicas de reconstrução 3D como as empregadas em Carranza et al. (32) e Magnor e Goldlucke (33) conseguem obter bons resultados, porém dependem de montagens especiais de *hardware* e têm alto custo computacional. O uso de imageamento estéreo para obter mapas de profundidade, como em, Nakamura e Tori (21), em conjunto com técnicas 2.5D, como Yamazaki et al. (31), representam uma solução que atende quase todos os requisitos, porém a geração do mapa de profundidades é, por si só, um processo computacionalmente custoso. A geração do avatar 2.5D a partir desta informação e o mapeamento de texturas inviabilizam sua aplicação em taxas interativas. Além disso, por particularidades da técnica de *microfacet billboards* a área em que o observador pode se mover sem enxergar discontinuidades nas imagens apresentadas é bastante restrita.

Dentre as técnicas apresentadas, a fusão de visões Seitz (4) se revela a mais adequada. Em sua versão mais simples depende de apenas duas câmeras cuja relação entre os parâmetros extrínsecos seja conhecida e de um processo simples de calibração para gerar imagens espacialmente coerentes de qualquer ponto que pertença à linha base da montagem. É também escalável, na medida em que é possível criar uma instalação com diversas câmeras tratando-as aos pares, e desta maneira tem-se apenas uma quantidade maior de problemas com solução conhecida para resolver antes de fazer o sistema Geração funcionar. Finalmente, não requer conhecimento prévio sobre a estrutura da cena e consegue lidar com questões de iluminação e realismo da imagem final, uma vez que se baseia apenas em pares de imagens capturadas da

cena real.

Adotando o uso da técnica de fusão de visões, o subsistema de Geração tem três atribuições:

Controle de câmeras e captura de imagens: acessar, controlar e sincronizar duas ou mais câmeras

Geração: gerar a imagem espacialmente coerente de acordo com o ponto de vista do observador recebido do subsistema Comunicação;

Interfaceamento: enviar a imagem gerada ao subsistema Comunicação.

O interfaceamento deve seguir o protocolo da camada de aplicação definido pelo subsistema Comunicação no momento em que envia a ele a imagem gerada e quando recebe aposição do observador. O controle das câmeras e a captura de imagens consiste em acessar fisicamente o conjunto de câmeras empregado e obter suas imagens, usadas como entrada para gerar as imagens espacialmente coerentes.

A atribuição de gerar as imagens coerentes é o que distingue o sistema de Geração da Janela 3D de sistemas análogos de outros sistemas de video-conferência. Sua finalidade é sintetizar imagens de pontos de vista quaisquer (na linha base da montagem de câmeras) a partir das imagens de câmeras reais, sem conhecimento prévio a respeito da estrutura física da cena. Tendo como entradas a posição do observador e o conjunto de imagens capturadas em um determinado instante o subsistema de Geração identifica qual é o par de imagens mais adequado como base para gerar a imagem sintética e renderiza uma visão intermediária da cena, que é enviada ao subsistema de Visualização.

4.3.3 Subsistema de Comunicação

O subsistema de comunicação tem as atribuições de:

Controle do fluxo de dados: coordenar a comunicação entre os subsistemas Visualização e Geração;

Definição do protocolo: estabelecer uma linguagem comum seguida pelos outros componentes da Janela 3D.

A Comunicação controla a troca de informações entre os outros componentes, que podem tanto estar sendo executados em computadores distintos a grandes distâncias quanto em um mesmo computador. Como recurso para isso ela define um protocolo de aplicação específico para troca de informações de imagem e posição, os dois tipos de dados usados na Janela 3D. Por pressupor a existência de uma infra-estrutura de rede de alta velocidade (Internet 2) não há preocupação com variações na taxa de transmissão, congestionamentos ou controle de

entrega em nível de aplicação. Além de usar uma plataforma de alta capacidade, a natureza da aplicação faz com que, em casos extremos, seja preferível perder dados do que arcar com o custo do reenvio.

Para fins de exemplo, consideremos que a transmissão de um quadro renderizado é atômica. Em caso de perda restariam duas opções: parar a exibição para o observados enquanto o(s) quadro(s) perdido(s) são reenviados, o que causaria congelamentos do vídeo e atraso, ou exibir os quadros que não foram perdidos e quando o(s) quadro(s) perdido(s) fossem recebido(s), exibi-lo(s), o que causaria incoerência temporal e visual. Em contrapartida, se o que foi perdido for apenas considerado pode haver um pequeno congelamento do vídeo, mas não haverá incoerência nem atraso, o que justifica a opção por esta abordagem.

4.4 Extensões e Integração

A arquitetura proposta é flexível para permitir extensões e integração com outros sistemas através de seus pontos de conexão. Desta maneira dá-se subsídio à continuidade do projeto, assim como à sua aplicação como parte de sistemas maiores, como o que está sendo desenvolvido no Projeto Video-Avatar do Interlab (1). Sugestões de extensão são feitas no Capítulo 8, e um teste de integração feito informalmente que gerou resultados expressivos foi o uso de visão estereoscópica a partir de visões virtuais.

4.5 Considerações Finais

Foi apresentada uma formalização do Sistema Janela 3D, que estabelece seu escopo, componentes e relações. Foram analisadas alternativas para desenvolvimento dos componentes e escolhidas as melhores opções dentro da proposta de cada um. Finalmente foram apresentadas as possibilidades de extensão e integração que a arquitetura proposta permite.

5 Experimentos Práticos

Nesta Seção serão descritos os experimentos práticos realizados para testar hipóteses e avaliar a viabilidade técnica das soluções encontradas. Estes experimentos foram elaborados de maneira semanticamente convergente em relação ao produto final Janela 3D. É importante destacar que o processo de desenvolvimento do trabalho como um todo compreende os experimentos aqui relatados, além de muitos cuja descrição foi suprimida para manter a concisão do texto, além de alguns não terem contribuição decisiva na evolução do trabalho.

Os resultados apresentados nas Seções 5.1, 5.4.1, 5.4.2 e 5.4.3 foram obtidos em um computador que possui as seguintes especificações: processador AMD (*American Micro Devices*) Athlon 64 X2 3600+, RAM (*Random Access Memory*) de 2GB (Gigabytes) DDR2 (*Double Data Rate 2*), câmera Logitech QuickCam 4000 e sistema operacional Linux *kernel* versão 2.6.24-19. Os testes foram feitos com imagens de resolução 320x240 *pixels*.

O teste da Seção 5.2 utilizou computador com processador AMD Athlon 64 X2 3600+, RAM de 2GB DDR2, sistema operacional Linux *kernel* versão 2.6.27-11 e imagens de resolução 320x240 *pixels* (interpoladas), ou 768x576 *pixels*, capturadas com uma câmera Sony DXC390P.

Para o teste apresentado na Seção 5.3 foram utilizados os seguintes equipamentos: computador com processador AMD Athlon 64 X2 3600+, RAM de 2GB DDR2, câmeras LG modelo LVC-S60HM com lentes auto-íris DC 3.5–8mm F1.4 e sistema operacional Linux *kernel* versão 2.6.27-11. Foram usadas com imagens de resolução 720x480 *pixels*.

A Seção 5.1 introduz um estudo de viabilidade inicial, descrevendo seus objetivos, processo de construção, base teórica e resultados obtidos. O teste de uso de um número maior de câmeras é relatado na seção 5.2 enquanto o uso de vídeo foi avaliado na Seção 5.3. Na Seção 5.4 são apresentados os testes realizados com diferentes sistemas de rastreamento de cabeça.

5.1 Estudo de Viabilidade do *View Morphing*

5.1.1 *Proposta*

O estudo de viabilidade buscava condensar o conhecimento levantado a respeito de *view morphing* em um programa que permitisse realizar testes iniciais e funcionasse como um arcabouço para experimentação de otimizações e propostas técnicas. A partir de uma implemen-

tação básica pretendia-se identificar o relacionamento entre os componentes mais importantes do sistema e isolar os gargalos de desempenho e qualidade visual das imagens.

5.1.2 Justificativa

Uma consolidação inicial das técnicas sugeridas é de vital importância no início da pesquisa. Além de validar a proposta dos trabalhos base, permite ter uma visão mais aprofundada do seu funcionamento. Além disso, o resultado deste desenvolvimento inicial é a base de *software* que permite a evolução dos demais componentes em direção à Janela 3D.

5.1.3 Base Teórica

Este experimento baseia-se nos trabalhos apresentados na Seção 3, especialmente em Seitz e Dyer (6) e Seitz (4), que introduziram a técnica de *view morphing* a partir de quaisquer montagens de câmeras.

5.1.4 Resultados

Os resultados obtidos com a prova de conceito foram satisfatórios. As imagens sintetizadas não apresentavam a qualidade visual esperada, porém permitiram observar uma série de aspectos relativos ao comportamento do algoritmo empregado. De modo geral elas eram compostas por áreas onde era possível distinguir claramente elementos das duas imagens base e por zonas de transição, nas quais havia mistura de linhas de *pixels* da imagem esquerda e da imagem direita sem, contudo, formar um trecho visualmente consistente. As Figuras 5.1, 5.2 e 5.3 ilustram os resultados obtidos com diferentes imagens base. Em 5.1d e 5.1e são mostradas as imagens base direita e esquerda, respectivamente. A Figura 5.1a mostra uma visão a distância 0.25 do epípólo esquerdo; a Figura 5.1b, visão no centro das duas visões base, a distância 0.5; finalmente, a Figura 5.1c a visão foi posicionada a distância 0.75. Nas Figuras 5.2 e 5.3 as imagens estão dispostas da mesma forma.

É possível notar que conforme a visão gerada computacionalmente se aproxima de um dos epípólos, os trechos contínuos são formados principalmente por elementos da imagem que o contém. Além disso nota-se que as descontinuidades ocorrem na direção das linhas epipolares, o que levanta a hipótese de instabilidade dos algoritmos empregados em certas regiões críticas das imagens.

Outra característica identificada durante os testes com a prova de conceito foi a influência

Figura 5.1 – Resultados da prova de conceito. Em 5.1a visão a distância 0.25. Em 5.1b visão equidistante dos epipólos e em 5.1c visão gerada a distância 0.75, todas em relação às imagens base 5.1d e 5.1e, esquerda e direita, respectivamente.

Figura 5.2 – Resultados da prova de conceito com pilha de toalhas. Em 5.2a visão a distância 0.25. Em 5.2b visão equidistante dos epipólos e em 5.2c visão gerada a distância 0.75, todas em relação às imagens base 5.2d e 5.2e, esquerda e direita, respectivamente.

Figura 5.3 – Resultados da prova de conceito com planta. Em 5.3a visão a distância 0.25. Em 5.3b visão equidistante dos epíolos e em 5.3c visão gerada a distância 0.75, todas em relação às imagens base 5.3d e 5.3e, esquerda e direita, respectivamente.

da distribuição dos pares de pontos equivalentes na qualidade final da imagem, como brevemente citado em Seitz (4). As imagens sintéticas geradas apresentam menos distorção nas áreas próximas aos pontos que foram usados como correspondências para o cômputo da matriz fundamental. Este efeito é ainda mais acentuado quando os pontos correspondentes são definidos dentro de uma área reduzida. As Figuras 5.4a e 5.4c mostram os pares de pontos equivalentes usados para calcular a matriz fundamental que foi usada para gerar a figura 5.4b, uma visão com distância 0.5. Pode-se perceber como a área próxima aos pares de pontos sofreu pouca distorção.

Figura 5.4 – Efeito da distribuição dos pontos correspondentes

A Figura 5.5b mostra o efeito causado pela concentração dos pontos equivalentes em uma pequena área das imagens base 5.5a e 5.5c. De modo semelhante ao caso anterior, a imagem

Figura 5.5 – Efeito da concentração dos pontos correspondentes

gerada é quase inalterada na região próxima aos pontos equivalentes e distorcida em direção aos epíolos em regiões afastadas dos ponto equivalentes.

5.1.5 Discussão

Os resultados obtidos no estudo de viabilidade foram interessantes sob diversos aspectos. Apesar de as imagens geradas não terem atingido a qualidade desejada foi possível obter parâmetros em relação ao custo computacional e identificar pontos do sistema onde cabem melhorias.

Em uma avaliação subjetiva, as imagens obtidas têm menor qualidade visual que as apresentadas em Seitz (4). As principais diferenças são em relação à continuidade de trechos de cor semelhante na imagem. Nota-se que pequenas manchas da cor de objetos aparecem desconexas destes e também há partes da imagem gerada em que trechos das duas imagens base aparecem intercalados, causando má apresentação visual.

É importante destacar que neste experimento buscou-se obter a base de resultados empregando a menor quantidade de recursos possível. Assim, seguiu-se à risca a recomendação de que apenas oito pontos correspondentes são necessários para obter a matriz fundamental, ao passo que Seitz (4) relata o uso de algumas dezenas deles nos resultados obtidos. Quando emprega-se 21 pontos são obtidos resultados melhores, como os mostrados na Figura 5.6.

Nota-se, portanto, que além da concentração dos pontos correspondentes também há relação entre a quantidade deles e a qualidade dos resultados obtidos. Isto deve-se principalmente à instabilidade do método numérico usado para resolver o sistema linear que obtém a matriz fundamental. Além disso os resultados eram obtidos na casa de centésimos de segundos, enquanto Timm (11) reporta que os resultados de Seitz (4) levavam alguns segundos para serem gerados.

Figura 5.6 – Resultados obtidos com 21 pontos correspondentes

5.2 Uso de Maior Quantidade de Câmeras

5.2.1 Proposta

Este teste busca investigar qual é o ganho que pode ser obtido pelo aumento do número de pares de câmeras, sem alteração na técnica de interpolação de visões. Procura-se identificar qual é possibilidade de aumentar a qualidade gráfica do resultado, o campo de visão e analisar o impacto disso no tempo de processamento.

5.2.2 Justificativa

Apesar de promissores, os resultados obtidos no teste descrito na Seção 5.1 ficaram aquém do esperado. Para um sistema que pretende fornecer vídeo em nível de qualidade de televisão é importante identificar maneiras de obter melhores respostas.

5.2.3 Base Teórica

Os resultados apresentados na Seção 5.1 mostram que a técnica empregada é válida,

porém é sensível ao distanciamento das câmeras, assim como variações entre a imagem obtida por cada uma delas. A proposta de usar várias câmeras permite desmembrar um problema grande em pequenos problemas de escopo reduzido e com condições em que foi mostrado que a interpolação de visões obtém resultados de maior qualidade. Para uma mesma distância original entre as câmeras base, a inclusão de apenas uma câmera permite reduzir pela metade a porção do espaço que deve ser contemplada por cada par de câmeras, o que representa uma simplificação importante do desafio inicial. No tocante às particularidades do algoritmo de interpolação de visões, há duas interpretações.

Caso seja mantido o comprimento da linha-base de uma montagem inicial, incluir uma nova câmera implica em converter esta linha base em duas, cada uma com a metade do comprimento anterior. Isso significa que as câmeras de um dado par estéreo ficarão fisicamente mais próximas e terão um campo de visão mais semelhante, além de haver maior correspondência nas imagens capturadas em relação a questões de iluminação e fatores do ambiente.

Outra possibilidade é alterar o comprimento da linha base inicial, assim a adição de uma câmera passa a significar uma nova linha-base, cujo comprimento é igual ao dobro da original. Com isso é possível gerar visões intermediárias de elementos que não podiam ser observados na montagem inicial.

5.2.4 Resultados

Os resultados obtidos nos testes efetuados, analisados de maneira subjetiva, foram satisfatórios. Em comparação com a prova de conceito apresentada na Seção 5.1 é possível notar um ganho de qualidade visual significativo pela adição de apenas uma câmera. A Figura 5.7 compara uma visão do ponto 0.25 gerada com duas câmeras 5.7a e com três 5.7b. Nos dois casos a linha-base abrange um ângulo de 30° . É possível perceber como a imagem gerada a partir de um número maior de câmeras apresenta menos artefatos.

Mais além, os resultados obtidos com a inclusão de um número maior de câmeras (mantendo o aproximadamente o comprimento total da linha-base) não apresentaram grande melhora sobre os realizados com três. A seguir são apresentados alguns exemplos que ilustram o que foi relatado. A Figura 5.8 apresenta um exemplo do uso de dois pares de câmeras, cujas imagens são 5.8a, 5.8b e 5.8c. Uma visão sintética do ponto 0.45 da linha-base composta pelas linhas-base dos dois pares de câmeras (abrangendo 30°) é mostrada na Figura 5.8d. Pode-se notar que uma comparação da Figura 5.8d com a Figura 5.9, gerada a partir de 12 câmeras abrangendo um total de 55° , não revela um salto de qualidade proporcional à quantidade de câmeras adicionadas.

A Figura 5.10d compara uma visão de 0.45 gerada a partir de três imagens (5.10a, 5.10b e 5.10c) com uma imagem renderizada do mesmo ponto, mostrada na Figura 5.11. Também

(a) Visão de 0.25 gerada a partir de duas câmeras

(b) Visão de 0.25 gerada a partir de três câmeras

Figura 5.7 – Comparação da geração de visões

é possível notar a semelhança do resultado sintético com o renderizado.

Outra questão relevante que deve ser discutida é a alteração no desempenho e no tempo de reposta quando uma quantidade maior de câmeras é usada. Muito embora haja um aumento no consumo de memória, não foi observada queda na taxa de atualização ou no tempo de resposta. Este era o comportamento esperado para o sistema, uma vez que não há mais

(a) Câmera esquerda

(b) Câmera central

(c) Câmera direita

(d) Visão sintética para a posição 0.45 da linha-base composta

Figura 5.8 – Amostra de resultado obtido com três câmeras abrangendo 30°

Figura 5.9 – Amostra de resultado obtido com doze câmeras abrangendo 55°

processamento a ser feito simultaneamente em razão da adição de câmeras à montagem. Isto se deve ao fato de só ser requerida uma atualização da imagem gerada por um dado par de câmeras quando o ponto de vista do observador está no trecho da linha-base coberta por ele. Desta maneira sempre é processado apenas um par de imagens não importando quantos existam.

(a) Câmera esquerda

(b) Câmera central

(c) Câmera direita

(d) Visão sintética para a posição 0.45 da linha-base composta

Figura 5.10 – Amostra de resultado obtido com três câmeras abrangendo 30°

Figura 5.11 – Renderização da cena vista do ponto 0.45

5.2.5 Discussão

Os resultados obtidos com imagens calibradas mostram que é possível obter ganho de qualidade pela adição de câmeras à montagem. Mostram também que o ganho não cresce na mesma medida do número de câmeras, havendo uma grande melhoria pelo uso de três câmeras quando antes se usava apenas duas, porém não havendo diferença tão expressiva entre conjuntos de três ou de doze câmeras. Desta maneira um número grande de câmeras se justifica apenas com o objetivo de ampliar o campo de visão, não havendo compromisso direto com a qualidade das imagens geradas.

Consolidando os resultados obtidos, este experimento permite identificar que o fator preponderante para a qualidade gráfica das imagens geradas é a semelhança entre as imagens que são observadas pelas câmeras reais. Esta afirmação fundamenta-se no fato de se obter melhores resultados sempre que as câmeras são fisicamente aproximadas, tanto quando se mantém o comprimento da linha-base quanto quando ele é aumentado, porém a quantidade de câmeras empregada faz com que a distância entre elas não seja grande.

5.3 Interpolação de Visões em Vídeo

5.3.1 Proposta

O intuito deste teste é submeter o arcabouço do Sistema Janela 3D a condições reais de utilização. Desta maneira é possível avaliar seu funcionamento e sua resposta às condições do ambiente, assim como quantificar aspectos relevantes do ponto de vista do usuário leigo. Além disso, em uma montagem real há uma série de variáveis que são introduzidas no problema que

se deseja resolver advindas, por exemplo, da inclusão de novos dispositivos de captura.

5.3.2 *Justificativa*

É necessário conhecer o comportamento do sistema frente a um cenário real de utilização. Muitas vezes características são ocultadas por conjuntos especialmente bons de amostras para teste, ou por um comportamento viciado dos testadores.

Além disso, uma vez que é inserido em um contexto de utilização maior ele deixa de ser uma peça técnica isolada e passa a compor um conjunto maior, devendo interagir de maneira adequada com usuários e produzir respostas que sejam condizentes com o esperado, independente de o desafio técnico de sua implementação ter sido ultrapassado.

5.3.3 *Base Teórica*

Este teste compartilha a base teórica com os descritos nas Seções 5.1 e 5.2. O que o diferencia são as condições do teste. Enquanto os dois primeiros eram focados em validar uma técnica matemática, este busca identificar o relacionamento desta técnica e dos dispositivos que integram o Sistema com um usuário leigo e com o ambiente em que está inserido.

5.3.4 *Resultados*

A inclusão de novos dispositivos no Sistema trouxe consigo novas características que se mostraram de extrema relevância ao longo do período de testes. A configuração utilizada, baseada em câmeras base diferentes, mostrou que, apesar de passível de utilização, há uma série de aspectos que devem ser considerados para que a Janela 3D possa atingir desempenho adequado.

As câmeras utilizadas conseguem adquirir imagens em taxas maiores, de maior resolução e estabilidade que *webcams*. Entretanto tratam-se de dispositivos focados em monitoramento de ambientes e, embora permitam ajustes importantes como cancelamento de controle automático de ganho, do balanço de branco, abertura da íris e compensação da luz de fundo, não possuem regulagem individual da curva de resposta para cada cor. Somando-se a isso a diferença de sensibilidade do CCD de cada câmera chega-se a um ponto de partida para compreender o comportamento do Sistema no teste em ambiente real. As Figuras 5.12a e 5.12b mostram imagens de uma mesma cena capturadas pelas duas câmeras, é possível notar a diferença das cores dos objetos nas duas imagens. Outras característica que pode ser

Figura 5.12 – Diferença de cor entre as câmeras

notada é a grande distorção radiométrica introduzida pelas lentes. Nas bordas das imagens é possível notar como elementos sabidamente retos, como armários e calhas de passagem de cabos, aparecem curvados.

A variação de resposta às cores e a distorção radiométrica conseguem tornar o problema muito mais complexo. Como foi dito na Seção 3.1, em dado momento o algoritmo de interpolação busca equivalências de cores sobre as linhas epipolares. Esta busca é afetada pelas duas características indesejáveis das câmeras. O problema mais óbvio decorre de o mesmo elemento não ser retratado com a mesma cor nas duas imagens, logo, a busca precisa ser flexível a variações tonais para conseguir alguma correspondência entre as imagens.

Além disso, toda a teoria matemática da geometria epipolar é baseada no modelo de câmera de orifício, que pressupõe que o plano retinal seja efetivamente plano, ou seja, o modelo não considera distorções como a radiométrica introduzida pelo conjunto de lentes. O efeito disso é que quanto mais perto das bordas da imagem maior é a chance de não se encontrar uma correspondência adequada. Isso devido ao fato de a busca ser feita ao longo de uma linha epipolar determinada analiticamente e considerando um modelo plano, porém amostrando *pixels* que foram capturados com distorção, em uma imagem onde linhas retas são, na verdade, curvas. A Figura 5.13 ilustra este efeito com a imagem de um tabuleiro capturada por uma das câmeras utilizadas.

5.3.5 Discussão

Os resultados permitiram identificar quais foram as diferenças de resposta do Sistema quando empregado em um cenário real de utilização. Foram identificados os principais fatores limitadores de qualidade, de maneira que é possível sugerir medidas para que seus efeitos sejam atenuados. De maneira que condiz com o posto na Seção 5.2, estes limitadores também são relacionados à semelhança das imagens, porém de maneira que não pode ser corrigida pelo simples posicionamento adequado das câmeras. O Capítulo 6 consolida os resultados deste

Figura 5.13 – Efeito da distorção radiométrica

teste e sugere maneiras de superar as dificuldades introduzidas pelo uso das câmeras reais.

5.4 Rastreamento de Cabeça

Esta Seção fala sobre as propostas de sistemas para rastrear a posição da cabeça do usuário do sistema de visualização. No todo da Janela 3D estes são os responsáveis por detectar a posição de quem visualiza remotamente o resultado do *view morphing*, fornecendo ao sistema de geração dados para gerar a projeção de maneira coerente.

Foram experimentadas três diferentes abordagens, duas baseadas em marcadores e uma que não necessita de nenhum dispositivo acoplado ao usuário. Na Subseção 5.4.1 é descrito o rastreamento baseado em marcadores coloridos; em 5.4.2 relatam-se as experiências com marcadores fiduciais e na Seção 5.4.3 são descritas as evoluções no sentido de livrar o usuário de objetos acoplados.

5.4.1 Marcadores Coloridos

Proposta

Neste experimento procurava-se desenvolver um sistema de rastreamento de três graus de liberdade, que fosse eficiente, de baixo custo e de uso simples. Estes são requisitos adequados a um usuário doméstico que deseje experimentar a Janela 3D ou ainda usá-la sem que seja necessário obter equipamentos ou marcadores especiais.

Justificativa

O rastreamento baseado em cores atende plenamente os requisitos do subsistema de visualização. O marcador mostrado na Figura 5.14 é uma instância deste tipo de ferramenta, baseada em etiquetas encontradas em papelarias. Manchas de tinta verde e vermelha(ou de outras cores, se o rastreador for parametrizado para detectá-las) em um papel branco teriam o mesmo efeito prático. Por empregar técnicas e materiais próximos do cotidiano do usuário a simplicidade de uso é atingida. O custo dos materiais empregados permite criar marcadores de preço irrisório. A necessidade de três graus de liberdade também é suprida com a informação obtida de um marcador deste tipo.

Base Teórica

Sistemas de rastreamento ou de sobreposição e mistura de cenas baseados em cores são largamente utilizados, como em programas televisivos ou em diversos trabalhos, entre eles Calife (34), Sanches (35). Apesar de objetivos distintos, a técnica de detecção de cores empregada nas ferramentas de *chroma keying* e em rastreadores é semelhante. As implementações mais comuns se baseiam em limiarização em canais específicos dos espaços de cor utilizados, como *RGB* ou *HSV*. Bergh e Lalioti (36) sugere uma abordagem baseada na geometria do cubo *RGB* que minimiza o número de operações e é geralmente mais robusta.

Resultados

Foi criado um sistema de rastreamento baseado nas cores verde e vermelho. A escolha destas cores se deu pelo fato de serem facilmente diferenciáveis entre si e pela facilidade de se obter objetos cuja intensidade destes tons seja grande. Seria possível obter os mesmos resultados usando apenas uma cor, porém o uso de duas torna a implementação mais simples e não representa um aumento significativo na dificuldade para replicar instâncias do marcador. A Figura 5.14 mostra uma versão de marcador adaptada a uma câmera. Esta montagem foi usada para produzir o vídeo de exemplo disponível em Trias (37). A detecção se baseia em limiarização de canais *RGB* devido à facilidade de implementação.

A dinâmica de utilização do sistema de rastreamento deve seguir determinados passos na ordem correta para que se obtenha resultados satisfatórios. Inicialmente é necessário calibrar limiares para adequar as intensidades tonais detectadas às condições de iluminação do ambiente. Com isso já é possível detectar a posição do marcador no sistema de coordenadas do plano de imagem da câmera. Após isso o usuário deve posicionar-se a uma distância confor-

Figura 5.14 – Marcador colorido posicionado ao redor da lente de uma câmera

tável da câmera para que seja definida a distância base entre o elemento vermelho e o verde. A partir deste momento a movimentação paralela ao eixo óptico da câmera também pode ser detectada comparando-se a distância atual entre os elementos coloridos com a que foi definida como base. Desta maneira consegue-se suprir a necessidade de três graus de liberdade.

Discussão

O sistema de rastreamento implementado conseguiu atingir os requisitos especificados. O algoritmo é sensível às condições de iluminação e a variações no sistema de captura. Como em todo processo de segmentação por cores, o usuário não pode usar roupas e nem o ambiente pode ter elementos da cor que está sendo rastreada. A calibração executada para cada ambiente consegue adequar o sistema às condições de iluminação na maioria dos casos, porém a sensibilidade a elementos da cor rastreada é incontornável.

Seria possível tornar esta ferramenta mais robusta otimizando algoritmos de segmentação e usando marcadores mais eficientes, por exemplo, não passivos. Tais possibilidades não foram exploradas pois, tratando-se de um sistema dependente de marcadores, há a possibilidade de uso de marcadores fiduciais que têm melhor desempenho no geral. Além disso considerou-se mais importante investir esforços em produzir uma solução sem marcadores, para tornar a interação mais natural.

5.4.2 Marcadores Fiduciais

Proposta

A idéia neste teste é criar uma outra ferramenta de rastreamento que atenda os mesmos requisitos de baixo custo, eficiência e de prover ao menos três graus de liberdade que o sistema baseado em marcadores coloridos, porém com menos sensibilidade às condições de iluminação e com maior estabilidade.

Justificativa

O uso de marcadores fiduciais é uma proposta alternativa em relação aos marcadores coloridos. O custo e os materiais envolvidos são semelhantes para ambos. Uma ressalva em relação aos fiduciais é que a grande maioria das pessoas não é familiarizada com este tipo de ferramenta e é fácil que incorram em erros simples como não conservar uma borda ao redor do desenho ou sobrepor outros elementos aos marcadores.

Base Teórica

Marcadores fiduciais são soluções de rastreamento largamente usadas em sistemas que privilegiam o baixo custo ou que por algum motivo não podem empregar outros tipos de rastreadores, magnéticos, por exemplo. Bibliotecas que implementam rastreamento de fiduciais, como ARToolkit (38), ARTag (39) e Tracker (40) são empregadas em diversos projetos, entre eles de Miranda et al. (41), Woods et al. (42), Billingham e Kato (43), Wagner et al. (44), Wagner et al. (45).

O fato de serem compostos por elementos de cores com muito contraste entre si (usualmente preto e branco) torna a detecção dos marcadores fiduciais mais fácil por sistemas de visão, ainda quem não sejam empregadas câmeras de alta qualidade. A Figura 5.15 mostra tal contraste. Além disso tais sistemas conseguem prover os três graus de liberdade necessários, e também rotação nos três eixos.

Resultados

O objetivo de suprir os mesmos requisitos do protótipo de rastreamento por marcadores coloridos, mas com maior estabilidade foi atingido. Graças ao grande contraste dos componentes do marcador e à implementação mais elaborada dos algoritmos de detecção foi possível obter a posição da cabeça do usuário sem flutuações. Além de permitir obter a translação em relação à origem do sistema de coordenadas do plano retinal da câmera também é possível obter a orientação relativa, o que, por hora, não é necessário, mas pode ser empregado em trabalhos futuros.

Figura 5.15 – Exemplos de marcadores fiduciais, ARToolkit (38)

Discussão

Há diferenças entre as ferramentas de rastreamento de marcadores fiduciais. Algumas são otimizadas para reduzir o processamento, outras buscam robustez ou até facilidade de uso no tocante à implementação. O teste foi realizado com base no ARToolkit (38), que busca ser fácil de usar, ainda que sua detecção não seja tão eficiente. A oclusão de uma parte do marcador, mesmo que pequena, faz com que este não seja mais detectado. Este efeito é especialmente indesejável quando a pessoa se aproxima dos limites da partição do espaço real que é vista pela câmera, já que quando a primeira parte do marcador deixa de aparecer este deixa de ser detectado, porém ainda resta espaço para que a pessoa se mova. O ARTag (39) e o Tracker (40) são soluções que buscam ser mais robustas do que fáceis de usar. O uso de marcadores codificados como códigos de barras bidimensionais permite que estes sejam reconhecidos mesmo quando grande parte está ocluída.

Tais ferramentas não foram testadas devido a dois motivos principais. A experiência anterior dos autores indica que elas são perfeitamente adaptáveis ao cenário de uso em discussão e que a codificação do programa adaptado não tem complexidade muito maior do que a já elaborada. Além disso, muito embora dêem margem a expansões no mecanismo de interação com o usuário e sejam estáveis, os marcadores fiduciais deve ser fixados no usuário. Como se dá preferência por sistemas que não precisem de nenhum dispositivo especial optou-se por investir esforços de pesquisa em soluções *markerless* como os classificadores em cascada aplicados à detecção de faces.

5.4.3 *Classificadores Características Haar-like*

Proposta

Criar um sistema de rastreamento que não necessite de marcadores no usuário, que permita obter a posição do usuário no sistema de coordenadas da câmera. A atualização destas posições deve ser feita em um intervalo de tempo que permita taxas interativas e deve haver estabilidade na detecção. Também não deve ser necessário empregar aparelhagem de alto custo ou que seja difícil de obter.

Justificativa

O sistema de geração da Janela 3D não precisa conhecer dados complexos a respeito do usuário, sendo suficiente a sua posição em relação à câmera. Tal dado pode ser obtido de diversas maneiras, usando técnicas e equipamentos diversos. Do ponto de vista da ergonomia do sistema, é desejável que a interação com o usuário se dê da forma mais natural possível tanto no momento em que este recebe estímulos quanto quando os envia. No contexto da visualização de uma cena tridimensional remota com correção de projeção esta naturalidade é associada à sensação de imersão do usuário, que não deve ser quebrada por mudanças bruscas na projeção ou por alterações em velocidade diferente da que ocorre no mundo real.

Já quando o usuário envia estímulos o principal fator que torna a interação mais natural é fazer com que este envio ocorra de maneira transparente. Ou seja, o usuário não deve ter que informar o sistema de que está querendo enviar um estímulo, o sistema deve ser capaz de detectar a intenção do usuário e responder prontamente. A mudança do ponto de vista é um caso em que este tipo de comportamento é relativamente fácil de implementar. Naturalmente as pessoas movimentam a cabeça quando querem ver a cena de uma outra perspectiva, por exemplo, para ver algo que está atrás de um obstáculo visual. Assim o rastreamento da cabeça do usuário, e não de elementos fixados a ela, se revela a solução ideal para obter interação natural. Para teste foi eleita a técnica de classificadores em cascata de características Haar-like devido à disponibilidade de bibliotecas prontas, à eficiência na detecção e ao tempo de execução relativamente baixo.

Base Teórica

Foram empregados classificadores de características *Haar-like* em cascata. Foi usado o método proposto em Lienhart e Maydt (46) que incrementa a base conceitual criada em Papageorgiou et al. (47) e estendida por Viola e Jones (48). O funcionamento é baseado em treinamento supervisionado de uma cadeia de classificadores que depois são aplicados a novas amostras e retornam um resultado binário que significa a pertinência ou não da amostra ao conjuntos de objetos apresentados como treinamento.

Características *Haar-like* são aquelas determinadas por contrastes na imagem. Elas são mais robustas que a análise *pixel-a-pixel* pois carregam informação da relação entre as regiões da imagem, e não apenas comparam valores pontuais. Elas podem, por exemplo, detectar arestas, linhas ou cantos. A Figura 5.16 mostra exemplos de características *Haar-like* retangulares.

Figura 5.16 – Exemplos de características *Haar-like*, adaptada de Viola e Jones (48)

O uso de conjuntos destas características permite expressar a estrutura de elementos mais complexos como contornos de olhos ou boca, por exemplo. Quando conjuntos deste tipo são empregados em cascata consegue-se identificar elementos ainda mais complexos como, no caso, o rosto. Existem diversos trabalhos que empregam características *Haar-like* para detectar diferentes tipos de elementos, assim como pesquisas que aliam classificadores a outras técnicas, por exemplo Menezes et al. (49) que rastreia faces usando também *eigenfaces* para identificá-las e Sirovich e Kirby (50). Já Hong et al. (51) alia classificadores a *Hidden Markov Models* para detectar e reconhecer faces.

Resultados

O sistema de rastreamento baseado em classificadores desenvolvido conseguiu satisfazer os requisitos especificados. Pela própria concepção não são usados marcadores e a posição do usuário é atualizada em taxas satisfatórias. Notou-se porém que é possível melhorar tal taxa

para que haja maior disponibilidade de informação.

O principal ponto em que cabe melhoria é a detecção da distância do usuário à câmera. Em momentos de estabilidade nas condições de captura, por exemplo com a câmera capturando uma fotografia de rosto impressa, em ambiente de iluminação controlada, nota-se que as estimativas para a área(em *pixels*) do rosto detectado sofrem uma flutuação ao redor do valor real. Esta área é usada da mesma maneira que a distância entre as marcas verde e vermelha apresentadas na Seção 5.4.1, capturando-se um valor inicial e calculando-se a movimentação relativa. Desta maneira, a flutuação na estimativa da área do rosto detectado provoca mudanças bruscas na translação da cena virtual, o que causa diminuição na sensação de imersão do usuário. A Figura 5.17 mostra a variação dos valores estimados ao redor da média obtida de um conjunto de 100 amostras capturadas sob iluminação constante e sem movimento relativo entre a foto de rosto rastreada e a câmera. Esta variação deve-se principalmente a variações tonais introduzidas pela própria câmera.

Figura 5.17 – Variação da detecção da profundidade por classificadores de características Haar-like em relação à média de 100 amostras

Discussão

Os resultados obtidos com classificadores foram satisfatórios. Embora seja desejável obter melhores taxas de atualização, o nível atingido já permite o uso em aplicações interativas. Além disso há otimizações simples que podem ser feitas no algoritmo existente para diminuir a área de busca e consequentemente o tempo gasto nela. Uma maior taxa de atualização permite também empregar filtragem para atenuar a flutuação e até pré-estimar a área que será detectada para o rosto. É importante destacar que tanto a diminuição da área de busca quanto aplicação de filtragem nas estimativas para a área do rosto não são maneiras de resolver os problemas encontrados, mas sim meios de contorná-los dentro de um domínio específico.

Seguindo esta linha de raciocínio foi feita uma comparação entre a oscilação dos valores detectados e o tempo de detecção para a imagem original e para uma imagem sobre a qual foi aplicada um filtro passa-baixa (duas iterações de filtro média com máscara 7x7).

Seja $F = (f_1, f_2, \dots, f_{100})$ um conjunto de cem quadros capturados de uma fotografia de rosto. A Figura 5.18 apresenta o tempo de detecção com e sem filtragem para cada quadro de F . É possível notar que apesar do aumento no tempo gasto, a oscilação dos valores detectados sofre uma diminuição desejável para o sistema de rastreamento. Além de outras possíveis otimizações em relação ao tamanho da imagem usada para detectar o rosto, o que pode equilibrar o tempo gasto na filtragem, também é importante citar que o efeito de um filtro passa-baixas pode ser obtido desfocando-se a câmera adequadamente, eliminando assim o acréscimo de custo computacional introduzido pela filtragem.

Na Figura 5.19 são apresentadas as variações $v(f_i)$ das estimativas para posição em Z, Y e Z, ao redor da média dos valores obtidos para cada eixo, dadas pela expressão:

$$v(f_i) = \mu_f - f_i$$

Figura 5.18 – Comparação do tempo gasto na detecção de um rosto com e sem filtragem

O principal ponto em favor desta abordagem é a naturalidade da interação. Desde a etapa inicial de calibração não é exigido qualquer ajuste ou procedimento não natural à vida do usuário. A detecção da posição do usuário em relação aos eixos X e Y do plano retinal é feita sem que haja necessidade de qualquer ajuste, bastando que este se posicione em frente à câmera. Para calibrar a medida de profundidade basta que ele se posicione na posição que será mantida como original e pressione uma tecla. A partir daí o usuário não precisa usar mais nenhum dispositivo anexado a ele ou manipular dispositivos de entrada tradicionais.

Figura 5.19 – Comparação da variação da detecção em X(5.19a), Y(5.19b) e Z(5.19c) com e sem filtro passa-baixas

5.5 Considerações Finais

Foram testadas duas grandes categorias de sistemas de rastreamento, os baseados em marcadores e os sem marcadores, aplicados ao rastreamento de cabeça. Dentro de cada categoria também foram feitas comparações entre as alternativas existentes visando buscar a melhor opção.

A solução com melhor taxa de atualização e maior estabilidade foram os marcadores fiduciais. Os marcadores coloridos atingiram resultados próximos, porém são mais instáveis. A opção pelo classificador de característica *Haar-like* se deu principalmente pela capacidade que eles têm de livrar o usuário de equipamentos acoplados, embora seja menos estável que os marcadores fiduciais. A Tabela 5.1 condensa os resultados dos testes efetuados. A interpretação para a quantidade de estrelas (★) é em relação à qualidade, assim, mais estrelas significam um resultado melhor e menos estrelas um resultado pior.

Os gráficos apresentados nas figura seguintes relacionam os resultados obtidos com as técnicas testadas. Na Figura 5.20 são comparadas as flutuações dos valores detectados para objetos de interesse estáticos detectados em ambiente controlado. A estabilidade do marcador fiducial é notável em relação às demais soluções testadas, apontando-o como melhor opção para caso sejam empregados marcadores.

Na Figura 5.21 são comparados os tempos gastos para a detecção dos objetos rastreados por diferentes métodos. Novamente nota-se a eficiência do algoritmo de detecção do marcador fiducial.

Tabela 5.1 – Comparação de sistemas de rastreamento

	E.A.	Est.	S.I.	T.A.	N.C.	F.R.
Marcadores Coloridos	★ ★	★ ★ ★	★ ★	★ ★ ★ ★ ★	★ ★ ★	★ ★ ★ ★
Marcadores Fiduciais	★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★	★ ★ ★
Classificadores de Características Haar-like	★ ★ ★ ★ ★	★ ★ ★ ★	★ ★ ★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★

Legenda:

E.A. : Equipamentos Acoplados

Est. : Estabilidade

S.I. : Sensibilidade à Iluminação

T.A. : Taxa de Atualização

N.C. : Necessidade de Calibração

F.R. : Facilidade de Reprodução

(a)

(b)

(c)

Figura 5.20 – Comparação da variação da detecção em X(5.20a), Y(5.20b) e Z(5.20c) entre as diferentes soluções testadas

Figura 5.21 – Comparação do tempo gasto por diferente métodos de detecção

6 Implementação

Nesta seção serão descritos os aspectos relacionados à implementação do protótipo da Janela 3D. Nela são apresentadas as características funcionais existentes, assim como qual foi o percurso para atingi-las, dos pontos de vista práticos e conceitual.

6.1 Infra-Estrutura Física

Esta Seção trata da concepção e execução da infra-estrutura física necessária para o funcionamento da Janela 3D. Aqui são abordados os temas relacionados à confecção do *hardware* empregado e do *software* associado diretamente a ele.

6.1.1 Sistema de Visualização

A montagem desenvolvida para o Sistema de Visualização é mostrada na Figura 6.1. No protótipo implementado o Sistema de Visualização é baseado em monitor LCD (*Liquid Crystal Display*) e uma *webcam* Logitech modelo QuickCam 4000. Outras plataformas de visualização, como projetores, monitores convencionais, *displays* espaciais, *etc*, poderiam ser empregadas da mesma maneira, desde que seja possível associar uma câmera a elas. Óculos de realidade virtual são passíveis de utilização apenas se possuírem um sistema próprio de rastreamento, ou se for usado rastreamento de marcadores, uma vez que a face do observador fica ocluída pelo dispositivo.

6.1.2 Sistema de Geração

Câmeras

No protótipo foi usado um par de câmeras LG modelo LVC-S60HM com lentes auto-íris, como a mostrada na Figura 6.2, alimentado por uma fonte AC 24V que atua também como dispositivo de sincronização. Estas câmeras permitem, além da sincronização por *hardware*,

ajustes como cancelamento de balanço de branco, controle de ganho e compensação de luz de fundo automáticos. Por questão de praticidade foram capturados alguns vídeos estéreo e estes eram usados nos testes, no lugar das câmeras diretamente ligadas ao computador.

6.2 Software Relativo à Infra-Estrutura Física

Esta Seção fala sobre os programas que compõem a infra-estrutura para o funcionamento da Janela 3D e detalha os processos em que eles estão envolvidos.

6.2.1 Retificação de Imagens

Como foi dito na Seção 5.3 há uma discrepância geométrica entre o que o modelo matemático espera que a câmera capture e o que ela efetivamente consegue capturar. Tal característica leva o algoritmo de interpolação de visões a buscar correspondências em conjuntos de *pixels* que não representam os mesmos objetos, gerando resultados inconsistentes.

Figura 6.1 – Montagem do Sistema de Visualização

Figura 6.2 – Câmera usada no Sistema de Geração

Embora não seja possível evitar a distorção, as imagens da câmera podem ser retificadas para que os objetos planos mostrados nelas voltem a ser planos. Desta maneira volta a ser consistente a restrição epipolar sobre os conjuntos de pixels das imagens. Zhang (52) propõe uma técnica para obter os parâmetros intrínsecos de uma câmera a partir de um conjunto de pontos, que podem ser extraídos automaticamente. Para isso é necessário apresentar um padrão à câmera em diferentes direções. A Figura 6.3 mostra um conjunto de imagens usado para calibração de uma das câmeras da montagem.

Figura 6.3 – Amostras para calibração de câmeras

A retificação das imagens foi implementada como uma das etapas de sua captura, seguindo o algoritmo proposto por Zhang (52). A captura das imagens do padrão para uma dada montagem de câmeras é feita manualmente, e um programa calcula os parâmetros intrínsecos de cada uma, armazenando-os para posterior uso na Janela 3D. A Figura 6.4 mostra o efeito da retificação aplicada sobre a imagem de um padrão.

Figura 6.4 – Amostras para calibração de câmeras

6.2.2 Calibração de Convergência

A calibração de convergência consiste em um procedimento simples, porém necessário. Ele decorre dos resultados apresentados na Seção 5.2, no tocante à importância da semelhança das imagens capturadas pelas câmeras. O objetivo é maximizar o campo de visão comum às duas câmeras de um par estéreo. Para isso uma das câmeras é usada como padrão e a posição e orientação da outra é ajustada até que tenha a maior semelhança possível com a primeira.

Outros métodos poderiam ser empregados para isso, por exemplo realizar um corte da área comum das imagens, assim as câmeras poderiam ser dispostas de qualquer maneira desde que houvesse uma área em comum. Entretanto, a detecção de elementos da cena nas duas imagens e a estimativa das correspondências não são problemas de solução intuitiva e de qualquer maneira perderia-se resolução das imagens. A opção pelo procedimento manual foi feita pois o automatizado demandaria trabalho de desenvolvimento com um suposto, mas não provado, ganho de tempo na calibração e para obter resultados da renderização, na melhor das hipóteses, iguais aos da calibração manual.

6.2.3 Correção de Cor

Como apresentado na Seção 5.3 a diferença entre a resposta das câmeras para uma mesma cor representa o problema mais grave para a aplicação da Janela 3D em ambiente real. A diferença entre os matizes das duas imagens faz com que o algoritmo de interpolação não encontre correspondências no par estéreo e interpole o valor que for próximo, ocasionando descontinuidades na imagem gerada.

Há algumas opções para resolver ou contornar este problema. Além de *hardware* específico, como câmeras com opção de calibração de cor, equipamentos específicos dedicados a uniformizar e sincronizar imagens, é possível empregar soluções por processamento de imagem. O principal fator definitivo na escolha da técnica aplicada é definir qual será a referência. Havendo um padrão de cor, sem textura e com iluminação constante, é possível posicionar as câmeras exatamente no mesmo ponto e gerar uma *lookup table* que corrige cada canal, de cada câmera, fazendo com que as imagens tenham cores iguais para objetos iguais. O fator limitante desta técnica é que padrões de cor pura, sem textura, com iluminação constante não são facilmente encontrados.

Uma outra possibilidade é adotar uma das câmeras como padrão e corrigir as cores da segunda, para que gere o mesmo valor que a primeira. Esta técnica *ad hoc* pode ser efetiva em casos particulares, mas é bastante restrita. Além de não haver um padrão absoluto, é necessário decidir com base em heurísticas pouco definidas qual das câmeras será adotada como base. Por fim, em uma montagem com diversos pares de câmeras a chance de a heurística falhar se torna maior, e a propagação de erros pode chegar a invalidar o procedimento.

Pela falta de padrão de cor e pela fragilidade da técnica relativa, a calibração de cor não chegou a ser implementada de maneira definitiva. Entretanto foram feitas duas experiências com ferramentas que causam distorção de cor, o que é intolerável em um sistema que pretende gerar imagens realistas, mas permite restringir o escopo do problema a ser resolvido e validar a idéia de que a diferença de cores é a causa das distorções. A primeira tentativa foi usar imagens em tons de cinza, o que gerou resultados de maior qualidade. Uma opção seria

trabalhar apenas sobre o canal de luminância, mas de qualquer maneira seria uma imagem em tons de cinza, que não é válida para apresentação nos parâmetros da qualidade desejada. A outra experiência foi efetuar uma equalização de histograma¹, o que também trouxe melhoras significativas para os resultados, mas também provoca alterações drásticas de cor. Amostras dos resultados obtidos são mostradas e discutidas no Capítulo 7.

6.2.4 Obtenção das Matrizes Fundamentais

O método de *view morphing* depende de conhecimento a respeito da relação entre as câmeras, condensado na matriz fundamental. A maneira mais simples de se obter esta matriz é numericamente, resolvendo-se um sistema linear cujas entradas são posições (no sistema de coordenadas da imagem da câmera) de pontos equivalentes nas duas imagens. Desta maneira, obter a matriz fundamental de um par estéreo consiste em indicar um conjunto de, ao menos, oito (devido à dimensão do sistema linear) pontos equivalentes não coplanares em cada imagem.

Esta é outra situação em que o desenvolvimento de um método automatizado poderia ser mais custoso do que realizar a tarefa manualmente, sem a garantia de melhores resultados. Assim optou-se por fornecer ferramentas para que o usuário possa fazer a seleção dos pontos com mais precisão e conforto no lugar de criar uma ferramenta de obtenção automática de matrizes fundamentais.

6.2.5 Sistema de Comunicação

A infra-estrutura física de comunicação está fora do escopo deste projeto, uma vez em que ele se insere em um projeto maior que parte do pressuposto da existência de conexões de alta capacidade (internet 2) para desenvolver novas aplicações. Os requisitos de capacidade que a Janela 3D espera que sejam atendidos são proporcionais à resolução do vídeo usada, sendo equivalentes à um *streaming* de vídeo no trecho entre o Sistema Geração e o Sistema Visualização.

6.3 Artefatos da Engenharia de Software

Nesta Seção são apresentados os artefatos de engenharia de *software* do Sistema. Por se tratar de um protótipo focado na demonstração da viabilidade de uma técnica matemática

¹equalizar de histograma consiste em fazer com que a imagem que originalmente tinha uma determinada faixa dinâmica passe a abranger a faixa dinâmica desejada.

para uma determinada classe de aplicações optou-se por não percorrer todas as fases de um processo de desenvolvimento de *software*. Desta maneira, foi gerada uma documentação básica para permitir a compreensão da estrutura interna e da dinâmica de funcionamento de determinados trechos do programa, considerados os mais relevantes.

6.3.1 Diagrama de Classes

O diagrama de classes apresentado nas Figuras 6.5 e 6.6 mostra a estrutura em alto nível do Sistema Geração. A divisão em duas Figuras foi feita para facilitar a leitura do diagrama, sendo que originalmente a Figura 6.6 estava à direita da 6.5. A classe `app::Morpher` representa a aplicação que é executada. Ela é composta por gerenciadores das grandes partes deste sistema: `ui::UIManager`, que trata da interface gráfica, `io::IOManager`, que trata do acesso a câmeras e configurações e `morphing::MorphManager`, que é responsável por gerenciar a interpolação. Foi prevista a realização de testes com técnicas de interpolação diferentes do *view morphing* em possíveis continuidades do projeto, e para isso foi criada a interface `morphing::Morpher` que permite a criação de quaisquer tipos de interpoladores sem alteração na estrutura do sistema. Da mesma maneira a interface `io::FrameServer` permite que sejam criadas fontes de imagens para interpolação diversas, sendo já previstas fontes locais e remotas.

Figura 6.5 – Diagrama de Classes - parte 1

Figura 6.6 – Diagrama de Classes - parte 2

6.3.2 Diagrama de Sequência

O diagrama da Figura 6.7 apresenta uma iteração de atualização do sistema. É importante destacar que embora seja usada notação UML (*Unified Modeling Language*) os objetos cuja linha de vida é mostrada não representam instâncias de classes, mas sim elementos importantes do sistema.

Figura 6.7 – Sequência de tarefas de um ciclo de atualização do Sistema

6.4 Considerações Finais

Foi apresentado um relato da implementação do protótipo da Janela 3D. Aspectos levantados no Capítulo 5 foram analisados sob uma visão da prática de utilização do sistema. Também foi apresentada uma breve introdução à arquitetura interna do sistema, com destaque aos pontos em que novas funcionalidades ou técnicas distintas podem ser agregadas para comparação.

7 Resultados

Neste Capítulo são analisados os resultados obtidos ao longo do desenvolvimento do projeto. São recapitulados os objetivos iniciais e estes são comparados com o que se obteve ao longo do período de desenvolvimento. Também são analisados os requisitos estabelecidos para os subsistemas da Janela 3D e seu nível de atendimento.

7.1 Macro Objetivos

Como relacionado na Seção 1.1, os macro-objetivos do projeto são:

1. **Sistema Visualização:** Desenvolver um sistema de rastreamento de usuário que seja de uso fácil, permita movimentação livre e minimize o uso de equipamentos acoplados ao observador;
2. **Sistema Geração:** Criar um sistema que gere imagens sintéticas realistas a partir de duas ou mais imagens obtidas em vídeos de uma cena real, em taxas interativas;
3. **Sistema Comunicação:** Empregar protocolos de rede conhecidos em conjunto com um protocolo de aplicação desenvolvido especificamente para atender às necessidades de comunicação entre os sistemas citados;
4. **Protótipo:** Desenvolver um protótipo que permita provar o conceito e realizar testes que subsidiem futuros projetos nesta linha.

A seguir eles serão analisados individualmente face aos resultados atingidos e será definido se foram ou não atingidos.

7.2 Sistema Visualização

A Seção 5.4 apresenta uma discussão a respeito de técnicas que podem ser empregadas para rastreamento. Cada qual possui suas características positivas e negativas, em relação aos aspectos técnicos e ergonômicos importantes para esta aplicação. Recapitulando a Seção 4.2, os requisitos para o Sistema Visualização eram: exigir o mínimo de equipamentos acoplados,

permitir movimentação livre, prover três graus de liberdade e aplicar componentes de prateleira com facilidade de uso.

O rastreamento sem marcadores atende a todos os requisitos e tem estabilidade aceitável para uma aplicação interativa. Tal afirmação baseia-se em taxas de atualização de sistemas de rastreamento com função semelhante e principalmente na transparência do sistema de rastreamento durante testes informais com usuários, ou seja, não foi possível notar a existência do sistema por nenhum tipo de atraso ou comportamento incomum. Os métodos baseados em marcadores atendem parcialmente o requisito de minimizar a quantidade de dispositivos acoplados ao observador, porém são mais leves, portáteis e adaptáveis que outros sistemas de rastreamento que precisam de marcadores. O nível de instabilidade do rastreamento de face é compensado, com sobra, pelo uso do marcador fiducial, sendo estas as duas opções mais recomendadas.

Sendo o desenvolvimento do Sistema Visualização um objetivo do projeto, e tendo todos os seus requisitos sendo, ao menos, parcialmente atendidos pelas opções apresentadas considera-se que o primeiro objetivo foi alcançado com sucesso.

7.3 Sistema Geração

Os testes realizados para desenvolvimento do Sistema Geração foram apresentados nas Seções 5.1, 5.2 e 5.3. O desenvolvimento se deu de maneira direta, inicialmente uma prova de conceito da técnica validou a possibilidade de uso. Em sequência confirmou-se que era possível obter ganho de qualidade pela adição de câmeras e que este ganho não é proporcional à quantidade de câmeras adicionada, sendo possível obter melhoras significativas pela adição de apenas uma. O último teste tinha como objetivo submeter o *software* obtido com os dois anteriores a um cenário real de utilização. Pelo significado prático dos resultados obtidos e pela característica decisiva das conclusões a que ele levou optou-se por fazer uma análise mais aprofundada nesta Seção.

O teste com vídeos de câmeras distintas permitiu reforçar a hipótese de que a semelhança das imagens capturadas por um par estéreo é um fator preponderante para qualidade visual da imagem sintetizada. Embora possam ser configuradas de maneira que as cenas retratadas sejam geometricamente próximas, a resposta dos sensores das câmeras não é igual para um mesmo sinal e desta maneira as cores são retratadas de maneira diferente de uma imagem para outra. O efeito disso pode ser visto na Figura 7.1, uma visão gerada do ponto médio da linha base entre as visões 5.12a e 5.12b. É possível notar como a busca por correspondências falha em determinados locais, gerando trechos descontínuos na imagem.

Duas medidas imediatas podem ser tomadas para melhorar a qualidade da imagem gerada: tornar os tons das imagens mais semelhantes e compensar a distorção radiométrica introdu-

Figura 7.1 – Efeito da interpolação de imagens com tons diferentes para objetos iguais

zida pela lente. Embora seja uma fator de deterioração matematicamente demonstrável, a discrepância entre o modelo de câmera de orifício representa um problema menor do que a diferença entre as cores no contexto da interpolação de visões. A Figura 7.2 mostra uma visão gerada do mesmo ponto que a anterior, porém empregando imagens reprojetadas. É possível notar que há menos descontinuidades, principalmente perto das bordas da imagem, porém o resultado ainda é insatisfatório.

Figura 7.2 – Efeito da interpolação de imagens com tons diferentes para objetos iguais, com correção da distorção radiométrica

As Figuras seguintes comparam a melhoria que é possível aproximando as cores e fazendo a correção da distorção radiométrica. Em 7.3a foi feita apenas uma equalização de histograma nas imagens e é possível notar como o resultado é melhor que o obtido com a imagem geometricamente corrigida, mas com diferenças de cor. Já a Figura 7.3b mostra o ganho que se obtém com a correção da distorção radiométrica sobre imagens equalizadas. há um aumento de qualidade, porém é incomparável com o que se consegue corrigindo as cores.

Embora tenha ocorrido uma melhora expressiva no resultado, o procedimento de equalização provoca alteração nas cores, gerando impressões não naturais, o que não é tolerável em um sistema que pretende gerar imagens realistas. Mesmo não sendo uma solução, o teste da

Figura 7.3 – Resultados de diferentes procedimentos de adequação das imagens à técnica empregada

equalização permite isolar o problema no escopo da semelhança entre as cores das imagens.

Os requisitos definidos para o Sistema Geração são objetivos: deve ser possível gerar imagens de qualquer ponto na linha-base, as imagens geradas devem ser fiéis à cena original e ele deve conseguir se adaptar à posição do observador. A geração de visões a partir de qualquer ponto da linha base é inerente à técnica de interpolação de visões. A partir daí é simples fazer com que o Sistema Geração se adapte à posição do observador, que é informada a ele pelo Sistema Visualização.

Uma discussão maior deve ser feita acerca do requisito a respeito da fidelidade à cena real. Do ponto de vista do usuário da aplicação este requisito foi parcialmente atendido, pois as imagens geradas são fiéis em termos de cor, porém apresentam trechos descontínuos e algumas vezes criam situações fisicamente impossíveis. Por outro lado, analisando o ponto de vista do desenvolvedor, este requisito foi atendido, ficando a melhoria do resultado obtido a cargo do uso de melhores equipamentos, que podem ser adquiridos no mercado, mas não estavam disponíveis para teste. Um exemplo da existência de solução prática pode ser dado por cenas de televisão em que mais de uma tomada é feita. Em um jogo futebol, por exemplo, são usadas diversas câmeras, mas não se nota diferença na cor do campo quando suas imagens são trocadas.

Desta maneira, uma vez que o teste relatado na Seção 5.1 validou a técnica e o da Seção 5.2 mostrou como gerar resultados de alta qualidade a partir de dados mais adequados à técnica empregada considera-se que o requisito de fidelidade à cena original foi atendido, ficando uma demonstração condicionada à disponibilidade de equipamentos existentes e amplamente utilizados, que apenas não puderam ser usados durante os testes feitos com vídeo. Assim, o segundo objetivo do projeto foi também atingido.

7.4 Sistema Comunicação

Por pressupor a existência de uma infra-estrutura de comunicação de alta capacidade, o Sistema Comunicação tem como requisito conseguir comunicar dados da posição do usuário e a imagem sintetizada. Além de simples, estes requisitos já foram atendidos muitas vezes, em muitos projetos. Desta maneira considera-se que o terceiro objetivo do projeto foi atingido com sucesso.

7.5 Desenvolvimento do Protótipo

O protótipo evoluiu do código gerado nos testes, tendo sua arquitetura definida com foco em flexibilidade e capacidade de integração, uma vez que as técnicas se mostraram aplicáveis. As imagens apresentadas nesta dissertação foram criadas com o protótipo desenvolvido, assim considera-se que o quarto e último objetivo do projeto foi atingido com sucesso.

7.6 Considerações Finais

Foram apresentados os resultados obtidos ao longo do desenvolvimento do projeto de maneira a proporcionar uma visão de sua evolução. Os objetivos iniciais foram analisados e comparados com o estado atual de desenvolvimento do trabalho, chegando-se à conclusão de que foram todos atingidos, faltando apenas uma demonstração prática do atendimento de um requisito de um dos subsistemas.

8 Conclusão e Trabalhos Futuros

Nesta dissertação foi apresentada a concepção e o desenvolvimento da Janela 3D, um sistema de telecomunicação visual inovador pelo fato de permitir visão tridimensional sensível ao ponto de vista do observador. A partir de um extenso levantamento bibliográfico foram definidos objetivos focados na obtenção de um protótipo que permitisse provar o conceito detalhado na descrição do Sistema e servir como base para a continuidade das pesquisas na área no Interlab (1). Experimentos práticos permitiram identificar quais são os pontos-chave em cada um dos componentes da Janela 3D, assim como sua influência na elaboração do produto final.

Muito embora os macro-objetivos tenham sido alcançados, é possível elencar aspectos em que melhorias são necessárias e tornarão o uso da Janela 3D mais simples e efetivo. Dentre as opções de sistema de rastreamento de usuário feitas os principais valores considerados foram facilidade de uso e taxa de atualização das estimativas. Marcadores fiduciais conseguem maiores taxas de atualização ao passo que obrigam o usuário a acoplar equipamentos a si, diminuindo a naturalidade da interação e alterando sua aparência. Já classificadores de características Haar-*like* independem de qualquer dispositivo acoplado, mas têm taxas de atualização menores e menos estáveis. Uma direção de melhoria é o aprimoramento do rastreamento do usuário permitindo maior linearidade e menores maiores taxas de atualização em soluções sem marcadores.

A geração de visões intermediárias mostrou-se uma solução prática e efetiva para o problema de apresentação de imagens com coerência tridimensional de ambientes dos quais não se conhece *a priori* nem estrutura e nem a aparência. Por outro lado a natureza do método revelou-se sensível a condições de captura comuns, exigindo determinados tipos de equipamento para que a qualidade de seu resultado em ambiente controlado seja obtida também em ambientes quaisquer. Outro aprimoramento que pode tornar a Janela 3D mais robusta é o uso de dispositivos físicos e técnicas de calibração e correção que aumentem a verossimilhança da reprodução de uma mesma cor nas imagens das câmeras utilizadas.

De maneira geral os objetivos principais foram plenamente atendidos, sendo que cabem otimizações e melhorias em pontos particulares de cada componente. Foi desenvolvida uma arquitetura de *software* que facilita a extensão e integração deste em sistemas maiores. É prevista a integração da Janela 3D no sistema desenvolvido no projeto Vídeo Avatar do TIDIA-Ae.

8.1 Trabalhos Futuros

Este trabalho abriu as portas para desenvolvimento de novas pesquisas e aplicações para a técnica por ele validada. Como sugestões de continuidade e extensão podem ser relacionadas:

- realizar testes que validem a hipótese de que com câmeras cujas respostas de cor podem ser reguladas o sistema consegue gerar resultados de alta qualidade também para vídeo;
- testar as possibilidades do uso das infinitas câmeras virtuais em sistemas de visão estereoscópica;
- integrar a Janela 3D ao Vídeo Avatar do projeto TIDIA-Ae.

8.2 Agradecimentos

O desenvolvimento deste projeto não seria possível sem o apoio para aquisição de infraestrutura fornecido pela FAPESP através do projeto TIDIA-Ae, processo número 2005/60618-1, e pela bolsa fornecida pelo CNPq sob o processo número 131531/2007-7.

Referências Bibliográficas

- [1] INTERLAB. Homepage do Laboratório de Pesquisa de Tecnologias Interativas. Disponível em: <<http://www.pcs.usp.br/~interlab/>>. Acesso em: 09 de fev. de 2009.
- [2] FAUGERAS, O.. **Three-dimensional computer vision: A Geometric Viewpoint**, 1st ed. MIT Press. 1993. 695 p.
- [3] HARTLEY, R.; ZISSERMAN, A.. **Multiple view geometry in computer vision**: ISBN: 0521540518, 2nd ed. Cambridge University Press. 2004.
- [4] SEITZ, S. M.. **Image-based transformation of viewpoint and scene appearance**. 1997. Tese(Doutorado) - University of Wisconsin - Madison.
- [5] HEYDEN, A.; POLLEFEYS, M.. **Emerging topics in computer vision**, 1st ed. Gerard Medioni and Sing Bing Kang. 2004.
- [6] SEITZ, S. M.; DYER, C. R.. **View morphing**. In: Conference on Computer Graphics and Interactive Techniques, 23. **Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques**, 1996, New York, NY, USA. ACM. p. 21-30.
- [7] Flock of Birds. Flock of Birds Homepage. Disponível em: <<http://www.ascension-tech.com/realtime/RTflockofBIRDS.php>>. Acesso em: 7 de fev. de 2009.
- [8] MILGRAM, P.; TAKEMURA, H.; UTSUMI, A.; KISHINO, F.. **Augmented reality: A class of displays on the reality-virtuality continuum**, SPIE Telemanipulator and Telepresence Technologies, v. 2351, p. 282-292.
- [9] SISCOOTTO, R. A.; TORI, R.. **Proposta de arquitetura para teleconferência baseada na integração de vídeo avatar estereoscópico em ambiente virtual tridimensional**. 2003. Tese(Doutorado) - Escola Politécnica da Universidade de São Paulo.
- [10] SECOND LIFE. Second Life Homepage. Disponível em: <<http://secondlife.com/>>. Acesso em: 5 de ago. de 2008.

- [11] TIMM, K. W.. **Real-time view morphing of video streams**. 2003. Tese(Doutorado) - University of Illinois.
- [12] BEIER, T.; NEELY, S.. **Feature-based image metamorphosis**, SIGGRAPH Comput. Graph., v. 26, n. 2, p. 35–42.
- [13] LEE, S.-Y.; CHWA, K.-Y.; SHIN, S. Y.. **Image metamorphosis using snakes and free-form deformations**. In: SIGGRAPH '95, . **Proceedings of the 22nd annual conference on Computer graphics and interactive techniques**, 1995, New York, NY, USA. ACM. p. 10.
- [14] WOLBERG, G.. **Digital image warping**. IEEE Computer Society Press. 1990.
- [15] CHEN, S. E.; WILLIAMS, L.. **View interpolation for image synthesis**. In: SIGGRAPH '93, . **Proceedings of the 20th annual conference on Computer graphics and interactive techniques**, 1993, New York, NY, USA. ACM. p. 10.
- [16] LAVEAU, S.; FAUGERAS, O.. **3-D scene representation as a collection of images and fundamental matrices**. Institut national de recherche en informatique et automatique, 1994. Relatório técnico número RR-2205.
- [17] TOMASI, C.; KANADE, T.. **Shape and motion from image streams under orthography: a factorization method**, International Journal of Computer Vision, v. 9, n. 2, p. 18.
- [18] OTT, M.; LEWIS, J. P.; COX, I.. **Teleconferencing eye contract using a virtual camera**. In: CHI '93, . **Conference companion on human factors in computing systems**, 1993, New York, NY, USA. ACM. p. 2.
- [19] POGGIO, T.; BRUNELLI, R.. **A novel approach to graphics**. 1992. Relatório técnico número AIM-1354.
- [20] WATT, A.. **3d computer graphics**. Pearson Education. 2000. 570 p.
- [21] NAKAMURA, R.; TORI, R.. **Vídeo-avatar com detecção de colisão para realidade aumentada e jogos**. 2008. Tese(Doutorado) - Escola Politécnica da Universidade de São Paulo. São Paulo - S.P. - Brasil.
- [22] GONZALEZ, R. C.; WOODS, R. E.. **Digital image processing**, 2nd ed. Prentice Hall. 2002.
- [23] SUBBARAO, M.; SURYA, G.. **Depth from defocus: A spatial domain approach**, International Journal of Computer Vision, v. 13, p. 271–294.

- [24] ZHANG, L.; NAYAR, S.. **Projection defocus analysis for scene capture and image display**, ACM Transactions on Graphics, v. 25, n. 3, p. 9.
- [25] LIAO, M.; WANG, L.; YANG, R.,; GONG, M.. **Light fall-off stereo**, IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.
- [26] BASRI, R.; JACOBS, D.,; KEMELMACHER, I.. **Photometric stereo with general, unknown lighting**, Int. J. Comput. Vision, v. 72, n. 3, p. 239–257.
- [27] WILLIAMS, L.. **Pyramidal parametrics**, SIGGRAPH Comput. Graph., v. 17, n. 3, p. 1–11.
- [28] OPENGL. OpenGL Homepage. Disponível em: <<http://www.opengl.org>>. Acesso em: 15 de jul. de 2008.
- [29] DIRECTX. DirectX Homepage. Disponível em: <<http://msdn.microsoft.com/en-us/directx/default.aspx>>. Acesso em: 15 de jul. de 2008.
- [30] HECKBERT, P. S.. **Survey of texture mapping**, IEEE Comput. Graph. Appl., v. 6, n. 11, p. 56–67.
- [31] YAMAZAKI, S.; SAGAWA, R.; KAWASAKI, H.; IKEUCHI, K.,; SAKAUCHI, M.. **Microfacet billboard**. In: EGRW '02, . **Proceedings of the 13th Eurographics workshop on Rendering**, 2002, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association. p. 169–180.
- [32] CARRANZA, J.; THEOBALT, C.; MAGNOR, M. A.,; SEIDEL, H.-P.. **Free-viewpoint video of human actors**, ACM Trans. Graph., v. 22, n. 3, p. 569–577.
- [33] MAGNOR, M.; GOLDLUCKE, B.. **Spacetime-coherent geometry reconstruction from multiple video streams**. In: , . **3DPVT '04: Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium**, 2004, Washington, DC, USA. IEEE Computer Society. p. 365–372.
- [34] CALIFE, D.. **Robot arena: uma infra-estrutura para o desenvolvimento de jogos com realidade aumentada espacial**. 2008. Dissertação(mestrado) - Escola Politécnica da Universidade de São Paulo. São Paulo - S.P. - Brasil.
- [35] SANCHES, S. R. R.. **A utilização da técnica do chroma key para composição de cenas em ambientes de realidade misturada**. Mar. 2007. 45. Dissertação(mestrado) - Centro Universitário Eurípedes de Marília. Marília - S.P. - Brasil. Mar.
- [36] BERGH, F. V. D.; LALIOTI, V.. **Software chroma keying in an immersive virtual environment**, South African Computer Journal, v. 24, p. 155–162.

- [37] LUCAS PADOVANI TRIAS. Demonstração do Rastreamento de Marcadores Coloridos. Disponível em: <<http://www.youtube.com/watch?v=9-MfMFght9w>>. Acesso em: 23 de jul. de 2008.
- [38] ARTOOLKIT. ARToolkit Homepage. Disponível em: <<http://www.hitl.washington.edu/artoolkit/>>. Acesso em: 23 de jul. de 2008.
- [39] ARTAG. ARTag Homepage. Disponível em: <<http://www.artag.net/>>. Acesso em: 24 de jul. de 2008.
- [40] STUDIERSTUBE TRACKER. Studierstube Tracker Homepage. Disponível em: <http://studierstube.icg.tu-graz.ac.at/handheld_ar/stbtracker.php>. Acesso em: 24 de jul. de 2008.
- [41] MIRANDA, F. R. D.; TORI, R.; BUENO, C. E. S.; TRIAS, L. P.. **Ar x-ray: Portable projector-based augmented exploration of buildings**. In: SVR 2008, . **Proceedings of X Symposium on Virtual and Augmented Reality**, 2008. p. .
- [42] WOODS, E.; MASON, P.; BILLINGHURST, M.. **Magicmouse: an inexpensive 6-degree-of-freedom mouse**. In: GRAPHITE '03, . **Proceedings of the 1st international conference on Computer graphics and interactive techniques in Australasia and South East Asia**, 2003, Melbourne, Australia. ACM. p. 285–286.
- [43] BILLINGHURST, M.; KATO, H.. **Real world teleconferencing**, CHI '99: CHI '99 extended abstracts on Human factors in computing systems, p. 194–195.
- [44] WAGNER, D.; REITMAYR, G.; MULLONI, A.; DRUMMOND, T.; SCHMALSTIEG, D.. **Pose tracking from natural features on mobile phones**. In: ISMAR 2008, . **Proceedings The 7th IEEE and ACM International Symposium on Mixed and Augmented Reality**, 2008. p. .
- [45] WAGNER, D.; LANGLOTZ, T.; SCHMALSTIEG, D.. **Robust and unobtrusive marker tracking on mobile phones**. In: ISMAR2008, . **Proceedings of The 7th IEEE and ACM International Symposium on Mixed and Augmented Reality**, 2008. p. .
- [46] LIENHART, R.; MAYDT, J.. **An extended set of haar-like features for rapid object detection**. In: ICIC2002, . **Proceedings of 2002 International Conference on Image Processing**, 2002. p. 900–903.
- [47] PAPAGEORGIOU, C.; OREN, M.; POGGIO, T.. **A general framework for object detection**, Sixth International Conference on Computer Vision, p. 7.

- [48] PAUL VIOLA AND MICHAEL JONES. Rapid object detection using a boosted cascade of simple features.
- [49] MENEZES, P.; BARRETO, J. C.; DIAS, J.. **Face tracking based on haar-like features and eigenfaces**. In: IAV2004, . **Proceedings of 5th IFAC Symposium on Intelligent Autonomous Vehicles**, 2004. p. 5–7.
- [50] Sirovich, L.; Kirby, M.. **Low-dimensional procedure for the characterization of human faces**, Journal of the Optical Society of America A, v. 4, p. 6.
- [51] HONG, K.; MIN, J.; LEE, W.; KIM, J.. **Real time face detection and recognition system using haar-like feature/hmm in ubiquitous network environments**. In: ICCSA2005, . **Proceedings of International Conference on Computational Science and Its Applications**. Springer Berlin / Heidelberg, 2005. p. 1154–1161.
- [52] ZHANG, Z.. **Flexible camera calibration by viewing a plane from unknown orientations**. In: , . 1999. p. 666–673 vol.1.