Interação baseada em gestos de mãos e cabeça para
ambientes de realidade virtual utilizando visão
computacional

Marcio Calixto Cabral

Dissertação apresentada como requisito parcial à
obtenção do grau de Mestre em Ciência da Com-
putação ao Instituto de Matemática e Estatística
da Universidade de São Paulo.

Orientador: Prof. Dr. Carlos Hitoshi Morimoto

SÃO PAULO
2005

à minha família.

Agradecimentos

Primeiramente gostaria de agradecer meu orientador, Professor Doutor Carlos Hitoshi Morimoto pela paciência, pela inestimável ajuda durante todo o desenvolvimento deste trabalho, pelas incontáveis correções e revisões realizadas e por horas e mais horas de conversa que tivemos a respeito não só da dissertação mas também de outros assuntos.

À minha família, que sempre me apoiou e incentivou na realização deste trabalho.

Aos meus amigos, que foram deixados de lado devido ao tempo consumido para a realização deste trabalho.

Ao meu orientador inicial, Professor Doutor Antonio Elias Fabris, pelo incentivo, desde os tempos de graduação, para a realização do mestrado e pelo conhecimento compartilhado.

Aos meus amigos do Laboratório de Sistemas Integráveis, CAVERNA Digital, pela ajuda e também pela troca de idéias durante a realização deste trabalho.

Ao Professor Doutor Marcelo Knörich Zuffo, pelo apoio incondicional dado para a realização deste trabalho.

À minha namorada, Luciana, pela paciência, compreensão e apoio durante a realização deste trabalho.

Resumo

O reconhecimento computacional de gestos em tempo real para interação com diversos tipos de ambientes de realidade virtual (RV) é uma área em constante evolução mas que ainda propicia um campo de pesquisa desafiador. Neste trabalho duas áreas de pesquisa têm grande influência: Visão Computacional e Interação Humano Computador. Através de métodos de Visão Computacional, o reconhecimento de gestos feitos pelo usuário possibilita a interação em ambientes de RV. O uso de gestos permite uma interface com boa usabilidade, pois é não invasiva e não requer longo tempo de preparação, comum aos dispositivos atuais utilizados nestes ambientes.

Nossa maior contribuição tecnológica foi o desenvolvimento de um rastreador de mãos e face em tempo real, através de métodos estatísticos para segmentação automática de regiões com cor de pele. Para distinguir as mãos da face, a região correspondente à face é encontrada através da Análise de Componentes Principais. Uma vez identificadas, estes objetos são rastreados usando filtros de Kalman e um grafo de consistência de estados que garante a robustez do sistema em situações de oclusão e sobreposição dos objetos. Além disso, para diminuir problemas com a variação da iluminação ambiente, desenvolvemos um processo de recalibração dinâmica do modelo de cor de pele.

Nessa dissertação também apresentamos o estado da arte dos dispositivos de interação em ambientes de RV, e desenvolvemos uma interface não invasiva baseada em gestos. Para validação do sistema desenvolvemos aplicações que foram testadas em ambiente de RV como *Powerwall* e CAVEs. Devido às grandes limitações de iluminação no ambiente CAVE, o desempenho do sistema foi insatisfatório para alguns dos testes realizados, porém em condições

melhores de iluminação, como em *Powerwalls*, mesmo com pequenas variações, o sistema se mostrou bastante robusto. Nossos resultados experimentais comprovam o grande potencial da utilização de gestos em ambientes de RV, onde apesar de um desempenho (velocidade) inferior a um mouse, todos os usuários se mostraram satisfeitos com esta nova modalidade de interação.

Abstract

Real time computer vision based gesture recognition for interaction in a myriad of immersive environments is a topic in constant evolution nowadays and it still constitutes a challenging research field. In this dissertation two research areas have great influence: Computer Vision and Human Computer Interaction. Using computer vision techniques, user's gestures are recognized in order to allow interaction in Virtual Reality (VR) environments. The use of gestures allows an interface with good usability because it is non invasive and it does not require a long setup time which is common to devices available today.

Our main contribution was the proposal and development of real time face and hands tracker, based on computer vision, that uses a statistical formula for automatic skin color segmentation. In order to differentiate between hands and face, the face region is detected using a Principal Component Analysis (PCA) technique. Once these regions are identified in the image, they are tracked using Kalman filters and a state consistency graph. This graph guarantees the system's robustness in situations where the face and hands get occluded or intersect with each other. Also, in order to attenuate problems with varying ambient illumination, we developed a dynamic skin color model recalibration.

In these dissertation we also developed a non invasive gesture based interface. We developed applications to validate the system, which were tested in Virtual Reality environments such as CAVEs and Powerwalls. Due to the restricted lighting conditions in the CAVE, the system's performance was unsatisfactory for some of the tests performed. However, in better lighting conditions, such as in Powerwalls, even with small variations, the system's robustness was good. Our experimental results show the great potential of using gestures in VR environments,

where even though the performance was worse than a mouse, all the users proved satisfied with this new interaction modality.

Sumário

1	Introdução	19
1.1	Motivação	22
1.2	Objetivos	24
1.3	Detalhamento do Problema	25
1.4	Contribuições	30
1.5	Organização do Texto	30
2	Interação em Ambientes de Realidade Virtual	33
2.1	Ambientes de Realidade Virtual	33
2.1.1	CAVEs	34
36		
2.1.3	Sistema semi-esférico	36
2.2	Telas Interativas	37
2.3	Sistemas de Rastreamento	38
2.3.1	Sistemas Eletromagnéticos	39
2.3.2	Sistemas Baseados em Câmeras	39
2.4	Reconhecimento de Gestos	40
2.5	Conclusão	41

3	Detecção e Rastreamento de Mãos e Face	43
3.1	Cor	44
3.1.1	Segmentação de cor de pele	45
3.2	Métodos de Segmentação de Pele baseados em Cor	48
3.2.1	Método Empírico	48
3.2.2	Método da Máxima Verosimilhança	49
3.3	Filtros de Kalman	53
3.4	Eigenfaces	54
3.4.1	Calculando Eigenfaces	55
3.4.2	Detecção de Faces	57
3.5	Conclusão	57
4	Sistema Interativo Baseado em Gestos	59
4.1	Visão Geral	59
4.2	Segmentação e Rastreamento de Mãos e Face	60
4.2.1	Treinamento do modelo de cor de pele	61
4.2.2	Segmentação de Cor de Pele	61
4.2.3	Rotulação das componentes conexas	62
4.2.4	Rastreamento de mãos e face	64
4.3	Interpretação de Gestos	67
4.3.1	Interface	68
4.3.2	Comandos	70
4.3.3	Limitações do Sistema	74
4.4	Conclusão	74

<i>SUMÁRIO</i>	11
5 Resultados Experimentais e Discussão	77
5.1 Descrição da implementação do protótipo	77
5.1.1 Equipamentos	78
5.1.2 Software e Algoritmos	78
5.1.3 Desempenho	80
5.2 Experimentos com a interface baseada em gestos	81
5.2.1 Ambiente Quiosque	82
5.2.2 Resultados	90
5.2.3 Ambiente CAVE	92
5.3 Discussão e Resultados	95
5.3.1 Quiosque	96
5.3.2 CAVE	97
5.3.3 Segmentação	98
5.3.4 Rastreamento	99
5.3.5 PCA	101
5.3.6 Interface baseada em Gestos	103
5.4 Conclusão	104
6 Conclusão e Trabalhos Futuros	105
A Mapeamento Tela-Imagem	109
B Questionário	113

Lista de Tabelas

5.1	Correlação Linear entre os coeficientes da Lei de Fitts	88
5.2	Coeficientes a e b da Equação 5.2 obtidos de forma empírica	90

Lista de Figuras

2.1	Exemplo de aplicação na CAVERNA Digital da USP	35
2.2	Equipamento <i>Vision Station</i> da <i>Elumens</i>	37
2.3	Imagem capturada por uma câmera posicionada sobre uma das telas da CAVERNA Digital. As condições de iluminação são desfavoráveis para captura de imagens com uma câmera comum.	41
3.1	Amostra de cor de pele de uma das pessoas utilizada para treinamento.	47
3.2	Histograma de distribuição de cor r (equação 3.1) e g (equação 3.2) da figura 3.1.	47
3.3	Resultados da aplicação da função $\lambda(r, g)$ com diferentes valores para x . a) $x = 0.1$ b) $x = 0.2$ c) $x = 0.3$	49
3.4	Resultados da aplicação da função $\lambda(r, g)$ com diferentes valores para x . d) $x = 0.4$ e) $x = 0.5$ f) $x = 0.6$	50
3.5	Imagem usada como entrada para teste dos algoritmos de segmentação	50
3.6	Distribuição de cor de pele no espaço r g	51
3.7	Função de Máxima Verossimilhança 3.8 aplicada a imagem de entrada (a , b , c e d) e seu respectivo Limiar (e , f , g e h). As áreas de intensidade mais clara indicam probabilidade maior de cor de pele	52
4.1	Fluxo de dados do sistema	60

4.2	A partir do canto superior esquerdo em sentido horário: imagem original, segmentação de cor de pele por limiarização, operação de erosão e operação de dilatação.	62
4.3	Posição inicial do sistema: mãos na altura do peito. a) CAVE b) Quiosque. . .	63
4.4	Grafo de consistência espaço-tempo. Os valores dentro de cada estado representam os objetos sendo rastreados.	65
4.5	Apenas uma região de busca quando há sobreposição de objetos	66
4.6	Rotulação de situações que o grafo de consistência identifica.	68
4.7	a) Posição Inicial b) Posição Norte c) Posição Leste	69
4.8	d) Posição Sul e) Posição Oeste	69
4.9	a) Posição de Calibração (x_0, y_0) b) Posição de Calibração (x_1, y_1)	72
4.10	a) Posição de Calibração (x_2, y_2) b) Posição de Calibração (x_3, y_3)	73
4.11	a) Trecho da ordenada x antes da aplicação do filtro e b) Depois da aplicação do filtro.	76
5.1	Câmera <i>Firewire Dragon Fly</i>	79
5.2	Máscara posicionada na frente do CCD para captura de quadros.	80
5.3	Problemas na utilização no filtro de Bayer. a) Embaçamento, b) Efeito de <i>Grid</i> e c) Cores Falsas	81
5.4	Participantes do experimento.	83
5.5	Ambiente Quiosque.	84
5.6	Amostra de seqüência de imagens colhidas pela câmera posicionada no ambiente Quiosque	85
5.7	Experimento de apontamento: Assim que um usuário ativa um botão (círculo) outro aparece. A diferença entre tamanhos de círculos e distância entre eles é utilizada no modelo da Lei de Fitts	86

LISTA DE FIGURAS

	17
5.8 Comparação da reta $T = a + bID$ para as duas modalidades	89
5.9 Aplicativo desenvolvido para os experimentos no ambiente Quiosque	91
5.10 Ambiente CAVE. a) Câmera posicionada acima da tela principal e b) Layout Virtual do Ambiente.	93
5.11 Amostra de sequência de imagens colhidas pela câmera posicionada no ambiente CAVE	94
5.12 Aplicação utilizada para teste na CAVERNA Digital	95
5.13 Usuário utilizando o sistema (face e mãos rastreados) e pessoa atrás não interferindo no rastreamento do sistema.	96
5.14 Pessoa com cabelos loiros: segmentação de cor de pele detecta cabelos loiros como cor de pele	99
5.15 Quadros tirados com intervalo de 1 segundo entre si. Os quadros mostram a variação na segmentação ocasionada pelo ajuste dinâmico do espaço de busca.	100
5.16 Valor medido para a ordenada X do centróide da mão esquerda e seu respectivo valor previsto pelo filtro de Kalman.	101
5.17 Objetos identificados como mão direita R , face H e mão esquerda L	102

Capítulo 1

Introdução

A utilização de gestos como forma ou auxílio à comunicação é um fenômeno encontrado em diferentes culturas. Até mesmo indivíduos cegos de nascença utilizam gestos para se comunicar. Dada a sua naturalidade e universalidade, nesse trabalho vamos explorar o uso de gestos para criar uma forma alternativa de interface para interação humano-computador.

Pesquisas na área de interação entre humanos e máquinas têm se tornado comuns. Com o advento de microprocessadores, monitores e componentes cada vez menores e mais potentes, já é possível encontrar sistemas computacionais sempre a nossa volta e que muitas vezes passam despercebidos. A redução de custo permite que essa tecnologia seja cada vez mais utilizada para tarefas corriqueiras, como por exemplo, em quiosques de informação em locais públicos, máquinas para venda automática de produtos, telefones celulares, televisores, DVD's em automóveis, projetores que cabem na palma da mão, câmeras digitais do tamanho de canetas, etc.

Com tantos dispositivos computacionais e interativos ao nosso redor, surge o problema de como controlar cada um deles de forma fácil, natural, e sem a necessidade de aprender uma interface complexa, ou carregar um dispositivo específico como um controle remoto universal.

Em particular, como a exibição de informação é fundamental para várias modalidades de interação, trataremos nesse trabalho do problema de interação com telas de projeção grandes típicas de ambientes de realidade virtual (RV), cuja tecnologia vem evoluindo muito rapidamente. Utilizaremos o termo *tela* como referência a qualquer superfície onde informações visuais podem ser exibidas, como monitores de computador, telas de projeção, e outros equipamentos menos convencionais como *powerwalls* e quiosques, e até sistemas mais complexos como o CAVE (*Cave Automatic Virtual Environment*) [14, 81].

Esses equipamentos tem uma importância particular para as áreas de Computação Gráfica e Realidade Virtual, onde a informação visual é, por vezes, o único estímulo sensorial. Desde as primeiras formas de interação propostas por Sutherland nos anos 50 [66, 67], avanços na área não pararam de acontecer, tanto em equipamentos (*hardware*) quanto em algoritmos (*software*). Placas gráficas de última geração permitem maior rapidez e qualidade na geração de imagens [26], e sistemas CAVE permitem maior imersão do usuário no ambiente simulado. Aglomerados gráficos com alta escalabilidade e desempenho superior mas com custo inferior ao de supercomputadores gráficos, alimentam sistemas CAVE como mostrado por Soares *et al.* [62]. Bibliotecas e algoritmos de sincronização têm sido desenvolvidos para geração de imagens em paralelo utilizando tecnologia de aglomerados gráficos como pode ser visto em Paiva *et al.* [28] e Gnecco *et al.* [25].

Ambientes de Realidade Virtual utilizam grandes telas de projeção para oferecer melhor e maior resolução para visualização e exploração. Outros estímulos sensoriais são comumente utilizados para estimular a audição e o tato. Ambientes Virtuais Imersivos (AVI) oferecem uma experiência virtual mais completa ao usuário, proporcionando uma melhor integração com outros modos sensoriais ao mundo criado por computador. *Head Mounted Displays* (HMD) e *Binocular Omni-Oriented Monitor* (BOOM) também são exemplos de dispositivos que possibilitam um campo de visão grande em um AVI.

Mas efetivamente, a utilização de dispositivos convencionais como teclado e mouse, adequados a mesas de trabalho, não são facilmente aplicáveis em RV. Até mesmo dispositivos como sensores de posição, equipamentos de rastreamento eletro-magnéticos e joysticks não proporcionam liberdade de movimentos ao usuário.

A preocupação crescente nos últimos anos com a forma de interação entre humanos e ambientes virtuais faz com que avanços e estudos sejam realizados nesta área. Há na literatura uma grande preocupação no desenvolvimento de formas alternativas de interação apropriadas para RV. Hodges e Bowman [4] apresentam um estudo sobre seis propostas de interfaces, apresentando suas principais qualidades e defeitos, e propõem uma solução híbrida para os problemas encontrados. Nesse trabalho nós exploramos a utilização de gestos como uma alternativa natural para a interação em ambientes de RV, sem a necessidade de utilizar dispositivos especiais, para executar tarefas de controle como ligar e desligar uma televisão, mudar de canal, escolher opções entre menus diversos de algum sistema, controlar a exibição de *slides* em uma palestra ou de fotos de um álbum exibidas por um projetor, interagir em aplicações de navegação em RV e manipular objetos em ambientes 3D, que fogem ao modelo tradicional de interação do *Desktop Computing*.

A identificação de gestos utilizando visão computacional permite ao usuário interagir com objetos virtuais em ambientes como CAVEs e Quiosques com a grande vantagem de não requerer que o usuário segure algum equipamento especial ou então coloque algum dispositivo em seu corpo, como *datagloves* [80] e mouses 3D, proporcionando ao usuário uma sensação de naturalidade maior ao interagir com o ambiente.

Além disso, a possibilidade do ambiente observar o usuário, sem que o mesmo precise utilizar algum equipamento específico, transforma completamente a forma como a interação é realizada atualmente. O ambiente pode detectar a presença do usuário e interpretar e reconhecer gestos para a realização de tarefas rotineiras como ligar um aparelho de som.

Neste trabalho apresentamos um sistema de reconhecimento de gestos baseado em visão computacional para interação em ambientes de RV, cada vez mais comuns em ambientes comerciais e residenciais. Apesar do sistema não ser restrito a esse equipamento, podendo ser utilizado com telas menores como em computadores *desktop*, o uso de ambientes de RV permite futuras extensões do sistema para a criação de ambientes interativos, que conseguem detectar e interpretar gestos de um usuário. Nosso sistema também requer curto tempo de preparação para uso, e após iniciado o sistema, a troca de usuários pode ser feita sem a necessidade de recalibração.

1.1 Motivação

O uso de telas grandes, inclusive em ambientes de RV, já foi visto apenas como uma forma de entretenimento, mas cada vez mais pessoas estão percebendo o verdadeiro potencial da utilização desta tecnologia como uma interface natural de comunicação entre humanos e máquinas, como citado por Coomans [11]. Mas a interação 3D ainda não alcançou um nível tecnológico apropriado para aplicações em RV como mostra Bowman *et al.* [5]. Em outro trabalho, Bowman [4] apresenta alternativas para formas de interação, onde atividades de manipulação e navegação em mundos virtuais são apresentadas utilizando menus. De acordo com Bowman, o uso de menus não é natural mas necessário em aplicações onde fatores como eficiência e velocidade são cruciais. Quando o assunto é treinamento e replicação do mundo real, metáforas para interação natural são mais apropriados.

Na evolução da área de Interação Humano Computador (IHC) [55, 16, 58], os teclados foram os dispositivos primários usados para entrada de dados. Com o advento do mouse, a metáfora WIMP (*Windows, Icons, Menus, Pointer*) tornou as interfaces gráficas conhecidas atualmente amplamente utilizadas. Boa parte das aplicações em ambientes de RV também

foram desenvolvidas utilizando essas metáforas como mostrado por Bowman [4].

Em grande parte das aplicações onde telas grandes são utilizadas, como Quiosques e CA-VEs, os teclados, mouses e outros dispositivos tradicionais como joysticks e dispositivos de rastreamento, ainda são usados para controle e navegação. Entretanto, esses dispositivos não são adequados para uso nestes ambientes devido à falta de naturalidade para interação 3D (é incômodo senão impossível utilizar um teclado e mouse na frente de uma tela grande, onde o usuário não tem apoio para o mesmo). Além da falta de naturalidade, o fato do usuário tirar sua atenção da tela ao utilizar o teclado e o mouse é importante também. Em uma interface baseada em gestos, o usuário não precisa mudar sua atenção da tela, pois não existe nenhum dispositivo a ser utilizado na interação que não suas próprias mãos. O sucesso de uma aplicação neste ambiente está intrinsicamente ligado à qualidade do paradigma de interação como citado por Coomans [11]. O futuro das interfaces humano computador irá refletir o avanço computacional contínuo, expandindo o acesso perceptual às informações, numa tentativa de reconciliar a facilidade de uso e o aumento das necessidades e capacidades computacionais.

A utilização de gestos vem se tornando uma alternativa atraente para interação e entrada de dados em ambientes de RV. Como exemplo de técnicas para reconhecimento de gestos, podemos citar a técnica proposta por Black e Jepson [3] capaz de rastrear objetos rígidos e não rígidos utilizando autovetores e autovalores para rastreamento. Hanning e Huang [79] conseguem determinar com robustez a posição e orientação de mãos utilizando redes bayesianas. O método combina o uso de cores e um modelo pré-computado para as mãos. O uso de luvas e dispositivos de rastreamento para captura dos movimentos das mãos do usuário é caro e invasivo como mostrado por Maciel e Shirley [45]. Bowman *et al.* [6] apresenta e analisa algumas técnicas de interação para menus e navegação em ambientes de RV.

As técnicas baseadas em visão computacional proporcionam uma alternativa promissora para a captura de gestos, utilizando mãos e rosto, pois tem custo relativamente baixo quando

comparada às outras e não são invasivas. Hong *et al.* [30] desenvolveram um sistema baseado em estados para reconhecimento e aprendizado de gestos. As mãos e o rosto do usuário são usados como características a serem rastreadas através de segmentação por cor. As informações coletadas durante a fase de treinamento de um determinado gesto são agrupadas então com outros gestos para inclusão de informações temporais. Essas informações são integradas para construção de uma máquina de estados finitos (*Finite State Machine - FSM*) para reconhecimento. Os autores afirmam conseguir taxas interativas em tempo real devido à eficiência computacional do FSM. Fong *et al.* [23] desenvolveram um sistema para identificação de gestos para controle da direção de robôs. A identificação de gestos é feita utilizando-se um filtro para cores normalizadas e visão estéreo. Entretanto, os autores afirmam que o sistema é não intuitivo para os usuários, dificultando seu uso.

1.2 Objetivos

O principal objetivo deste trabalho é o estudo de formas alternativas e mais naturais para interação em ambientes de RV, através do desenvolvimento de um sistema não-invasivo, capaz de ser utilizado em lugares diversos, desde aplicações em escritórios até residências. Para isso, técnicas de visão computacional são utilizadas para segmentar e rastrear as mãos e a cabeça do usuário e posteriormente interpretar os seus movimentos para controlar e interagir em ambientes de RV, onde interfaces não-invasivas são mais adequadas. Interfaces não-invasivas são aquelas onde não há a necessidade do usuário vestir ou usar equipamentos que interfiram na fluidez de gestos e movimentação de seu próprio corpo. Por outro lado, interfaces invasivas são aquelas que requerem que o usuário vista ou utilize equipamentos e dispositivos para manipulação que limitam ou atrapalhem a sua movimentação e flexibilidade, como *datagloves* [80].

Isso envolve dois temas diferentes: detecção e rastreamento do usuário em seqüências de vídeo em tempo real; e interpretação destes gestos de acordo com a interface criada para controle da aplicação. Todo o processo deve ser executado a taxas interativas de atualização para não comprometimento da imersão do usuário no ambiente de RV.

A tarefa de detectar e interpretar o movimento das mãos e da face do usuário como forma de interface é uma área complexa e desafiadora. Por outro lado, é interessante restringir o menos possível os movimentos do usuário, bem como evitar o uso de dispositivos adicionais e marcadores para rastreamento. Por isso, optou-se por rastrear mãos e rosto sem o uso de marcadores.

No desenvolvimento do protótipo, algumas limitações serão consideradas. Por exemplo, assumimos que o usuário deverá estar dentro da área de visão da câmera que captura as imagens, o que inclui as mãos e a face do usuário. Para melhorar o desempenho, é necessário que o usuário esteja utilizando roupas de manga comprida, para que apenas suas mãos e face estejam visíveis para o sistema de captura. A iluminação do ambiente será artificial.

1.3 Detalhamento do Problema

O reconhecimento de gestos de mãos é uma tarefa complicada devido a grande variedade de possibilidades existentes de movimentação das mãos, de forma aleatória, e a não existência de uma linguagem pré-definida. Uma alternativa para tornar um sistema de reconhecimento de gestos não-invasivo é utilizando um sistema de visão computacional. Neste caso, é necessário rastrear as mãos do usuário, por exemplo, baseado na cor de pele, evitando-se o uso de luvas ou marcadores para facilitar a segmentação. Alguns trabalhos sobre rastreamento baseado em cor reportam problemas como similaridade do fundo com a tonalidade da pele, como descrito por Ahmad [1], problemas em rastrear mais de um objeto com a mesma tonalidade, complexidade

e quantidade de gestos detectados tornando o sistema lento e impossibilitando seu uso para aplicações em tempo real.

A iluminação também é um fator importante no caso de ambientes de RV, como na CAVERNA Digital (Zuffo *et al.* [81]), onde o sistema de projeção pode causar alterações na tonalidade da pele vista pela câmera. Isso dificulta o rastreamento pois o brilho médio de aplicações em sistemas como a CAVERNA Digital é de 25 lux como afirmam Moesland *et al.* [52]. Sistemas de rastreamento enfrentam problemas sob condições de iluminação variáveis. A mesma área de pele pode ser detectada como cores diferentes sob iluminações diferentes como apresentado por Störing *et al.* [65]. Este é um problema comum conhecido em visão computacional como constância de cor. O problema de constância de cor também foi descrito por Birgitta *et al.* em [49]. Diferentes fontes de luz bem como diferentes intensidades de luz e diferentes cores emitidas pelos projetores em ambientes virtuais imersivos afetam a percepção da cor de pele do usuário.

Mesmo com o rastreamento efetuado de forma robusta, o agrupamento das componentes encontradas nas imagens se faz necessária devido a suscetibilidade a ruídos na classificação. O agrupamento em regiões conexas (essas regiões conexas sem forma são chamadas de *blobs*) é uma alternativa viável. A técnica desenvolvida por Wren *et al.* [77] apresenta resultados robustos utilizando *Minimum Description Length* (MDL). Usualmente, antes de se agrupar os *pixels* em regiões, um filtro de embaçamento é aplicado sobre a imagem, para que regiões conexas próximas, ou regiões conexas que estejam separadas por ruídos ou má classificação anterior, sejam consideradas como parte de um mesmo agrupamento. Através da segmentação e rastreamento, é possível estimar parâmetros dos mesmos como posição, área e centro de massa. Isso possibilita empregar tais dados no reconhecimento de gestos.

Além disso, dado que a mão seja detectada apropriadamente em tempo real, a definição de uma interface sem redundância e de fácil execução e memorização pelo usuário é necessária

para a robustez do sistema. Movimentos bruscos podem prejudicar a captura e interpretação, dependendo da velocidade de rastreamento do sistema, como mostrado por Hong *et al.* em [30]. Eles também afirmam que a identificação de término e início de cada gesto também é algo que deve ser levado em conta, pois a princípio o usuário não estará gesticulando para controle durante toda a execução da aplicação.

Apesar de disponíveis comercialmente para a comunidade, sistemas para rastreamento de mãos, cabeça e corpo do usuário ainda sofrem com problemas de acurácia, latência, custo e suscetibilidade a interferências externas como descrito no relatório de interfaces virtuais de Youngblut *et al.* [78] elaborado em 1996. Youngblut afirma que soluções híbridas são alternativas viáveis para superar as limitações individuais de cada técnica. Métodos mais recentes como o proposto por Sminchisescu e Triggs [61], são eficazes mas não funcionam em tempo real. A utilização de câmeras de alta frequência e resolução, como as câmeras da Vicom [44], permitem atingir processamento em tempo real. Mas atualmente são caras e impõem que o usuário utilize marcadores nas partes do corpo que deverão ser rastreadas.

O rastreamento de movimentos para posterior análise e reconhecimento dos gestos pode ser feita através de duas formas principais: visualmente através de captura de vídeo com câmeras ou através de dispositivos rastreadores de posição eletromagnéticos. Os primeiros dispositivos são uma alternativa de custo mais baixo e não-invasiva. Porém são muito suscetíveis a problemas de oclusão e variação de iluminação. Já no caso dos dispositivos eletromagnéticos, a precisão é muito grande e não há problemas de oclusão. Porém o custo dos mesmos é elevado e este tipo de equipamento, além de apresentar interfaces invasivas, é muitas vezes desconfortável para o usuário.

Apesar de vivermos e interagirmos num mundo 3D, o mundo físico contém muitas variáveis para interação, que ainda não são possíveis de serem representadas totalmente em uma simulação computacional. A adaptação de paradigmas utilizados até então para 2D pode até

ser implementada, mas não de forma satisfatória a suprir as necessidades e requerimentos que aplicações de RV possuem. Desta forma, novas interfaces, baseadas em paradigmas distintos dos já existentes, precisam ser desenvolvidas.

Três tarefas básicas são identificadas na grande maioria das aplicações em ambientes interativos como descrito por Bowman *et al.* [4]: locomoção, seleção e manipulação. Locomoção, ou também controle do ponto de vista, se refere a tarefa do usuário interativamente se posicionar e mudar seu ponto de vista no ambiente. Seleção consiste na escolha de um objeto específico a ser manipulado. E finalmente, manipulação, que consiste na utilização de comandos para manipular o objeto selecionado.

Quase a totalidade de aplicações está mais preocupada na translação de um ponto ao outro. Em outro trabalho do grupo [5], eles classificam navegação em três categorias. *Exploração* é a navegação sem objetivo específico. *Manobra* é a navegação onde precisão e acurácia em espaços curtos é necessária. E *Busca*, que é a localização e ida para um ponto específico do ambiente.

Além destas três categorias básicas presentes na quase totalidade de aplicações na área, cinco são as categorias definidas por Bowman *et al.* [5] que estão presentes na maioria dos trabalhos em técnicas de interação:

- **Movimento Físico:** O movimento de corpo do usuário é usado como forma de navegação. Este tipo de técnica só pode ser empregada onde a área de rastreamento é grande o suficiente para acomodar movimentos como esse. Bicicletas estacionárias também podem se encaixar nesta categoria.
- **Manipulação Manual de Ponto de Vista:** Os movimentos de mão do usuário são usados para efetivamente executar a locomoção. Apesar de causar fadiga, esta técnica é fácil de ser aprendida pelo usuário.

- **Direcionamento:** Especificação contínua de direção de movimento, com um usuário apontando uma direção com o dcdo. Esta técnica é bastante eficiente pois é intuitiva.
- **Navegação com Destino:** Conhecida também como tele-transporte, consiste em ir diretamente para o ponto de destino indicado pelo usuário, pulando imediatamente para a nova posição. De acordo com estudos feitos por Bowman em [4] este tipo de técnica deixa o usuário “perdido”, alterando o seu senso de localização no ambiente.
- **Planejamento de Rota:** O caminho é especificado pelo usuário e o sistema se encarrega de locomover o ponto de vista por este caminho. Possibilita que o usuário controle a navegação e ainda assim fique livre para realizar outras tarefas.

Técnicas para seleção e manipulação 3D em ambientes de RV devem ser objetivas o suficiente para serem capazes de selecionar o objeto e movimentá-lo (translação e rotação). Sendo esta modalidade de interação extremamente usada tanto em 2D como em 3D, a definição de técnicas de interação apropriadas tem um grande efeito na qualidade final da interface.

Dentre as várias técnicas existentes, algumas assumem proximidade ao objeto, o que implica que o usuário está a uma distância tal que o permite tocá-lo. Outras técnicas não se baseiam nesta hipótese e permitem que o usuário manipule o objeto, mesmo ele estando bastante afastado do usuário.

A grande variedade de técnicas disponíveis torna a escolha difícil e complicada pois cada uma tem suas vantagens e atrativos com relação às outras. Entretanto, algumas diretrizes básicas podem ser levadas em consideração. Nenhuma das técnicas pode ser considerada como sendo a melhor, e a restrição de interação e de alguns graus de liberdade é importante para o sucesso da interface.

1.4 Contribuições

A principal contribuição deste trabalho é a criação de uma técnica baseada em visão computacional para reconhecimento de gestos em ambientes de RV, como Quiosques e CAVEs. Esta técnica compreende a junção de segmentação por cor de pele, detecção e rastreamento de formas, informação contextual e técnicas para interação 3D através de gestos definidos neste trabalho.

Outras contribuições deste trabalho incluem:

- Amplo levantamento bibliográfico, com o estudo dos principais métodos de segmentação de cor de pele e reconhecimento de gestos para interação em ambientes virtuais imersivos;
- Experimentos com segmentação de cor de pele em diferentes espaços de cor para ambientes com pouca iluminação e iluminação variável;
- Desenvolvimento de uma técnica robusta para o rastreamento de mãos e face, capaz de trabalhar sob condições de baixa iluminação e com pequenas variações de iluminação;
- Desenvolvimento de uma técnica de interação utilizando gestos para ambientes com telas interativas como CAVEs e também Quiosques;
- Implementação de aplicativos de teste implementando os esquemas propostos para avaliação do método;
- Experimentos realizados com usuários para avaliação da interface desenvolvida.

1.5 Organização do Texto

Esta dissertação está organizada da seguinte forma: no capítulo 2 será apresentada uma breve revisão bibliográfica sobre ambientes de RV, sistemas de rastreamento e detecção de gestos. O

capítulo 3 resume a base teórica que será utilizada pelo sistema de reconhecimento de gestos. O capítulo 4 descreve o sistema de reconhecimento de gestos implementado. O capítulo 5 descreve os experimentos realizados com o sistema desenvolvido e discute alguns resultados encontrados. Por fim, o capítulo 6 traça nossas conclusões, mencionando algumas direções e o desenvolvimento de trabalhos futuros.

Capítulo 2

Interação em Ambientes de Realidade Virtual

Vários trabalhos já foram propostos para realizar o rastreamento de gestos baseados em visão computacional para ambientes de RV. Neste capítulo iremos abordar um panorama geral das principais áreas correlatas.

2.1 Ambientes de Realidade Virtual

A Realidade Virtual é considerada a interface mais completa entre homem e máquina. Utilizando RV, os sentidos do usuário são extensamente estimulados por informações de um computador, sendo capaz de transmitir muito mais informações do que uma interface WIMP.

Sistemas para RV colocam o usuário dentro de um ambiente virtual, para que o usuário tenha uma experiência próxima da realidade, muito além da experiência oferecida através do uso de dispositivos comuns como monitores de vídeo. Como a visão é em geral o principal sentido explorado por estes ambientes, as grandes telas fornecem uma melhor sensação de

imersão. Há várias propostas de ambientes que exploram diferentes formas de projeção, como CAVEs, Quiosques e *Powerwalls*, descritos neste capítulo.

Ambientes imersivos não-invasivos são descritos como aqueles onde a interface computacional naturalmente envolve o usuário, sem que o usuário necessite utilizar dispositivos presos ao seu corpo. Para que isto seja atingido, telas que cubram a maior parte da visão do usuário são necessárias. Imersão é uma extensão da técnica conhecida como oclusão de informação, onde os dispositivos de entrada e saída não são visíveis ao usuário, isolando o usuário desses equipamentos computacionais [54]. Desta forma, dado que o usuário não “vê” o computador, a pré-concepção de que esta o usuário está interagindo com um computador não existe. Lombardi [43] afirma que: “... no futuro, ambientes imersivos serão a principal forma de interação encontradas porque os usuários não irão aceitar as formas incômodas dos dispositivos atuais de realidade virtual, como HMDs (Head Mounted Displays).”. O CAVE é um exemplo de ambiente de RV que busca satisfazer essas condições.

2.1.1 CAVEs

O primeiro sistema CAVE de projeção cúbica foi desenvolvido na Universidade de Illinois em Chicago pelo grupo da pesquisadora Carolina Cruz-Neira [14]. Sua primeira proposta foi a colocação de monitores de vídeo ao redor da cabeça do usuário, mas logo se percebeu que telas e projetores seriam bem mais práticos e eficientes.

Os sistemas cúbicos apresentam maior campo de visão (até 360 graus na horizontal e 270 graus na vertical), capacidade de resolução de imagens acima de 2000×2000 e também são menos restritivos com relação a movimentação do usuário. Isto implica em melhor qualidade de imersão quando comparados, por exemplo, aos capacetes de realidade virtual HMDs como primeiramente idealizados por Sutherland em [67], que são pesados e apresentam sérios problemas de rastreamento, navegabilidade, campo de visão reduzido e podem induzir ao mal

Figura 2.1: Exemplo de aplicação na CAVERNA Digital da USP

estar no usuário. Sistemas mais leves e robustos que estão sendo desenvolvidos atualmente procuram reduzir esses problemas.

O primeiro sistema baseado nesses conceitos a ser construído na América Latina foi a CAVERNA Digital. A CAVERNA Digital, desenvolvida por Zuffo *et al.* [81], é um sistema de realidade virtual que consiste de uma sala onde cada parede é uma tela de projeção. Assim, o observador que se encontra dentro do ambiente tem um grande grau de imersão.

Dentre as possíveis aplicações da CAVERNA Digital, podemos destacar a visualização de lençóis petrolíferos para prospecção de petróleo em águas profundas, as maquetes digitais na indústria automobilística, a análise meteorológica como a visualização de massas de ar, passeios arquiteturais em locais históricos ou ainda mesmo nem construídos, além de simulações cirúrgicas na medicina, planetário virtual, dentre muitas outras. Veja na figura 2.1

uma aplicação de prototipagem de automóvel.

Apesar de apresentar características de alto nível de complexidade, a interação na CAVERNA Digital e em ambientes baseados em sistemas de projeção cúbicos ainda é algo não muito explorado, principalmente através do uso de câmeras. Todos os testes baseados em sistema CAVE realizados neste trabalho foram feitos na CAVERNA Digital.

2.1.2 Quiosques ¹

Quiosques são sistemas compostos por telas grandes e projetores para exibição de imagens. Muitas vezes, vários projetores são utilizados para cobrir apenas uma tela, proporcionando melhor resolução. As principais aplicações de Quiosques são a visualização e a exploração de dados científicos (meteorológicos, geográficos, etc) e a navegação em mundos virtuais, que podem ser realizadas interativamente, através de dispositivos mais complexos como rastreadores ou até mesmo mouse e teclado.

2.1.3 Sistema semi-esférico

Os sistemas de projeção semi-esféricos [13] ao invés de possuírem uma tela plana, possuem uma tela esférica e a projeção é feita por um projetor disposto entre o usuário e a tela. Estes sistemas possuem algumas vantagens por não apresentarem descontinuidades na imagem, porém a presença do projetor impede que o usuário se aproxime muito da tela. Veja na figura 2.2 o equipamento *Vision Station*, desenvolvido e atualmente comercializado pela *Elumens Corporation*,

As adaptações que foram executadas neste sistema permitiram criar um sistema de cola-

---

¹Powerwalls: semelhantes aos quiosques, powerwalls podem possuir mais de uma tela de projeção, sendo portanto uma alternativa intermediária entre CAVE's e quiosques. Uma aplicação em Quiosque pode ser vista com mais detalhes no trabalho de Regh *et al.* [56].

Figura 2.2: Equipamento *Vision Station* da *Elumens*

boração portátil muito eficiente. Uma câmera foi implantada na parte superior da semi-esfera. Embora apresente uma grande inclinação, esta funciona perfeitamente para capturar imagens e reenviá-las para outro ambiente. Um joystick com retorno de força foi utilizado para navegar no ambiente.

2.2 Telas Interativas

Os tópicos mais relevantes para a criação de um ambiente imersivo são visualização e interação. As interfaces são as ferramentas que permitem a troca de informações com o ambiente (ou tela). A forma de interação define como o usuário e o sistema se comunicam. A interface em si não deve desviar a atenção do usuário, de forma que a atividade primária (o que deve ser

feito) deve prevalecer sobre a interface (como deve ser feito), como descreve Coomanns em [11]. Taxonomias de interfaces estão amplamente difundidas e sendo utilizadas fora de RV, porém elas são limitadas e portanto inapropriadas para aplicações em RV.

O enfoque dos trabalhos mais expressivos que atuam nas áreas de IHC e RV, se concentra na navegação e procedimentos de interação como mostram Gabbard e Hix em [24]. Outros trabalhos, como o de Kaurs [37], promovem melhorias nos procedimentos de interação em AVI's sob os princípios de usabilidade das interações, contribuindo para o alcance da qualidade de interação destes sistemas. Apesar de pouco, representam os primeiros passos de uma caminhada para o desenvolvimento de sistemas mais simplificados e objetivos.

Coomans [11] propôs um modelo de taxonomia baseado em definições previamente formuladas em IHC para RV que define características fundamentais como Interação, Visualização, Simulação e Imersão. Dentro da taxonomia proposta, Interação em RV significa que o computador precisa entender comandos naturais de serem executados pelo usuário. Além disso, para que o usuário perceba que seus comandos estão sendo executados como resultado de suas ações, o tempo de resposta do sistema também é importante. Para todas essas características, dispositivos apropriados precisam estar disponíveis.

2.3 Sistemas de Rastreamento

Pode-se classificar as formas de entrada de dados em ambientes de Realidade Virtual como invasivas e não-invasivas, como já mencionado. Dentre as técnicas invasivas podemos considerar as luvas de dados (*datagloves*) e dispositivos de rastreamento (*Tracking Devices*) que precisam ficar em contato com o usuário. Já as técnicas não-invasivas são aquelas que utilizam câmeras, ultrasons e dispositivos a *laser*. Uma breve descrição destes equipamentos é apresentada a seguir.

2.3.1 Sistemas Eletromagnéticos

Cutler *et al.* [15] propõem um sistema baseado no trabalho de Guiard [27] onde as mãos podem ser usadas em conjunto ou separadamente para manipular naturalmente modelos virtuais. A interface usada para captura são rastreadores de posição (luvas e canetas) acarretando um custo elevado e tornando o sistema invasivo, porém a metodologia adotada para manipulação de modelos tri-dimensionais se mostra satisfatória e natural. Fels e Hinton [18] usam luvas com sensores de movimento e posição para detecção de sinais de gestos que compõem a linguagem surdo-mudo utilizando redes neurais. Além disso, o uso de rastreadores magnéticos para entrada de dados está sujeita à interferência magnética.

2.3.2 Sistemas Baseados em Câmeras

Sistemas baseados em câmeras são intrinsicamente menos invasivos, mesmo com o uso de marcadores espalhados pelo corpo do usuário ou o uso de objetos específicos que visam simplificar os algoritmos de visão, como canetas ou apontadores com cores marcantes que facilitam o rastreamento. Entre os sistemas desenvolvidos com essa tecnologia de rastreamento visual podemos destacar o de Dorfmueller [17], o de Ribo *et al.* [57] e o de Welch *et al.* [76]. Soluções comerciais também se encontram disponíveis como as da *Ascension Technologies* [12] chamada *LaserBird* e da IGT descrita em [70]. Porém, o uso de marcadores torna o sistema não robusto o suficiente para a maioria das aplicações porque requerem um grande tempo extra de montagem e configuração. Shamaie e Sutherland [60] desenvolveram um sistema, melhorado posteriormente por Shamaie [59], que utiliza Análise de Componentes Principais (*Principal Component Analysis*), abordagem multi-escala e pareamento de grafos para rastreamento mesmo na presença de oclusão das mãos.

2.4 Reconhecimento de Gestos

O objetivo de um sistema de reconhecimento de gestos é identificar gestos humanos específicos. O gesto de uma mão ou corpo pode ser definido por uma posição específica do corpo, ou então pelo movimento físico definido por ele em 2 ou 3 dimensões. Uma vez que o gesto é identificado em uma sequência de vídeo, ele é transformado em símbolos, que podem ser usados para controlar um dispositivo. Como exemplo, Hong *et al.* [30] descrevem um sistema que automaticamente reconhece gestos para implementar uma aplicação que simula uma brincadeira conhecida como “*rei mandou*” (*Simon Says*). Gestos simples, como um aceno de mão, são reconhecidos através de uma máquina de estados finitos (*Finite State Machine*).

O conceito de gestos abrange uma definição muito grande do ponto de vista biológico ou sociológico. Pesquisadores tentam definir e classificar gestos de acordo com o que se observa no dia a dia. Pessoas freqüentemente utilizam gestos para se comunicar. Gestos são usados em situações corriqueiras como apontar para uma pessoa para chamar sua atenção e até descrever emoções e sentimentos como visto por Kendom [40]

Gestos associados a fala são referenciados como gesticulação. Gestos que funcionam independentemente da linguagem falada são referenciados como autônomos. Gestos autônomos também podem representar comandos de movimento, ou então serem usados no lugar da fala como em outro trabalho de Kendom [39] e também no trabalho de Malandro *et al.* [47].

O reconhecimento de gestos em ambientes imersivos, como CAVEs, Quiosques, *Powerwalls* e sistemas desta natureza, é um problema bastante desafiador devido às condições severas de iluminação e posicionamento da câmera (veja a figura 2.3 que mostra as condições normais de iluminação da CAVERNA Digital). Para conseguir extrair dados 3D de posicionamento do usuário, diversas técnicas multi-câmeras para estimação da posição do corpo foram publicadas nos últimos anos. A maioria delas tenta encaixar um modelo articulado de corpo humano em dados 3D adquiridos através das câmeras [51, 10], mas a captura rápida destas estruturas

Figura 2.3: Imagem capturada por uma câmera posicionada sobre uma das telas da CAVERNA Digital. As condições de iluminação são desfavoráveis para captura de imagens com uma câmera comum.

articuladas está longe de ser trivial. Kehl e Van Gool [38] propuseram uma redução no número de gestos identificados para simplificação do problema. Em seu sistema, Kehl reconhece apenas gestos de apontamento. Leubner [42] também utiliza geometria de múltiplas vistas (*multi-view geometry*) em um *powerwall* com projeção traseira mas o sistema é limitado a uma velocidade de 8 quadros por segundo.

2.5 Conclusão

Ambientes Imersivos em Realidade Virtual significam que o usuário está envolto neste sistema computacional de forma a não perceber os equipamentos envolvidos na interação. A utilização de visão computacional como forma de interação baseada em gestos substitui os dispositivos invasivos como *datagloves*, que prejudicam a imersão do usuário.

A utilização de múltiplas câmeras torna o sistema mais robusto porém também mais complexo. Técnicas mais simples como subtração de fundo são computacionalmente bastante eficientes mas assumem um fundo constante, que não ocorre em geral em ambientes imersivos e interativos. O próximo capítulo descreve a técnica de rastreamento baseada em cor de pele, que foi utilizada no desenvolvimento de nossa interface baseada em gestos.

Capítulo 3

Detecção e Rastreamento de Mãos e Face

A detecção e o rastreamento de objetos [32, 29, 69, 19] é um problema fundamental de visão computacional. Descreveremos neste capítulo os princípios relacionados ao processo de segmentação e o rastreamento de mãos e face, tais como espaços de cor, segmentação de cor de pele, filtros de Kalman e análise de componentes principais. Nosso trabalho de reconhecimento de faces com PCA e filtros preditivos serão baseados nos trabalhos e estudos já desenvolvidos pelo Laboratório de Tecnologias de Interação (LaTIn) do Departamento de Ciência da Computação do IME-USP.

Após a detecção das mãos e da face do usuário, o rastreamento desses objetos é realizado utilizando filtros de Kalman e, para aumentar a robustez do sistema, implementamos também um filtro para consistência de faces baseado na análise de componentes principais dos objetos detectados.

3.1 Cor

O conceito de "cor", embora aparentemente imediato, envolve muitas noções de ordem física, psicológica e fisiológica, pois depende da iluminação do ambiente, da cor da região em volta do objeto e do sistema de visão humano, o que dificulta enormemente sua modelagem, como explicado por Foley *et al.* [22]. Embora os comprimentos de onda na faixa de luz visível sejam grandezas contínuas, quando um ser humano observa tais comprimentos de onda, estes são imediatamente agrupados em classes, fazendo uso de noções psicológicas e exemplos prévios de cada indivíduo na separação entre cada uma das cores visíveis, ou na forma como é usualmente colocado, na classificação de cores. Transportando tal problema ao universo computacional, o processo de classificação de cores pode ser visto como a aplicação de funções altamente não-lineares a uma função de intensidades luminosas sobre os diversos pixels que compõem uma imagem digital, visando adequá-los a parâmetros conhecidos (classes).

A definição quantitativa de cor que permite associar um conceito como "vermelho" a uma quantidade física exige a escolha de um modelo apropriado. Gomes e Velho [74] descrevem como essa escolha depende do contexto da aplicação, pois um modelo válido perceptualmente pode gerar resultados não precisos quando usado como um modelo computacional. A ausência ou presença de luz é o que causa a sensação de cor. Luz é um fenômeno natural, mas a percepção de cor depende da interação da luz com o nosso sistema visual e por isso é um fenômeno psicofísico.

O ser humano percebe cores através de diferentes processos psico-químicos. Pode-se ainda definir três processos de formação de cor: aditivo, subtrativo e pigmentação, que precisam ser entendidos para a criação de modelos matemáticos adequados.

1. Formação de cor aditiva: Isaac Newton, físico inglês, verificou experimentalmente que a luz branca é composta de componentes coloridos e que o prisma simplesmente as separa,

diferentemente do que se acreditava na época, que o prisma criava as cores.

2. Formação de cor subtrativa: cor subtrativa é formada quando a luz vista é transmitida através de um filtro de luz ou de uma tintura. Um filtro de luz é um material parcialmente transparente e sólido que absorve parte da luz que recebe e transmite o resto dependendo do comprimento da onda. Portanto, um filtro vermelho deixa passar apenas a parte vermelha do espectro, bloqueando todas as outras partes. Tintura é um líquido colorido ou então uma solução ou uma dispersão de um componente colorido. Como um filtro, bloqueia a radiação dependendo de sua cor.
3. Formação de cor por pigmentação: um pigmento consiste de partículas coloridas em dispersão em um líquido, ou espalhadas sobre uma superfície. Essas partículas podem absorver, refletir ou transmitir a luz que chega até elas.

O processo de captura de imagens através de uma câmera se dá pelo método aditivo. Através desse processo é possível utilizar cores como características representativas para objetos de interesse nas imagens capturadas. É tarefa inicial da visão computacional separar algumas cores - que representem objetos de interesse nas imagens amostradas - das demais cores (fundo da imagem). Contudo, muitas questões permanecem abertas no trato com cores, tais como a forma de representação das cores e as ferramentas ideais capazes de promover a separação desejada.

3.1.1 Segmentação de cor de pele

Diferentemente do que a intuição nos diz, pessoas de diferentes raças possuem aproximadamente a mesma cor de pele. A diferença reside na intensidade da cor ao invés da cor propriamente dita, como mostra Yang *et al.* [35]. Este trabalho também demonstra que a cor de pele forma aglomerados que podem ser caracterizados por uma distribuição normal

no espaço de cor normalizado. Diferenças de cor de pele entre pessoas podem ser reduzidas normalizando a intensidade e controlando a iluminação.

A escolha da representação do espaço de cores cumpre um papel fundamental no desempenho da aplicação considerada, determinando aspectos como eficiência e robustez.

Há diferentes tipos de espaços para representação de cor. Na literatura podemos encontrar trabalhos na área de segmentação de cor, mais especificamente cor de pele, utilizando alguns os formatos disponíveis. Dependendo das restrições do ambiente, condições de iluminação e do tipo de câmera, determina-se o formato apropriado. Martinkauppi [48] apresenta um relato de diversos trabalhos na literatura em segmentação de cor de pele utilizando diferentes espaços de cor. Terrillon e Akamatsu [71] também apresentam um estudo comparativo de espaço de cor. O fator mais importante, de acordo com eles, é encontrar o melhor modelo que distingua em um dado espaço, o que representa ou não cor de pele.

Yang *et al.* [35] mostram que a atenuação da diferença de intensidade e conseqüente agrupamento das cores podem ser caracterizados por distribuições normais, que podem ser calculados normalizando-se as cores no espaço RGB.

A composição da cor de pele é restrita. Esta composição pode ser descrita como sendo uma combinação de sangue (vermelho) e melanina (amarelo) de acordo com Fleck *et al.* em [21]. Se o espaço RGB for usado para representação de cor de pele, além da cromaticidade, estará representada também a sua luminância, que pode variar ao longo da superfície da pele de uma pessoa devido a luz ambiente. Chang e Robles [9] e Cai *et al.* [7] mostram que a luminância não é apropriada para detectar cor de pele, e ela pode ser removida da representação tradicional RGB utilizando-se cores normalizadas.

De acordo com Yang e Waibel [34], é possível normalizar o espaço de cores RGB para o chamado espaço cromático, também conhecido por “cores puras”, com as seguintes equações:

$$r = \frac{R}{R + G + B} \quad (3.1)$$

$$g = \frac{G}{R + G + B} \quad (3.2)$$

Figura 3.1: Amostra de cor de pele de uma das pessoas utilizada para treinamento.

Essa normalização reduz o espaço de cores em uma dimensão. Outros métodos incluem a utilização de espaços de cor diferentes do espaço RGB e do espaço cromático. Sob condições normais de iluminação, as componentes de cor r e g agrupam-se em pontos específicos, como pode ser observado no histograma de cor 3.2 da imagem 3.1 contendo apenas cor de pele obtidos a partir das imagens usadas para treinamento.

Figura 3.2: Histograma de distribuição de cor r (equação 3.1) e g (equação 3.2) da figura 3.1.

O uso da cor da pele para detecção de faces e mãos em imagens e seqüências de imagens é robusto sob condições variadas de intensidade de iluminação e computacionalmente eficiente. Também é robusto no caso de oclusão e mudança de escala e orientação como descrito por Störing *et al.* em [65]. Um dos pontos fracos é a mudança de iluminação, que pode afetar o resultado final do algoritmo. De acordo com Birgitta em [49], uma solução é transformar o espaço RGB em diferentes espaços de cores onde a intensidade e cromaticidade podem ser analisadas separadamente. Yang e Waibel [34] também resolve este problema de brilho convertendo o espaço de cores RGB para o espaço cromático.

3.2 Métodos de Segmentação de Pele baseados em Cor

Os métodos mais comuns na literatura para segmentação de objetos baseados em cor podem ser classificados como empíricos e estatísticos. Chai e Ngan [8] e Hsu e Jain [31] apresentam técnicas empíricas que se mostram eficientes para o rastreamento de cor de pele. Estas técnicas são baseadas na comparação dos *pixels* com um certo intervalo de valores que definem cor de pele em um determinado espaço de cor.

3.2.1 Método Empírico

A detecção de um determinado pixel como cor de pele pode ser feita através de um método simples como a limiarização de sua distância ao modelo de cor.

Neste método, o treinamento é efetuado calculando-se uma média dos valores (3.1 e 3.2) de uma região conhecidos *a priori* pertencer à pele do usuário a ser rastreado. Estes valores serão usados na limiarização. Um fator importante também são as condições de iluminação do ambiente, que são medidas e devem ser as mesmas no momento da captura deste conjunto de treinamento e durante a execução do algoritmo. Os parâmetros intrínsecos da câmera

como brilho, contraste, abertura e ganho também devem ser os mesmos tanto no momento da captura quanto durante a execução do algoritmo.

Portanto, dada uma imagem de treinamento que contém apenas pixels de cor de pele, a média dos valores de (3.1) e (3.2), chamados agora de $\bar{r}$ e $\bar{g}$ respectivamente é calculada.

Impondo-se um limiar x aos valores $\bar{r}$ e $\bar{g}$, temos que a função $\lambda(r, g)$ é definida por:

$$\lambda(r, g) = \begin{cases} Pele & \text{se } \bar{r} + x > r > \bar{r} - x \text{ E } \bar{g} + x > g > \bar{g} - x, \\ \overline{Pele} & \text{caso contrário} \end{cases} \quad (3.3)$$

Figura 3.3: Resultados da aplicação da função $\lambda(r, g)$ com diferentes valores para x . a) $x = 0.1$ b) $x = 0.2$ c) $x = 0.3$

Os resultados são satisfatórios com o valor de 0.4 para o limiar obtido empiricamente. Nas figuras 3.3 e 3.4 temos o resultado da segmentação com a aplicação da função $\lambda(r, g)$ com diferentes valores para x . A imagem 3.5 é a representação original em RGB que foi capturada pela câmera.

3.2.2 Método da Máxima Verossimilhança

Nesta seção iremos descrever a implementação realizada com base no trabalho de Yang e Waibel [34]. Pelo que se pode observar na figura 3.2.2, a distribuição de cor de uma região de cor

Figura 3.4: Resultados da aplicação da função $\lambda(r, g)$ com diferentes valores para x . d) $x = 0.4$ e) $x = 0.5$ f) $x = 0.6$

Figura 3.5: Imagem usada como entrada para teste dos algoritmos de segmentação

de pele pode ser aproximada por um modelo Gaussiano. Yang e Waibel também observaram que distribuições para diferentes “pedaços” de pele de diferentes pessoas apresentam a mesma forma no espaço de cor cromático.

Além disso, mesmo quando as condições de iluminação do ambiente são alteradas depois da análise do conjunto de treinamento, a distribuição de cor de pele sob diversas iluminações são semelhantes. Portanto, de acordo com Yang, a distribuição de cor de pele pode ser representada por um modelo Gaussiano, como foi demonstrado em [35]. A distribuição de cor de pele é aproximada por uma gaussiana com parâmetros:

$$N(m, \sum^2) \quad (3.4)$$

Figura 3.6: Distribuição de cor de pele no espaço r g

onde $m = \begin{bmatrix} \bar{r} & \bar{g} \end{bmatrix}$ com média e covariância calculados da seguinte forma:

Média:

$$\bar{r} = \frac{1}{N} \sum_{i=1}^N r_i \quad (3.5)$$

$$\bar{g} = \frac{1}{N} \sum_{i=1}^N g_i \quad (3.6)$$

Covariância:

$$\Sigma = \begin{bmatrix} \sigma_{rr} & \sigma_{rg} \\ \sigma_{gr} & \sigma_{gg} \end{bmatrix} \quad (3.7)$$

onde σ_{ab} é a covariância de a e b .

Dado então este modelo, a função de *Máxima Verosimilhança* de cada pixel pode ser definida da seguinte forma:

$$MV(r, g) = e^{-0.5(x-m)'\Sigma^{-1}(x-m)} \quad (3.8)$$

onde $x = \begin{bmatrix} r & g \end{bmatrix}$.

Dessa forma temos que este modelo transforma uma imagem de entrada colorida em uma imagem em tons de cinza, onde cada pixel representa a probabilidade do mesmo ser cor de pele. Quanto maior a probabilidade, maior o valor nominal dele.

Os resultados são apresentados nas figuras 3.7 abaixo. Percebe-se que os resultados não são tão satisfatórios. Devido a semelhança da cromaticidade do fundo da sala com a cromaticidade da cor de pele, partes da sala são reconhecidas como sendo pele.

Figura 3.7: Função de Máxima Verossimilhança 3.8 aplicada a imagem de entrada (*a*, *b*, *c* e *d*) e seu respectivo Limiar (*e*, *f*, *g* e *h*). As áreas de intensidade mais clara indicam probabilidade maior de cor de pele

Uma vez detectadas as regiões candidatas a mãos e face através da segmentação de cor de pele, utilizamos filtros de Kalman para o rastreamento desses objetos na imagem.

3.3 Filtros de Kalman

Em 1960, R. E. Kalman publicou um artigo descrevendo uma solução recursiva para o problema de filtro linear discreto. Desde então, o filtro de Kalman tem sido alvo de extensos estudos e aplicações [36]. Abaixo apresentamos uma breve introdução teórica baseada em [75]. Para uma introdução mais completa e detalhada veja [50].

O filtro de Kalman se endereça ao problema de tentar estimar o estado de $x \in R^n$ de um processo de controle discreto que é governado pela equação linear de diferença estocástica definida por

$$x_k = A_{k-1}x_{k-1} + w_{k-1}, \quad (3.9)$$

$$z_k = Hx_k + v_k, \quad (3.10)$$

As variáveis aleatórias w_k e v_k representam o ruído da previsão do estado e o ruído da observação do estado, respectivamente. A matriz H é a matriz de observação e A a matriz de transição, que modela o comportamento do sistema no decorrer do tempo. z_k é a observação do sistema no tempo k e x_k é o estado do sistema no tempo k . Assume-se que x_k e z_k são independentes uma da outra e com distribuição de probabilidade normais

$$\begin{aligned} p(w) &\equiv N(0, Q) \\ p(v) &\equiv N(0, R) \end{aligned} \quad (3.11)$$

O filtro de Kalman estima um processo, ou estado de um processo, utilizando um controle de retorno: o filtro estima o estado deste processo em um determinado momento e obtém um retorno na forma de medidas de ruídos. Desta forma, as equações para o filtro de Kalman caem em dois grupos: equações de atualização de tempo e equações de atualização de medidas. As

equações de tempo são responsáveis pela projeção no futuro do estado corrente e também pela estimativa da matriz de covariância do erro para obter estimativas *a priori* para o próximo passo. As equações de atualização são responsáveis pelo retorno, isto é, pela incorporação das novas medidas na estimativa obtida *a priori* para obter uma estimativa *a posteriori* melhorada. As equações de tempo também podem ser encaradas como equações de previsão enquanto que as equações de medida podem ser encaradas como equações de correção. No nosso caso, estamos estimando as posições de mãos e face do usuário. Para cada um dos objetos, estamos considerando o seu centro de massa como o localizador do mesmo. Por exemplo, os parâmetros estimados para o objeto face, são suas coordenadas (x, y) . Desta forma, o nosso sistema implementa 3 filtros de Kalman distintos, um para cada objeto rastreado: face, mão esquerda e mão direita.

Antes de inicializar os filtros de Kalman, é necessário verificar se os objetos a serem rastreados realmente correspondem a um usuário. Para isso, verificamos se o *blob* correspondente à face se assemelha a um modelo de face através de uma técnica baseada em PCA, também conhecida por *eigenfaces*.

3.4 Eigenfaces

Este método desenvolvido por Turk *et al.* [73, 72] transforma imagens de faces em um conjunto reduzido de imagens de características, chamadas de *eigenfaces*. As características codificadas correspondem a variação média das faces obtidas a partir das imagens de treinamento. A princípio, o trabalho proposto por Turk *et al.* [73] trata o reconhecimento de faces a partir de um conjunto de treinamento mas a solução proposta para detecção de faces é computacionalmente cara. Este problema é solucionado em [72] por Turk *et al.* utilizando-se um método mais eficiente que será descrito a seguir.

A partir deste conjunto de imagens de face usadas para treinamento, as *eigenfaces* são obtidas. Elas são as componentes principais deste conjunto ou espaço de faces. Cada imagem deste conjunto pode ser definido como uma combinação linear das *eigenfaces*. Matematicamente falando, estas *eigenfaces* correspondem aos auto-vetores da matriz de covariância do conjunto de imagens de face.

Para classificarmos uma imagem como face é necessário primeiro a construção das *eigenfaces* a partir de um conjunto de imagens de treinamento. O número de imagens de treinamento costuma ser pequeno, como mostrado por Turk *et al.* [73]. Como temos o espaço das *eigenfaces*, dada uma imagem a ser classificada, esta imagem deverá ser projetada neste espaço.

Turk propõe uma comparação da distância euclidiana entre a imagem de entrada e as imagens de treinamento para a classificação de faces. No nosso caso, utilizamos uma análise diferente da proposta por Turk. Nós fazemos uma comparação entre as diferentes componentes conexas encontradas na imagem. Aquela que apresentar a menor distância ao espaço de faces, dentro de um limiar para o qual as faces estão definidas, será classificada como face.

3.4.1 Calculando Eigenfaces

Dada uma imagem definida por $I(x, y)$ de tamanho $N \times N$, definimos o vetor Γ como sendo o vetor de dimensão N^2 desta imagem I . Uma imagem típica de 256×256 descreve um vetor de dimensão 2^{16} , ou equivalentemente, um ponto no espaço 2^{16} -dimensional. Um subconjunto das imagens de treinamento, por exemplo, pode ser mapeado para uma coleção de pontos neste espaço.

A matriz de covariância C do espaço de faces do conjunto de treinamento $I = [I_1, I_2 \dots I_M]$ pode ser calculada da seguinte forma:

$$C = \frac{1}{M} \sum_{n=1}^M \Phi_n \Phi_n^T \quad (3.12)$$

onde

$$\begin{aligned}\Phi_i &= I_i - \Psi \\ \Psi &= \frac{1}{M} \sum_{n=1}^M I_n\end{aligned}\tag{3.13}$$

C também pode ser escrita da forma $C = AA^T$, onde $A = [\Phi_1, \Phi_2 \dots \Phi_M]$. Esta matriz C é uma matriz de dimensão $N^2 \times N^2$.

Como em geral $M < N^2$, então existirão apenas $M - 1$ auto-vetores significativos sendo que os demais terão auto-valores associados nulos. Logo basta calcular os auto-vetores significativos a partir de uma matriz $M \times M$. As imagens com os maiores autovalores associados formam o espaço de faces. Sejam os auto-vetores v_i de AA^T :

$$A^T A v_i = \lambda v_i\tag{3.14}$$

Multiplicando por A dos dois lados

$$AA^T A v_i = \lambda A v_i\tag{3.15}$$

Então, $A v_i$ são os auto-vetores de $AA^T = C$. Logo, basta encontrar os auto-vetores v_l para $l = 1..M$ mais significativos de $L = A^T A$.

A partir destes vetores v_l obtemos os vetores u_l de C

$$u_l = \sum_{n=1}^M v_n \Phi_n, l = 1..M\tag{3.16}$$

Estes auto-vetores u_l são chamados de *Eigenfaces* devido a aparência de face que assumem ao serem salvos na forma de imagem.

3.4.2 Detecção de Faces

Após o cálculo do espaço de faces, definido pelas *eigenfaces*, a detecção de faces é feita projetando-se a imagem de entrada neste espaço. Consegue-se isto transformando-se esta nova imagem em seus componentes *eigenfaces*.

Esta operação é feita da seguinte forma. Primeiro

$$\omega_k = u_k^T (\Gamma - \Psi) \quad (3.17)$$

para cada um dos $k = 1, \dots, M'$, onde Ψ é a imagem média do conjunto de treinamento, Γ é a imagem de entrada e u são os *eigenvectors* calculados a partir do conjunto de treinamento. Estes pesos, w_i , formam um vetor de padrões $\Omega^T = [w_1, w_2 \dots w_{M'}]$ que representam a contribuição de cada *eigenface* na representação da imagem Γ .

A partir dos valores de w_i pode-se calcular a distância euclidiana entre a imagem de entrada ajustada pela média $\Phi = \Gamma - \Psi$ e a projeção desta imagem no espaço de faces $\Phi_f = \sum_{n=1}^M w_i u_i$. Esta distância é dada por:

$$\epsilon^2 = \| \Phi - \Phi_f \|^2 \quad (3.18)$$

Nosso sistema utiliza como medida de comparação entre diferentes objetos, a distância euclidiana mostrada em 3.18. O objeto que apresentar a menor distância, é considerado face.

3.5 Conclusão

A normalização do espaço de cores RGB se mostrou uma alternativa viável para a minimização dos problemas ocasionados pela variação de iluminação de ambientes de RV. No espaço normalizado, a segmentação de cor de pele através do método empírico foi a que apresentou melhores

resultados e será utilizada para a implementação do protótipo.

Filtros de Kalman e análise de componentes principais são técnicas robustas, utilizadas por diversos trabalhos na literatura para rastreamento e identificação de faces, respectivamente. A partir de análise de componentes principais é possível identificar qual dos objetos encontrados na imagem é a face do usuário. Localizando-se a face, assume-se que as mãos esquerda e direita são, respectivamente, os objetos a esquerda e direita da face.

O próximo capítulo descreve a implementação do sistema utilizando as técnicas mostradas neste capítulo, além de algumas outras técnicas que se mostraram necessárias para se obter um melhor desempenho.

Capítulo 4

Sistema Interativo Baseado em Gestos

Neste capítulo descreveremos o processo de reconhecimento de gestos utilizado pelo sistema implementado. As técnicas descritas no capítulo anterior servem de base teórica para o sistema desenvolvido. Aspectos de implementação do sistema como posicionamento de câmera, inicialização do sistema, segmentação e rastreamento, e definição da interface baseada em gestos são descritos.

4.1 Visão Geral

O sistema implementado tem por finalidade reconhecer gestos em ambientes imersivos distintos, com alguma variação de iluminação. Cada ambiente possui suas particularidades. A extensão a outros locais pode permitir o uso do sistema em ambientes mais convencionais como por exemplo a sala de TV de uma residência.

Alguns aplicativos foram desenvolvidos especificamente para avaliar e testar o sistema de reconhecimento de gestos. Os testes foram realizados em 2 ambientes, CAVE e Quiosque.

A figura 4.1 mostra um diagrama de blocos correspondente ao processo de captura, seg-

mentação, rotulação, rastreamento e identificação dos gestos.

Figura 4.1: Fluxo de dados do sistema

4.2 Segmentação e Rastreamento de Mãos e Face

O processo de segmentação necessita de uma fase de treinamento para definição do modelo estatístico de cor de pele que é utilizado durante a execução do sistema.

4.2.1 Treinamento do modelo de cor de pele

Para o treinamento do modelo de cor de pele foram utilizadas imagens de 16 pessoas de diversas etnias. Destas imagens foram retiradas manualmente pedaços de cor de pele para formar o conjunto de treinamento como descrito na seção 3.1.1.

O modelo de cor de pele é obtido utilizando o método empírico descrito na seção 3.2.1, que é mais robusto à mudança de iluminação que o espaço RGB. Este modelo de cor de pele é aplicado sobre a imagem RGB de entrada, e o resultado obtido é uma imagem binária que corresponde a imagem original com as áreas de cor de pele segmentadas.

4.2.2 Segmentação de Cor de Pele

Durante a execução do sistema as imagens a serem segmentadas são pré-processadas por um filtro mediana 3×3 e suavizadas por um filtro gaussiano 5×5 . Pixels com cor de pele são classificados utilizando o modelo definido na fase de treinamento, obtendo assim uma imagem binária.

Operações morfológicas de erosão e dilatação, nesta ordem, são aplicadas nesta imagem binária para eliminação de ruídos. Os filtros de erosão e dilatação utilizados foram de 5×5 . Veja na figura 4.2 os resultados das operações morfológicas de erosão e dilatação. Estas operações têm por objetivo reduzir os ruídos na imagem, destacando apenas os pixels de cor de pele.

A imagem filtrada pelos operadores morfológicos é utilizada como entrada para o algoritmo de determinação de componentes conexas. Para cada componente, são calculados o centróide e o quadrilátero delimitador. Dentre estas componentes conexas, as que possuem área muito pequena (em torno de 40 pixels) são descartadas. As restantes passam por um processo de rotulação para identificar a face e as mãos.

Figura 4.2: A partir do canto superior esquerdo em sentido horário: imagem original, segmentação de cor de pele por limiarização, operação de erosão e operação de dilatação.

4.2.3 Rotulação das componentes conexas

Para identificar cada uma das componentes conexas restantes após a eliminação de ruídos, assumimos que o usuário se posiciona à frente da câmera com suas mãos ao nível do peito. Isto facilita a inicialização do rastreamento, pois podemos restringir a busca da face e das mãos a regiões chave, como mostra a figura 4.3. Apenas as componentes conexas cujos centróides se encontrem dentro das regiões chave são utilizadas. O rastreamento é inicializado quando existirem componentes conexas adequadas em cada uma das regiões chave.

Para a face, a componente conexa que mais se aproxima do modelo de face utilizando o algoritmo PCA é utilizada para rastreamento. Adotamos o mesmo modelo de face desenvolvido por Régis de Abreu Barbosa, outro aluno do LaTIn, que vem pesquisando técnicas de

Figura 4.3: Posição inicial do sistema: mãos na altura do peito. a) CAVE b) Quiosque.

detecção de faces. Ele utiliza um banco de imagens com 16 faces, das quais as 8 imagens mais significativas foram utilizadas para a construção do espaço de faces, de acordo com a técnica de Eigenfaces descrita na seção 3.4.

Um candidato a face é definido pelo quadrilátero delimitador de uma componente conexa. Se um quadrilátero corresponde à imagem de uma face, então a distância entre esta imagem e sua projeção no espaço de faces deve ser pequena (como visto na seção 3.4.2). O candidato que apresentar a menor distância euclidiana é utilizado para rastreamento.

Para as mãos, a componente conexa de maior área é utilizada para rastreamento. Assim que a face é encontrada, as componentes conexas dentro de suas respectivas regiões chave a direita e a esquerda da face são rotuladas, respectivamente, como mão direita e mão esquerda.

Além da rotulação, encontrar a face através do PCA permite que as regiões de cor de pele desta componente conexa sejam utilizadas para refinar o modelo de cor de pele. Desta forma, o modelo de cor de pele estará levando em conta a cor de pele do usuário que está utilizando o sistema e as condições de iluminação do ambiente. Este sistema de adaptação de cor de pele, personalizado para cada usuário, ajuda a reduzir ruídos e pixels que são falsos candidatos a

cor de pele.

4.2.4 Rastreamento de mãos e face

O rastreamento é iniciado assim que todos os objetos são rotulados. Um filtro preditivo linear (filtro de Kalman), por objeto, é utilizado para rastreamento. O filtro de Kalman (como visto na seção 3.3) permite estimar a posição do objeto no próximo passo de processamento. A posição dos objetos detectados em uma imagem são utilizados para atualizar os parâmetros de observação do filtro na equação 3.10.

Um dos principais problemas que este tipo de filtro apresenta é que eles não são robustos a movimentos que levem a oclusão entre objetos. Devido à natureza dos gestos usados na comunicação do ser humano, as mãos e face se ocluem e se sobrepõem com frequência. Oclusões acontecem, por exemplo, quando uma das mãos se esconde atrás do corpo do usuário ou então de seu próprio braço.

A fim de manter a rotulação de cada objeto durante o rastreamento, utilizamos um grafo de consistência espaço-temporal, descrito na figura 4.4.

As letras H , L e R representam, respectivamente, a face, a mão esquerda e a mão direita do usuário. Quando o sistema está em seu estado inicial, a predição do filtro de Kalman define, para cada objeto, uma região de busca para a próxima imagem. Os nós superiores do grafo de consistência representam os estados de sobreposição de mãos e face e os nós inferiores representam os estados de oclusão, onde um ou mais objetos “desaparecem”.

Com relação aos estados de oclusão, consideramos que a face nunca é ocluída, apenas as mãos podem sofrer oclusão por parte do corpo ou membros do usuário. Uma vez que R ou L sejam ocluídos pelo corpo ou membros do usuário, o respectivo filtro de Kalman não é atualizado. O sistema espera então por alguns segundos para que este objeto “desaparecido” reapareça na mesma região que ele sumiu ou então próximo dela. Se o objeto não for detectado

Figura 4.4: Grafo de consistência espaço-tempo. Os valores dentro de cada estado representam os objetos sendo rastreados.

dentro deste período, o sistema procura por um outro objeto próximo da região onde ele desapareceu. Somente quando H desaparece, o rastreamento não é capaz de se recuperar. Neste momento, o sistema espera que o usuário se coloque na posição inicial novamente.

Quando as regiões de busca de objetos diferentes se sobrepõem, uma única região de busca é utilizada (ver figura 4.5) para prever a posição de ambos os objetos. Esta região é o quadrilátero delimitador da união das duas regiões. Quando um único objeto é encontrado dentro desta região, ocorre a mudança de estados, significando que dois objetos se aglutinaram. O

conteúdo dos nós que correspondem aos estados de oclusão mostram as situações dos objetos, por exemplo, o objeto $L + R$ corresponde a aglutinação de L com R . O valor X indica uma situação de dúvida entre R e L .

Figura 4.5: Apenas uma região de busca quando há sobreposição de objetos

A situação mais comum encontrada durante a interação é quando as duas mãos se aglutinam, formando um objeto $L + R$. Este objeto pode se dividir em dois (R e L) e voltar para o estado inicial. Ao se dividir, o objeto mais a esquerda é considerado L e o outro objeto é considerado R . É possível que a mão esquerda e a mão direita não correspondam a L e R , respectivamente. É possível que L seja encontrado à direita de R mas o sistema não tolera esta configuração por mais de 3 segundos. Após esse período de tolerância, os rótulos são trocados para que R esteja à direita de L .

Uma das mãos ou até mesmo as duas mãos podem ser levadas a face. Logo, os objetos L e H ou então R e H podem se sobrepor. Quando isso ocorre, um novo estado do grafo é atingido, com o objeto denominado $H + L$ ou $H + R$. Um objeto $H + R$ ou $H + L$ pode se dividir em (H e R) ou (H e L), respectivamente, retornando o sistema para o estado inicial. Como a face é menos ativa que as mãos é esperado que o usuário leve as mãos a face e também mova suas mãos para longe da face e não o contrário (mova a face em direção as mãos). Então,

quando o objeto $H + R$ ou $H + L$ se separa, da mesma forma como descrito acima para a separação do objeto $R + L$, os objetos dentro da região de busca serão renomeados da seguinte forma: aquele que estiver se movendo com velocidade maior da posição anterior a separação é nomeado de L ou R .

Caso o objeto $H + R$ se sobreponha ao objeto L ou o objeto $H + L$ se sobreponha ao objeto R , ocorre a mudança para o estado $H + R + L$. Quando este objeto se divide em dois, o sistema não tem como saber se apenas uma das mãos se separou ou então se as duas mãos estão se separando de H , mas juntas. Devido a essa ambigüidade o sistema muda para o estado $H + X$ e $L + R$. Somente quando um destes dois objetos se divide novamente é que o sistema pode voltar ao estado inicial. O processo de separação de $H + X$ e $L + R$ é o mesmo descrito anteriormente para o caso $H + R$ ou $H + L$ e $L + R$.

Veja na figura 4.6 alguns exemplos dos estados que o grafo identifica.

4.3 Interpretação de Gestos

As informações obtidas a partir do rastreamento de mãos e face, ou seja, as coordenadas dos centróides destes objetos, são utilizadas para a interpretação de gestos. Desenvolvemos uma interface baseada em gestos capaz de identificar determinados gestos pré-definidos. Gestos pré-definidos são posições específicas que as mãos e a face do usuário podem atingir ao realizar um movimento. Estes gestos são utilizados para interação em ambientes de RV. Nosso trabalho é baseado na técnica de manipulação manual proposta por Bowman *et al.* [5]. Esta técnica consiste no mapeamento da informação dos movimentos do usuário para manipular objetos e o cursor, o que torna esta técnica mais fácil e intuitiva de aprender.

Figura 4.6: Rotulação de situações que o grafo de consistência identifica.

4.3.1 Interface

Os gestos pré-definidos que podem ser utilizados pelo usuário são executados com as mãos esquerda (L) e direita (R). Estas posições serão usadas para descrever comandos, de acordo com o ambiente e aplicativo, como será descrito no próximo capítulo.

Ambas as mãos podem atingir cinco posições chave diferentes: Descanso (ou posição inicial), Norte, Sul, Leste e Oeste, todas relativas a posição da face (F). Estas posições são identificadas pelos centróides dos objetos correspondentes. O centróide de um objeto é dado por (γ_x, γ_y) em coordenadas de tela, onde γ pode ser H , R ou L , respectivamente para face,

mão direita e mão esquerda do usuário. Para a mão esquerda (L), as posições chave são atingidas quando ocorrem as seguintes situações:

Figura 4.7: a) Posição Inicial b) Posição Norte c) Posição Leste

Figura 4.8: d) Posição Sul e) Posição Oeste

1. Descanso: Dada pela posição de (L_x, L_y) quando o sistema de rastreamento é inicializado (veja figura 4.7a);
2. Norte: Dada quando L_y é maior do que H_y (veja figura 4.7b);
3. Sul: Dada quando L_y é menor do H_y (veja figura 4.7c);
4. Leste: Dada quando L_x é maior do H_x (veja figura 4.8a);

5. Oeste: Dada quando L_x é menor do H_x (veja figura 4.8b).

A mão direita controla o movimento do cursor. Desta forma, dependendo do modo ativado pela mão esquerda, a mão direita irá controlar o cursor e, além disso, controlar o respectivo modo.

Com a definição desta interface, é possível descrever comandos para ambientes de RV.

4.3.2 Comandos

O sistema de interface de gestos desenvolvido simula, tanto em modo Quiosque quanto em modo CAVE, uma interface de “apontar e clicar” (point and click interface). A mão direita (R) do usuário funciona como um controlador da posição do cursor e a mão esquerda (L) funciona para ativação de outros comandos, como um clique, uma mudança de estado da interface e outros comandos que serão apresentados. Dessa forma, as posições definidas na seção anterior são combinadas de forma a disponibilizar ao usuário uma gama de comandos que possam ser utilizados nas aplicações disponíveis. As analogias destes comandos são descritas a seguir.

Com a mão esquerda (L), para permitir que o usuário possa descansar seus braços, as posições Sul e Descanso não são utilizadas para ativar comandos. A posição Norte é usada para mudança do modo que controla o comportamento do cursor e usada também para navegação, com comandos como translação, rotação e escala. A posição Oeste é usada para mudança de modo dentro de um modo específico selecionado como, por exemplo, dentro do modo de rotação, temos rotação x-y, rotação x-z ou rotação y-z. A posição Leste é usada para simular um clique.

A mão direita (R) é usada para controlar as operações selecionadas pela mão esquerda (L). Também é utilizada para controle da posição mouse/cursor. Com isso, é necessário mapear o movimento de R para corresponder aos movimentos do mouse. Assumindo-se que o objeto representativo da mão direita varia uniformemente quando o usuário movimenta sua mão em

um quadrilátero arbitrário, confinado ao lado direito de seu corpo, o centróide de R irá variar monotonicamente através da tela.

Este quadrilátero pode ser mapeado para as coordenadas do espaço de tela a partir das coordenadas do espaço de imagem usando a seguinte transformação:

$$\begin{aligned} [x_a, y_a, w_a] &= [x, y, w]A \\ x_s &= \frac{x_a}{w_a} = \frac{A_{11}x + A_{21}y + A_{31}}{A_{13}x + A_{23}y + A_{33}} \\ y_s &= \frac{y_a}{w_a} = \frac{A_{12}x + A_{22}y + A_{32}}{A_{13}x + A_{23}y + A_{33}} \end{aligned} \quad (4.1)$$

onde A é uma matriz de transformação 3×3 , (x_s, y_s) são as coordenadas da tela e (x, y, w) são as coordenadas de imagem [41]. A matriz A é calculada a partir de um procedimento de calibração. A partir deste procedimento é que se pode inferir os valores da matriz A .

Como cada usuário tem características físicas diferentes, um procedimento de calibração é utilizado para que o sistema se adapte às diferentes proporções anatômicas de cada usuário. Este procedimento é executado depois do início do sistema de rastreamento mas antes do início do sistema de reconhecimento de gestos. A calibração consiste na aquisição de quatro posições da mão direita do usuário que devem corresponder aos cantos da tela. Estas quatro posições definirão a região de atuação da mão direita. Portanto, enquanto o usuário movimentar a mão direita dentro desta região, o cursor se movimentará de acordo com a transformação descrita na equação 4.1. Quando o usuário atinge ou passa o limite delimitado por esta região, o cursor não se movimenta. As figuras 4.9 e 4.10 mostram as quatro posições chave que foram capturadas para mapeamento da mão direita, que é utilizada como cursor, utilizadas para interação no ambiente Quiosque.

A partir destas quatro posições pode-se inferir o valor de A , pelos pares $(x, y) \longleftrightarrow (x_s, y_s)$,

fazendo:

$$\begin{aligned}
 (x_0, y_0) &\longleftrightarrow (0, 0) \\
 (x_1, y_1) &\longleftrightarrow (1, 0) \\
 (x_2, y_2) &\longleftrightarrow (1, 1) \\
 (x_3, y_3) &\longleftrightarrow (0, 1)
 \end{aligned}
 \tag{4.2}$$

e considerando coordenadas normalizadas da tela. O apêndice A fornece o procedimento detalhado do mapeamento tela-imagem.

Figura 4.9: a) Posição de Calibração (x_0, y_0) b) Posição de Calibração (x_1, y_1)

Mesmo com os procedimentos de calibração, a movimentação do cursor através da mão direita (R) apresenta tremulações, que prejudicam a utilização da interface. Isso ocorre devido a grande área de visão da câmera, onde pequenos deslocamentos dos objetos na imagem de entrada provocam grandes variações na posição do cursor. Além disso, a região de alcance máximo das mãos é pequeno se comparado a uma tela de alta resolução.

Logo, pequenos erros de estimação na posição dos objetos criam tremores nitidamente grandes no cursor e no controle das aplicações. Para amenizar este comportamento foi utilizado um filtro digital de resposta impulsiva finita (Finite Impulse Response - FIR) com uma

Figura 4.10: a) Posição de Calibração (x_2, y_2) b) Posição de Calibração (x_3, y_3)

componente de frequências passa baixas.

Finite Impulse Response - FIR

O filtro usado na implementação é descrito pela seguinte equação diferencial:

$$a(0) * y(t) = [b(0), b(1), \dots, b(n)]^T [x(t), x(t-1), \dots, x(t-n)] \quad (4.3)$$

onde $n = 5$, $y(t)$ e $x(t)$ são, respectivamente, a saída e a entrada do filtro no instante t , o coeficiente $a(0)$ é 1, e os coeficientes b foram obtidos empiricamente. Os coeficientes a e b , também chamados de pesos, são por definição a resposta impulsiva do filtro. O valor n é a duração da resposta impulsiva do filtro. Os valores utilizados em nossos experimentos para compor o filtro são $b[0..5] = \{0.0174, 0.025, 0.0295, 0.0295, 0.025, 0.0174\}$.

Este filtro reduz de forma significativa o ruído e o tremor do cursor sem comprometer a resposta em tempo real que o sistema oferece em resposta as ações do usuário. Veja o resultado da aplicação do filtro em um trecho da ordenada x de movimentação da mão direita do usuário na figura 4.11.

4.3.3 Limitações do Sistema

O sistema possui as seguintes restrições:

- Apenas um usuário pode utilizar o sistema de cada vez;
- As mãos e face do usuário devem ser as únicas partes do corpo que podem estar visíveis;
- A cor do vestuário não pode se assemelhar com tons de cor de pele;
- As condições de iluminação iniciais devem ser próximas as usadas para a fase de treinamento de cor;
- O usuário deve estar posicionado de frente para a câmera.

Por outro lado:

- O sistema é robusto a pequenas variações de iluminação;
- O fundo pode ser arbitrário;
- O usuário pode ter altura ou dimensões anatômicas bastante variáveis;
- Não utilização de câmeras especiais;
- A câmera não precisa ser calibrada e não precisa estar em altura ou ângulo específico para o usuário, contanto que o usuário esteja de frente para ela.

4.4 Conclusão

Este capítulo descreve a implementação e o funcionamento do sistema de reconhecimento de gestos juntamente com os detalhes de implementação de cada ambiente.

A câmera utilizada é capaz de capturar a 30 quadros por segundo mas para aumentar o tempo de exposição do quadro e conseqüentemente melhorar a imagem, o sistema captura imagens a 15 quadros por segundo, sem que isso influencie na sensação de resposta imediata para o usuário. Além disso, a conversão da imagem em tons de cinza para a imagem colorida é feita através de interpolação, seguindo o padrão do filtro de Bayer.

A segmentação é realizada transformando-se a imagem de entrada para o espaço de RGB normalizado, diminuindo-se um pouco os problemas encontrados com a variação de iluminação dos ambientes. Para identificação inicial dos objetos segmentados, os mesmos são submetidos a análise de componentes principais para determinação do objeto que representa a face do usuário. Com os objetos identificados, o filtro de Kalman prevê as posições dos objetos, restringindo a janela de busca para o próximo quadro.

Os gestos são identificados através de posições chave, atingidas pelas mãos esquerda e direita, e face. A partir destes gestos, uma interface é definida para cada aplicativo e ambiente, proporcionando uma nova forma de interação nestes ambientes.

O próximo capítulo apresenta os resultados experimentais encontrados a partir de experimentos realizados com diversos usuários.

a

b

Figura 4.11: a) Trecho da ordenada x antes da aplicação do filtro e b) Depois da aplicação do filtro.

Capítulo 5

Resultados Experimentais e Discussão

Para permitir uma análise mais detalhada e fundamentada do sistema, alguns experimentos foram realizados. Além destes experimentos, ao longo do desenvolvimento deste trabalho, diversas pessoas participaram de experimentos por curtos períodos de tempo. Este capítulo apresenta os detalhes e os resultados encontrados a partir destes experimentos.

5.1 Descrição da implementação do protótipo

Os aspectos de implementação que foram adotados para o sistema de reconhecimento de gestos serão descritos. Esses aspectos dizem respeito às fases de segmentação e detecção de cor, rastreamento dos objetos encontrados e posterior análise do rastreamento das regiões de cor de pele para reconhecimento de gestos. Antes porém, vale a pena destacar aspectos mais básicos de implementação como hardware, sistema operacional utilizado, versões de bibliotecas e protocolos, de forma a permitir a replicabilidade do sistema.

5.1.1 Equipamentos

O computador utilizado durante a implementação e testes do sistema foi um laptop PC com processador Pentium IV de 2.8Mhz, com 512 MB de memória RAM. O intuito foi de elaborar uma plataforma que oferecesse desempenho suficiente para que o sistema fosse executado em tempo real e que sua mobilidade não fosse prejudicada.

A interface de comunicação *IEEE 1394*, também conhecida por *Firewire* [63] foi utilizada por sua largura de banda. A velocidade de transmissão do padrão *IEEE 1394* de 1995 é de 400Mbps. O *Hotswap* também é uma característica interessante, pois permite a conexão e desconexão das câmeras, sem necessidade da reinicialização do sistema operacional. O tamanho dos cabos pode ser um problema se a câmera ficar mais que 4.5 metros do computador. Mas existem soluções como repetidores de sinal. Além disso a nova especificação *Firewire* chamada de *IEEE 1394b* [64] usa conectores óticos, permitindo um alcance de até 100 metros.

O potencial de banda *Firewire* foi explorado utilizando-se câmeras padrão Firewire 1394 da marca *Point Grey Research* modelo *Dragon Fly* ([33]) (ver figuras 5.1a e 5.1b). Essas câmeras capturam imagens com resolução de 640×480 pixels em tons de cinza. O filtro (*Color Filter Array* CFA) utilizado pelas câmeras para posterior conversão a imagens RGB é o filtro de Bayer [2]. Veja na figura 5.2 como funciona esse filtro.

5.1.2 Software e Algoritmos

O sistema para captura e segmentação foi implementado no ambiente Linux, utilizando-se a distribuição **Fedora Core 2**. Esta distribuição já oferece suporte para interfaces *IEEE 1394*.

As bibliotecas de acesso aos dispositivos *Firewire* utilizadas foram encontradas em repositórios na internet da comunidade de software livre. Essas bibliotecas são robustas e desenvolvidas em C.

Figura 5.1: Câmera *Firewire Dragon Fly*

O sistema foi implementado e testado em um laptop Pentium IV. Este laptop funciona como um servidor de dados, representando os gestos identificados pelo sistema. Tanto a aplicação utilizada na CAVERNA Digital quanto a aplicação utilizada no Quiosque utilizam este laptop como fonte de dados. Um protocolo simples de comunicação de dados foi utilizado, contendo as posições chave identificadas. Desta forma, as aplicações recebem informações como Norte, Sul, Leste, Oeste e as coordenadas (x, y) das mãos, já mapeadas para as coordenadas da tela.

Filtro de Bayer

Grande parte das câmeras digitais utilizam-se de um único sensor CCD (*Charged Coupled Device*) por questões de custo. Como pode-se ver na figura 5.2 apenas uma espectro de cor por pixel é capturado. Como precisamos de três pixels (RGB) para a composição de uma imagem colorida, um algoritmo de reconstrução de cores chamado de *Color Demosaicing* é necessário para gerar as cores faltantes para cada pixel.

A desvantagem desta alternativa é a perda de resolução e o custo necessário para decodificação, pois a informação espacial e a informação de cor estão misturadas para cada pixel. De acordo com Süsstrunk [68], alguns problemas aparecem com o uso desta alternativa como

Filtro de Bayer

Figura 5.2: Máscara posicionada na frente do CCD para captura de quadros.

embaçamento, efeito de *grid* e cores falsas. Veja em (a), (b) e (c) na figura 5.3, os problemas encontrados de embaçamento, efeito de *grid* e cores falsas, respectivamente.

Dentre as alternativas disponíveis para reconstrução da imagem colorida, a implementação de utilizada foi a de *Nearest Neighbors* por ser a de melhor desempenho. Infelizmente observamos o efeito de *grid* nas imagens convertidas mas não foi detectado problemas na segmentação de cor de pele mesmo com este problema.

O vídeo de entrada é composto por quadros RGB de tamanho 640×480 de resolução. Para melhor desempenho do sistema, todas as operações realizadas sobre os quadros utilizam uma resolução menor, de 320×240 .

5.1.3 Desempenho

Utilizando um equipamento relativamente barato para processamento e aquisição de imagens, um laptop Pentium IV, o sistema consegue capturar e rastrear mãos e face a 30 quadros por

Figura 5.3: Problemas na utilização no filtro de Bayer. a) Embaçamento, b) Efeito de *Grid* e c) Cores Falsas

segundo imagens com resolução de 320×240 e 24 bits de cor. Esta velocidade é mais do que aceitável para um protótipo.

Subjetivamente, o rastreamento proporciona uma movimentação *macia* do cursor no ambiente de Quiosque. Já no ambiente CAVE, o rastreamento é prejudicado devido às severas condições de iluminação. Mesmo assim, é possível movimentar o cursor e navegar pelo ambiente virtual.

Apesar de utilizarmos um filtro FIR (veja seção 4.3.2) de tamanho cinco, o atraso gerado não é perceptível pelo usuário, deixando o rastreamento mais estável.

5.2 Experimentos com a interface baseada em gestos

Testamos o paradigma criado em dois ambientes de RV: CAVE e Quiosque. Para estes teste, conduzimos dois experimentos no ambiente Quiosque e um experimento no ambiente CAVE. O propósito destes experimentos é mostrar que a técnica desenvolvida pode ser aplicada para

interação em ambientes como Quiosques e CAVEs, e pode se tornar vantajosa se comparada a outras opções disponíveis. No ambiente Quiosque, seis usuários participaram de dois experimentos controlados. Além destes experimentos, diversos usuários participaram das fases de teste e implementação ao longo do projeto. No ambiente CAVE, apenas o autor realizou experimentos controlados. Vale ressaltar que além destes experimentos com usuários, outros usuários puderam acompanhar a fase de desenvolvimento e testes, que foram cruciais para a melhora tanto em desempenho quanto em usabilidade. Os experimentos realizados tem o objetivo de avaliar a usabilidade do sistema. Os dados foram coletados através de avaliação subjetiva e análise comparativa entre o uso de mouse e o uso de gestos para uma mesma tarefa, utilizando para este último um modelo baseado na lei de Fitts [20]. A lei de Fitts será descrita com maiores na seção 5.2.1. Todos os usuários são estudantes e não receberam incentivo financeiro para participar do teste. A média de idade foi de 25 anos. Todos os estudantes possuíam alguma experiência ou contato com a área de Realidade Virtual e Computação Gráfica. Todos também tem em média 7,3 anos de experiência na utilização de mouse/teclado. Veja na figura 5.4 os participantes do experimento.

A seguir descreveremos os experimentos específicos realizados em cada ambiente: Quiosque e CAVE.

5.2.1 Ambiente Quiosque

A configuração do ambiente Quiosque é composta por um projetor convencional e uma tela de 2×2 metros como pode-se observar nas figuras 5.5a e b. A iluminação neste ambiente pode ser melhor controlada do que na CAVERNA Digital, proporcionando melhores condições para o sistema de reconhecimento de gestos. As imagens capturadas por esta câmera têm o aspecto mostrado na figura 5.6. O usuário interage estando de frente para a câmera. É esperado pelo sistema que ele esteja mostrando ambas as mãos, na altura do peito, de frente para a câmera,

Figura 5.4: Participantes do experimento.

como mostrado na figura 4.3. Observe que a qualidade das imagens é pobre e que o fundo é arbitrário.

Para este ambiente, a interface abaixo foi criada:

- A mão direita (R) controla a posição do cursor;
- Quando a mão esquerda (L) se posiciona na posição Norte, o modo de controle é seqüencialmente alterado. Os modos de controle disponíveis são: escala, rotação e translação;
- Quando (L) se posiciona na posição Oeste, o status do modo de controle atual é seqüencialmente alterado. Por exemplo, se o modo de controle atual for Escala, então o status de modo varia entre Escala X-Y, Escala X-Z e Escala Y-Z;
- Quando (L) se posiciona na posição Leste, um clique é gerado.

Figura 5.5: Ambiente Quiosque.

Exemplificando um modo de operação, por exemplo, quando um L se posiciona na posição Oeste:

- Rotação X-Y passa a ser Rotação X-Z ou;
- Translação X-Z passa a ser Translação Y-Z;

Para analisar o desempenho desta interface vamos comparar o desempenho dos gestos com um dispositivo de apontamento tradicional (mouse) utilizando a lei de Fitts, que será descrita na próxima seção.

Lei de Fitts

A lei de Fitts é utilizada comumente como uma medida usual do desempenho de dispositivos de apontamento. Ela foi proposta em 1954 por Fitts [20] e fornece uma estimativa do tempo necessário para um usuário executar um movimento manual em direção a um alvo em função da distância do alvo e também do tamanho do mesmo. A lei de Fitts original é representada pela equação

Figura 5.6: Amostra de seqüência de imagens colhidas pela câmera posicionada no ambiente Quiosque

$$T = a + b \log_2\left(\frac{2A}{W}\right) \quad (5.1)$$

onde T é o tempo de execução do movimento, A é a amplitude, ou seja a distância percorrida entre um alvo e outro, e W é o tamanho ou diâmetro do alvo. Os coeficientes a e b são determinados empiricamente através de regressão linear. O termo $\log_2\left(\frac{2A}{W}\right)$ é chamado de *Índice de Dificuldade* (ID) da tarefa. Este primeiro modelo de Fitts foi originalmente desenhado para avaliação de desempenho de usuários tocando chapas metálicas.

Mackenzie *et al.* [46] propuseram uma variação da formulação de lei de Fitts que permitia melhor aderência a dados empíricos, ou seja, não permite que ID seja negativo. A equação de

Figura 5.7: Experimento de apontamento: Assim que um usuário ativa um botão (círculo) outro aparece. A diferença entre tamanhos de círculos e distância entre eles é utilizada no modelo da Lei de Fitts

Fitts modificada fica assim

$$T = a + b \log_2\left(\frac{A}{W} + 1\right) \quad (5.2)$$

O inverso do coeficiente b é chamado de *Índice de Desempenho* (IP) e pode ser utilizado para comparação de desempenho entre diferentes dispositivos de apontamento. No nosso caso, mouse e sistema de gestos.

Descrição do experimento

Os usuários foram instruídos a utilizar o sistema de forma natural, sendo enfatizado antes e durante a realização dos experimentos que o sistema estava em teste, e não o usuário participando do experimento. Nenhum dos usuários presenciou o teste dos demais. Cada usuário passou por uma fase de treinamento onde foram descritos os objetivos e o funcionamento do sistema através de demonstração. Em seguida, o usuário era convidado a utilizar o sistema para familiarização. A etapa de familiarização era executada até que o usuário tivesse total domínio da interface e durava de 2 a 4 minutos.

Para esta experimento, o usuário é instruído a ativar botões de diferentes tamanhos em diferentes locais da tela. A medida que o usuário ativa um botão, outro botão aparece para seja ativado. E assim sucessivamente até o final da tarefa. Nos experimentos com os usuários, ao todo, eram ativados 10 botões. O tempo de realização deste experimento foi cronometrado para cada usuário. A ocorrência de erros (ativação do clique quando o ponteiro não está em cima do botão) foi descartadas para a análise deste experimento.

Para uma boa avaliação do sistema, foram colocados alvos de diferentes tamanhos e cujas posições de aparecimento fossem diferentes umas das outras. Com isso, asseguramos que o processo cognitivo necessário para cumprir a tarefa consiga incluir a busca por um determinado alvo, a cada passo, como acontece em aplicações em um ambiente real. Vale lembrar também que a maior variabilidade de alvos e distâncias garante uma maior variabilidade dos parâmetros da lei de Fitts.

A sessão de controle foi realizada da mesma forma, com os mesmos dados de tamanho e distância de alvo, mas com a utilização de mouse. O mouse foi escolhido como dispositivo de comparação na sessão de controle pois a maioria dos usuários de sistemas computacionais possui experiência na utilização do mesmo.

Modalidade	ID e T	A e T	W e T
Mouse	0,3969	0,7969	-0,0601
Gestos	0,4410	0,4049	-0,0133

Tabela 5.1: Correlação Linear entre os coeficientes da Lei de Fitts

Resultados

Para identificarmos em nossos experimentos como o mouse se compara ao sistema de gestos, calculamos o coeficiente de correlação linear entre ID e T, entre A e T e entre W e T, mostrados na tabela 5.1.

Os coeficientes a e b da equação 5.2 foram calculados através de mínimos quadrados a partir da equação $T = a + bID$ para as duas modalidades.

A média do tempo de execução das 10 ativações de botões pelos usuários foi de 7,3 segundos para o mouse e de 26,1 segundos para o sistema de gestos. Os fatores que podem ter levado a um demora maior para o cumprimento da mesma tarefa são a não familiaridade com o sistema em uso diário, ou seja, não houve tempo suficiente para a aquisição de habilidade motora necessária para a nova técnica, e também o ruído apresentado pelo sistema no posicionamento do cursor, dificultando a ativação do botão.

O coeficiente b indica a inclinação das retas. Este índice é positivo em ambos os casos, o que indica que a medida que o nível de dificuldade aumenta, o tempo de execução aumenta para os dois dispositivos. Mas o aumento de dificuldade não é o mesmo em ambos os casos. No caso do mouse, o aumento de dificuldade é mais acentuado, pois o valor de b para mouse é aproximadamente duas vezes o valor de b para o sistema de gestos. Logo, o valor de ID onde as retas se cruzam não está dentro de um patamar utilizável. Isto quer dizer que o sistema de gestos gastará sempre mais tempo do que o mouse para uma determinada tarefa com o mesmo ID como esperado pois o tempo de experiência de uso para o mouse é muito superior

ao tempo de experiência de uso do sistema de gestos para todos os indivíduos utilizados no experimento. Além disso, o universo de indivíduos utilizados no experimento é restrito a um número muito pequeno e tendencioso pois todos pertencem a um mesmo grupo, no caso, de estudantes. Deve-se levar em conta também que a amplitude de movimento de mãos e braços na utilização do sistema de reconhecimento de gestos é maior do que na utilização do mouse.

Verificamos valores que indicam correlação linear entre ID e T tanto na modalidade mouse quanto na modalidade de gestos. A correlação é ainda maior na modalidade de gestos. Para os valores de A e T, também verificamos que existe a correlação linear. Já para os valores de W e T podemos verificar que pode não existir uma dependência linear entre o W e T, o que significa que o ruído do sistema de gestos não foi traduzido em maior tempo de ativação de botões.

Figura 5.8: Comparação da reta $T = a + bID$ para as duas modalidades

Modalidade	a	b	IP
Mouse	240,43	215,31	0,082
Gestos	121,83	649,61	0,042

Tabela 5.2: Coeficientes a e b da Equação 5.2 obtidos de forma empírica

Avaliação subjetiva

Para esta tarefa, o usuário é instruído a utilizar o aplicativo de manipulação de objetos 3D, como descrito na seção 5.2.1, utilizando para isso o sistema de reconhecimento de gestos. Durante a realização desta fase de familiarização, o usuário tinha liberdade total de movimentação, ou seja, não havia nenhum roteiro específico a ser seguido.

Este experimento não foi cronometrado e consistia de uma avaliação subjetiva do sistema, feita por cada usuário. Os usuários foram encorajados a fazer comentários durante e após a utilização do sistema. Além disso, durante e após a realização das tarefas, comentários e avaliações subjetivas dos usuários foram coletadas.

5.2.2 Resultados

Para esta tarefa de critérios subjetivos, pudemos observar como os usuários se saíam utilizando o sistema de forma arbitrária. Comentários foram colhidos sobre o sistema como um todo, incluindo aspectos de utilização, conforto e usabilidade.

Ao utilizar o sistema, 50% dos usuários indicaram como fator positivo a robustez do sistema a movimentos bruscos e rápidos. Eles podiam movimentar as mãos a qualquer velocidade, sem perda do sistema de rastreamento.

Ao observar a movimentação das mãos dos usuários neste experimento, pudemos observar que todos eles, quando queriam movimentar um objeto 3D na tela para frente e para trás,

Figura 5.9: Aplicativo desenvolvido para os experimentos no ambiente Quiosque

moviam suas mãos também para frente e para trás. Tal recurso não foi implementado visto que o sistema reconhece apenas movimentos no plano 2D. Mas fica claro que a naturalidade de trazer e empurrar um objeto é algo a ser incorporado por este sistema, com o aumento do grau de reconhecimento de 2D para 3D.

Metade dos usuários que utilizaram o sistema reclamaram de cansaço na utilização. Realmente, a movimentação das mãos envolve a movimentação dos braços como um todo, e ao utilizar o sistema por mais de 10 minutos, os usuários tendem a ficar cansados. Pudemos reparar que usuários mais experientes tendem a relaxar o braço, deixando-o na posição Sul quando não o estão usando para controle do sistema, o que reduz a fadiga mencionada.

Apenas um dos seis usuários reclamou da restrição do sistema em apenas aceitar usuários que estejam utilizando roupas que cobrem totalmente os braços.

Um dos usuários era canhoto. Como a mão direita controla os movimentos do cursor, a princípio a hipótese de usuários canhotos apresentarem menor destreza na movimentação do cursor não foi confirmada, pois este usuário não relatou desconforto nem falta de destreza. Contudo, experiências com um maior número de usuário precisam ser realizadas.

A utilização de usuários com experiência no uso de mouse e já familiarizados com realidade virtual e computação gráfica pode ter levado a uma avaliação tendenciosa do sistema. Entretanto, acreditamos que as características de usuários experientes em RV e com o uso de dispositivos como rastreados magnéticos não influenciou na análise de resultados.

5.2.3 Ambiente CAVE

A CAVERNA Digital conta com um sistema de cluster [62] para alimentar as projeções. Utilizando-se este sistema, é possível exibir imagens nas telas utilizando-se máquinas diferentes. Para a sincronização das informações de ponto de vista e consistência de dados foi utilizada a biblioteca chamada Glass [28].

Desta forma é possível escolher o que é projetado em cada uma das telas. Atualmente a tela do chão é utilizada para a projeção de uma imagem branca, que ajuda a eliminar ruído na segmentação porque provê melhor condições de captura de imagem para a câmera posicionada na CAVERNA Digital. Prevê-se o uso da tela do chão também como uma alternativa para informar o usuário o status do sistema de reconhecimento de gestos. Estas informações serviriam para indicar ao usuário dados como posição das mãos, eventuais falhas na identificação dos objetos e também indicação para voltar ao estado inicial caso o sistema de rastreamento se perca.

O layout da CAVERNA Digital pode ser visto na figura 5.10b. Uma câmera é posicionada

acima da tela principal no ambiente CAVE (veja figura 5.10a) e fornece imagens como as mostradas na figura 5.11.

Figura 5.10: Ambiente CAVE. a) Câmera posicionada acima da tela principal e b) Layout Virtual do Ambiente.

Descrição do experimento

A aplicação de CAVE desenvolvida para este experimento (veja figura 5.12) é uma exploração virtual de um modelo 3D da catedral de Sibenik, na Croácia. A interface desenvolvida para controle desta aplicação é a seguinte:

- A mão direita (R) controla o ponto de vista da câmera;
- A mão esquerda (L) controla a translação da câmera. Quando (L) se posiciona na posição Norte, a câmera se movimenta na direção do ponto de vista determinado por (R). Quando (L) se posiciona na posição Sul, a câmera se movimenta ao contrário da direção do ponto de vista.

Figura 5.11: Amostra de seqüência de imagens colhidas pela câmera posicionada no ambiente CAVE

Resultados

A utilização de gestos no ambiente CAVE deixa o usuário mais a vontade para visualização do ambiente 3D. Mas, com a falta de iluminação adequada para a câmera e também devido a grande variação de iluminação intrínsecas a este ambiente, o sistema de gestos fica menos robusto do que nas aplicações no ambiente Quiosque. Esta variação ocasiona tremulação do ponto de vista da câmera, prejudicando a navegação no mundo 3D. Mesmo com a tela do chão sendo usada para compensar a iluminação, projetando-se uma imagem branca nela, os

Figura 5.12: Aplicação utilizada para teste na CAVERNA Digital

resultados não são tão bons quanto no ambiente Quiosque.

Uma observação feita é a naturalidade dos usuários ao utilizarem o sistema. Isso faz com que, por exemplo, ao querer movimentar o ponto de vista da câmera para a frente, eles movimentem as mãos para frente, ao invés de movimentar para a posição Norte, como observado também nos experimentos no ambiente Quiosque. Isso indica que a utilização de rastreamento 3D, ou seja, a utilização de múltiplas câmeras é um tema a ser estudado e incorporado ao sistema.

5.3 Discussão e Resultados

Discutiremos nesta seção os resultados encontrados nos ambientes Quiosque e CAVE. Discutiremos também separadamente os processos que compõem o sistema como um todo, como segmentação, rastreamento e interface baseada em gestos.

5.3.1 Quiosque

O ambiente Quiosque, devido a suas melhores condições de iluminação e possibilidade de manter a iluminação controlada, foi o que apresentou melhor robustez ao sistema de segmentação, rastreamento e reconhecimento de gestos. O fundo não é modificado nem controlado, aumentando a portabilidade e usabilidade do sistema.

Com base nos testes realizados, o sistema suporta a presença de outras pessoas na área de captura da câmera, após o sistema ter sido iniciado. A presença de pessoas passando atrás do usuário utilizando o sistema interferem apenas caso a pessoa esteja tão próxima da câmera que as componentes conexas de suas mãos e face sejam tão grandes quanto as do usuário utilizando o sistema. A robustez a estas interferências se devem ao uso do grafo de consistência. Veja na figura 5.13 um exemplo onde um indivíduo se posiciona atrás do usuário utilizando o sistema.

Figura 5.13: Usuário utilizando o sistema (face e mãos rastreados) e pessoa atrás não interferindo no rastreamento do sistema.

Podemos ressaltar também que a calibração do sistema é simples, rápida e não requer recalibrações freqüentes.

5.3.2 CAVE

O ambiente CAVE oferece grandes desafios para aplicações de visão computacional. A iluminação é variável e não controlada. Uma possibilidade de melhora neste aspecto é a utilização de outros dispositivos para captura de imagens, que disponham de mais recursos para trabalhar sob estas condições.

Com isso, o ruído na segmentação é propagado por todas as etapas do sistema, causando menor robustez no reconhecimento de gestos neste ambiente.

Apesar disto, a liberdade proporcionada pela navegação utilizando-se apenas as mãos foi interessante. Os sistemas atuais disponíveis na CAVERNA Digital para interação não dispunham de tal característica. Uma das vantagens apontadas pelos usuário foi a não preocupação com os fios e cabos de outros sistemas bem como a melhor visualização com a eliminação destes cabos, que muitas vezes ficavam na frente das projeções.

O treinamento para a identificação de faces utilizando PCA não foi realizada para a CAVE. O PCA, a princípio, funciona apenas para faces que estejam de frente para a câmera, sem nenhuma rotação ou inclinação. Atualmente, membros do LaTIn estão estudando técnicas para superar estas dificuldades.

A definição da interface e os ajustes necessários nos parâmetros dos aplicativos exigiram um esforço significativo. Mesmo assim, consideramos que a robustez do sistema no ambiente CAVE precisa ser melhorada. Em particular, demoramos muito no ajuste para que a interface reagisse bem aos ruídos ocasionados pela segmentação.

A calibração do sistema na CAVERNA Digital também é simples, rápida e não requer recalibrações freqüentes.

5.3.3 Segmentação

A segmentação baseada em cor de pele funciona em ambientes com boa iluminação, mesmo com pequenas variações, como o Quiosque. Já no ambiente CAVE, a baixa iluminação causa problemas para a câmera, e suas imagens se tornam muito ruidosas. O ajuste de cor personalizado para cada usuário, feito através da identificação da região da face através do PCA, ajuda a refinar o modelo de cor, tornando o sistema mais robusto.

Há vários métodos para melhorar a qualidade de imagem, como aumentar a resolução da câmera, aumentar sua capacidade de trabalhar em baixas condições de iluminação (poucos lumens) e pré-processamento das imagens, como operações morfológicas e filtros. Entretanto, o custo computacional desses métodos pode inviabilizar o desempenho necessário para interação.

Outro ponto a ser melhorado é a capacidade do sistema de rastrear todo o corpo (braço, tronco, pernas, etc) além das mãos e do rosto, permitindo assim que o usuário utilize qualquer vestimenta.

Pessoas com cabelos mais claros tem seus cabelos detectados como cor de pele. Veja exemplo na figura 5.14.

Correção Dinâmica de Cor de Pele

A correção automática do espaço de cor de pele para busca apresentou resultados satisfatórios. O algoritmo para cálculo de atualização deste espaço é chamado a cada 30 quadros capturados, ou aproximadamente 1 segundo. A figura 5.15 mostra os resultados obtidos. Nota-se o aumento da região de cor de pele correspondente a face sendo segmentada dos quadros 1 até o quadro 4. Estes quadros foram gravados em intervalos de 1 segundo.

A aparição de ruídos com o ajuste dinâmico do espaço de cor de pele é um problema que aparece com frequência. Mas estes ruídos são excluídos durante a busca por componentes conexas pois a área deles é inferior ao limiar imposto pelo sistema. Estes ruídos aparecem

Figura 5.14: Pessoa com cabelos loiros: segmentação de cor de pele detecta cabelos loiros como cor de pele

quando o fundo da imagem, próximo a região do rosto, tem uma cor semelhante a cor de pele.

5.3.4 Rastreamento

A utilização de informações sobre as coordenadas de seu centróide (x, y) e também o quadrilátero que o contém confere robustez e desempenho pois a quantidade de informações é reduzida. Com estas informações é possível detectar a colisão de objetos, calcular o espaço de procura para o próximo passo e calcular a posição do cursor na tela. A detecção de colisão ficaria mais complexa se fosse utilizado todo o contorno dos objetos. A baixa resolução de imagem utilizada diminui a precisão no rastreamento, causando um tremor no cursor. Felizmente este não é um fator limitante do sistema.

O filtro de Kalman contribui para a robustez e velocidade do sistema de rastreamento por sua simplicidade e velocidade, se comparado com filtros mais complexos, não lineares e que utilizam mais de uma variável. Para cada uma das coordenadas, mão esquerda, mão direita e face foram utilizados filtros distintos pois seus movimentos são independentes uns dos outros.

Figura 5.15: Quadros tirados com intervalo de 1 segundo entre si. Os quadros mostram a variação na segmentação ocasionada pelo ajuste dinâmico do espaço de busca.

Observa-se que o filtro não prevê corretamente movimentos muito bruscos de movimentação dos objetos (observe a figura 5.16 entre os quadros 100 e 120 e também entre 160 e 180).

O conjunto de treinamento utilizado para a identificação de faces pelo algoritmo PCA apresentou menos problemas para o ambiente de Quiosque porque a face do usuário ficava em um ângulo mais próximo ao ângulo das faces do conjunto de treinamento. Devido à particularidades do PCA, em alguns momentos, quando o usuário não está totalmente de frente para a câmera, o PCA não consegue identificar a face corretamente. No ambiente CAVE a situação se agrava porque a câmera não consegue capturar a face totalmente de frente pois está posicionada muito acima do usuário.

Heurísticas de manipulação e apontamento podem ser uma alternativa para melhorar a manipulação de objetos. Por exemplo, quando o usuário clica sobre uma região vazia mas próxima a um objeto, o sistema poderia automaticamente selecionar esse objeto mais próximo.

Em um sistema de apontamento indireto como este, pode ser difícil obter uma uniformidade na movimentação dentro da área de trabalho. A calibração de uma área específica, definida

Figura 5.16: Valor medido para a ordenada X do centróide da mão esquerda e seu respectivo valor previsto pelo filtro de Kalman.

pelo próprio usuário no momento da calibração do sistema de rastreamento ajuda a reduzir variações e a melhorar a acurácia.

Apesar de robusto, o sistema de rastreamento ainda sofre com a interferência de outras pessoas posicionando-se na área de alcance da câmera. Se alguma pessoa passar entre a câmera e o usuário, o sistema pode se perder. Basicamente ele se perde quando alguma área de cor de pele se aproxima ou se aglutina a alguma área de cor de pele já sendo rastreada. Caso algum objeto sem cor de pele fique entre a câmera e o usuário, dificilmente o sistema se perderá pois o sistema é robusto à oclusão.

5.3.5 PCA

A utilização de PCA para identificação de qual objeto é a face do usuário ainda apresenta problemas e restrições com relação a diferenças de iluminação, orientação e escala da face

(rotação e tamanho com relação as imagens utilizadas para treinamento). Embora o PCA não seja muito sensível a rotações, devido a particularidade da interação permitida pelo sistema proposto, dificilmente o objeto face está corretamente alinhado de acordo com as imagens de treinamento.

Figura 5.17: Objetos identificados como mão direita *R*, face *H* e mão esquerda *L*.

Apesar disso, o algoritmo se apresenta robusto quando utilizado no ambiente Quiosque, pois neste caso, o usuário estará posicionado de frente para a câmera. Na figura 5.17 temos os objetos identificados com a seguinte coloração: mão direita - azul; face - vermelho; mão esquerda - amarelo. Respectivamente, a distância calculada entre estes objetos (imagens) e o espaço de faces calculada foi: 0.7893, 0.7689, 0.7939. Percebe-se que neste caso inicial, a face apresenta a menor distância para o espaço de faces e, logo, foi classificada como tal.

O grupo de pesquisa do LaTIn [53] está trabalhando nesta área de PCA para melhorar a robustez à variação de rotação e escala da imagem de face.

5.3.6 Interface baseada em Gestos

A simplicidade de identificação de posições chave para interfaces baseadas em gestos contribuiu para a robustez do sistema como um todo. Ao contrário de sistemas que utilizam cadeias de Markov ocultas (*Hidden Markov Models*), a rapidez deste tipo de alternativa proporciona melhores resultados referentes à habilidade de tratar gestos em tempo real.

Como os gestos das mãos definidos para atingir as posições chave são relativos a posição da face do usuário, o sistema fica robusto a movimentação do usuário em frente a câmera, não exigindo que ele/ela fique estático/a.

Por outro lado, a simplicidade dos gestos reconhecidos não permite que o sistema identifique gestos mais complexos, como por exemplo, um movimento circular de uma das mãos poderia ser identificado como sendo um comando individual.

O grafo de consistência foi essencial na eliminação dos problemas de oclusão na movimentação das mãos do usuário. Devido a natureza de interação, as mãos e face do usuário se sobrepõem com frequência. Gestos como junção de mãos esquerda e direita ou, levar a mão esquerda a face foram rastreados corretamente sem que o sistema se perdesse. Esta robustez atingida foi crucial para a utilização do sistema como um dispositivo de entrada de dados.

Após alguns minutos de uso intermitente, alguns usuários apontaram como problema a fadiga dos braços pelo uso corrente ao utilizar o sistema. Infelizmente isso é algo que pode ocorrer e varia de usuário para usuário. Apesar do sistema utilizar posições de conforto ao usuário para interação, a movimentação contínua para uso de uma interface não usual pode causar fadiga. Isso também pode ser observado em usuários que utilizam mouses e teclados ininterruptamente. Foram testadas regiões de trabalho de menor amplitude para a interação do usuário, mas também sem sucesso no que diz respeito a fadiga.

5.4 Conclusão

A utilização do modelo de Fitts mostra que o sistema apresenta características semelhantes ao da utilização de mouse. Mas o tempo gasto para realização de uma tarefa foi maior que o tempo gasto com o mouse para a realização da mesma tarefa. Creditamos a isso o fato dos usuários utilizados para teste terem muitos anos de experiência com mouse ao contrário do sistema proposto, onde os mesmos tiveram apenas alguns minutos de familiarização.

A recalibração dinâmica de cor de pele se mostra robusta auxiliando na redução dos problemas encontrados com a variação de iluminação. A utilização de um filtro de resposta finita para redução de ruídos na movimentação do cursor permitiu melhorar de forma significativa a tremulação do cursor, melhorando o desempenho do sistema. O uso de análise de componentes principais ainda não está adequado pois a variação de classificação obtida ainda é grande, devido a movimentação constante da face do usuário, prejudicando sua detecção. Apesar disso, os experimentos realizados mostram a robustez do sistema de reconhecimento de gestos.

Capítulo 6

Conclusão e Trabalhos Futuros

Este trabalho apresenta uma técnica de reconhecimento de gestos capaz de ser utilizada em diferentes tipos de ambientes de RV e com algumas restrições em ambientes imersivos como a CAVERNA Digital. Mais especificamente, este estudo concentrou-se em desenvolver uma interface baseada em gestos que é robusta e capaz de ser utilizada em diferentes ambientes, sendo proposta, como principal contribuição um sistema baseado em visão computacional capaz de implementar a interface desenvolvida em tempo real. As técnicas desenvolvidas e apresentadas foram avaliadas de forma subjetiva e também quantitativa. O problema de reconhecer movimentos de gestos de um usuário, sem que o mesmo precise utilizar dispositivos invasivos, para interação em ambientes de RV foi estudado e uma solução apresentada. A extensão deste sistema para a utilização em qualquer tipo de ambiente, para realização de tarefas simples como ligar e desligar um televisor, mostrou-se viável.

As dificuldades inerentes do uso de visão computacional como iluminação e latência grande podem ser superados para a utilização de gestos como uma interface para ambientes de RV. O sistema implementado é capaz de reconhecer gestos com as mãos numa configuração realista e em tempo real. Após uma curva de aprendizado, um usuário experiente pode manipular

objetos e navegar por mundos virtuais com a velocidade e conforto de outros dispositivos populares. A habilidade de manipular diretamente objetos parece ser mais confortável para alguns usuários que utilizar meios indiretos como mouse ou *joystick*. Com um pouco de engenharia, este sistema pode ser utilizado de forma mais corriqueira em interações nestes ambientes. Soluções para melhorar a robustez no ambiente CAVE parecem ser relativamente diretas, como por exemplo, substituição da câmera utilizada para melhorar a qualidade de imagem.

Diversos experimentos foram realizados para todas as etapas do sistema. Para segmentação de cor de pele, a solução mais robusta e com maior desempenho foi a utilização do espaço RGB normalizado e limiarização com valores de treinamento. Este espaço mostrou-se robusto a pequenas variações homogêneas de iluminação. Mas quando utilizado no ambiente CAVE, mostrou-se menos robusto. Percebe-se claramente neste trabalho que a segmentação de cor de pele tem papel fundamental no reconhecimento de gestos. A escolha de câmeras mais adequadas a condições adversas de iluminação deve ser levada em consideração em trabalhos futuros. Mesmo assim, os experimentos foram capazes de mostrar a robustez do sistema bem como o desempenho em tempo real alcançado para o reconhecimento de gestos baseado em visão computacional.

A combinação de Filtros de Kalman com o grafo de consistência possibilitou maior robustez ao sistema de rastreamento. Apesar do sistema ter sido desenvolvido para utilização por apenas um usuário, assumindo que outras pessoas não estariam na área de captura da câmera, a combinação proposta mostrou-se robusta. A tentativa de utilização de Análise de Componentes Principais para identificação da face do usuário não mostrou-se adequada, devido a grande variação de ângulos que a face do usuário pode assumir durante a execução do sistema. O problema é ainda mais grave no ambiente CAVE, onde esta variação é maior devido ao posicionamento da câmera. Estudos no laboratório LaTIn já estão sendo conduzidos

para robustez na utilização de algoritmos de Componentes Principais.

As interfaces desenvolvidas para o sistema de gestos mostraram-se adequadas e os experimentos realizados com diferentes usuários comprovam sua viabilidade. Comparando-se através da Lei de Fitts atividades de apontamento entre mouse e a interface, a mesma apresentou características semelhantes a do mouse.

O sistema de reconhecimento de gestos desenvolvido e também as aplicações criadas para os experimentos e testes foram baseadas nas técnicas apresentadas, objetivando na construção de todo o sistema, que é um interpretador de gestos baseado em visão computacional. A pesquisa para o desenvolvimento deste apresentou respostas para diversas dúvidas que surgiram durante o andamento do trabalho. E inerente a todo trabalho científico, esta mesma pesquisa levantou outras dúvidas e trabalhos futuros a serem investigados. São eles:

- A segmentação de cor de pele depende das condições de iluminação. Infelizmente utilizamos apenas uma câmera para testes neste trabalho. Diferentes câmeras, com capacidade de captura a iluminação mais baixa podem ser mais adequadas.
- Melhorar a robustez à variação de iluminação para o sistema de segmentação de cor de pele.
- Adaptar o algoritmo de análise de componentes principais para identificação de faces em imagens onde a face não está totalmente de frente para a câmera.
- Realizar experimentos com um número maior de pessoas, incluindo medidas definidas pelos padrões de usabilidade em IHC.
- Realizar experimentos em outros ambientes fora do laboratório.
- Rastreamento utilizando mais de uma câmera (3D).

- Reconhecer gestos de mais de uma pessoa para aplicações colaborativas.

Apêndice A

Mapeamento Tela-Imagem

Kjeldsen [41] mostra que o jeito mais fácil de se calcular estes coeficientes é primeiro calcular os coeficientes da matriz inversa de A , A^{-1} .

$$A^{-1} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Para esta matriz, as equações de transformação são:

$$\begin{aligned} x &= \frac{A_{11}x_s + A_{21}y_s + A_{31}}{A_{13}x_s + A_{23}y_s + A_{33}} \\ y &= \frac{A_{12}x_s + A_{22}y_s + A_{32}}{A_{13}x_s + A_{23}y_s + A_{33}} \end{aligned} \tag{A.1}$$

Combinando a Equação A.1 com os pares 4.2 temos:

$$\begin{aligned}
a_{11} &= x_1 - x_0 + A_{13}x_1 \\
a_{21} &= x_3 - x_0 + A_{23}x_3 \\
a_{31} &= x_0 \\
a_{12} &= y_1 - y_0 + A_{13}y_1 \\
a_{22} &= y_3 - y_0 + A_{23}y_3 \\
a_{32} &= y_0 \\
a_{13} &= \begin{vmatrix} \delta x_3 & \delta x_2 \\ \delta y_3 & \delta y_2 \\ \delta x_1 & \delta x_2 \\ \delta y_1 & \delta y_2 \\ \delta x_1 & \delta x_3 \\ \delta y_1 & \delta y_3 \end{vmatrix} \\
a_{23} &= \begin{vmatrix} \delta x_1 & \delta x_2 \\ \delta y_1 & \delta y_2 \end{vmatrix} \\
a_{33} &= 1
\end{aligned} \tag{A.2}$$

onde

$$\delta x_1 = x_1 - x_2$$

$$\delta x_2 = x_3 - x_2$$

$$\delta x_3 = x_0 - x_1 + x_2 - x_3$$

$$\delta y_1 = y_1 - y_2$$

$$\delta y_2 = y_3 - y_2$$

$$\delta y_3 = y_0 - y_1 + y_2 - y_3$$

De acordo com Kjeldsen [41], para uma transformação perspectiva, a matriz adjunta da

matriz de transformação é equivalente a sua inversa. Então A' :

$$A' = \begin{bmatrix} a_{22}a_{33} - a_{23}a_{32} & a_{13}a_{32} - a_{12}a_{33} & a_{12}a_{23} - a_{13}a_{22} \\ a_{23}a_{31} - a_{21}a_{33} & a_{11}a_{33} - a_{13}a_{31} & a_{13}a_{21} - a_{11}a_{23} \\ a_{21}a_{32} - a_{22}a_{31} & a_{12}a_{31} - a_{11}a_{32} & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix}$$

é o mapeamento das coordenadas da imagem para a tela unitária. Multiplicando-se A' por T , uma matriz de transformação que mapeia as coordenadas unitárias de tela para as coordenadas corretas de tela, temos que

$$A = A'T$$

é a matriz de transformação das coordenadas de imagem para as coordenadas de tela.

Apêndice B

Questionário

1. Idade:

2. Sexo:

Masculino

Feminino

3. Destro ou canhoto?

Canhoto

Destro

4. Você se considera um usuário avançado ou iniciante nas áreas de Realidade Virtual e Computação Gráfica?

Iniciante

Avançado

5. Quais interfaces abaixo você conhece e já utilizou?

Track-of-Birds

Spaceball

Nenhuma

Referências

- [1] Subutai Ahmad. A usable real-time 3d hand tracker. *28th Proceedings Asilomar Conference on Signals, Systems and Computers*, pages 1257–1261, 1995.
- [2] Bryce Bayer. Color imaging array, 1976. US Patent 3,971,065.
- [3] Michael Black and Allan Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV (1)*, pages 329–342, 1996.
- [4] Doug Bowman. Interaction techniques for common tasks in immersive virtual environments. Technical report, Georgia Institute of Technology, June 1999.
- [5] Doug Bowman, Ernst Kruijff, Joseph La Viola, and Ivan Poupyrev. An introduction to 3-d user interface design. *Presence: Volume 10 - Number 1*, pages 96–108, 2001.
- [6] Doug Bowman, Chadwick Wingrave, Joshua Campbell, and Vinh Ly. Using pinch gloves for both natural and abstract interaction techniques in virtual environments. *Proceedings of HCI International*, pages 629–633, 2001.
- [7] James Cai, Arthur Goshtasby, and Cien Yu. Detecting human faces in color images. In *Image and Vision Computing - Volume 18*. Wright State University - University of Illinois, 1999.

- [8] Douglas Chai and King Ngan. Automatic face location for videophone images. *IEEE - Automatic Face Location for Videophone Images*, pages 137–140, November 1996.
- [9] Henry Chang and Ulises Robles. Face detection. Technical report, University of Stanford, 2001.
- [10] German Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. *Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition - CVPR*, 2003.
- [11] Mark Coomans. Towards a taxonomy of virtual reality user interfaces. *International Conference on Information Visualisation - ICIV*, 1997.
- [12] Ascension Technology Corporation. URL: [http:// www.ascension-tech.com](http://www.ascension-tech.com), Acesso em Junho de 2004.
- [13] Elumens Corporation. URL: [http:// www.elumens.com](http://www.elumens.com), Acesso em Novembro de 2004.
- [14] Carolina Cruz-Neira, Daniel Sandin, and Tom DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the cave. In *SIGGRAPH*. ACM SIGGRAPH, Julho 1993.
- [15] Lawrence Cutler, Bernd Frolich, and Pat Hanrahan. Two-handed direct manipulation on the responsive workbench. In *Symposium on Interactive 3D Graphics*, pages 107–114, 191, 1997.
- [16] Alan Dix, Janet Finlay, Gregory Abowd, and Russel Beale. *Human Computer Interaction*, volume 2nd edition. Prentice Hall, 1998.

- [17] Klaus Dorfmueller. An optical tracking system for VRAR-applications. In *Eurographics Workshop on Virtual Environments - EGVE99*, June 1999.
- [18] Sidney Fels and Geoffrey Hinton. Glove-Talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on Neural Networks*, 4(1):2–8, January 1993.
- [19] Paul Fieguth. Color-based tracking of heads and other mobile objects at video frame rates. *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition - CVPR*, 1997.
- [20] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology: Human Perception and Performance*, 47:381–391, 1954.
- [21] Margaret Fleck, David Forsyth, and Chris Bregler. Finding naked people. In *ECCV (2)*, pages 593–602, 1996.
- [22] James D. Foley, Andries van Dam, Steven Feiner, and John Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, 1997.
- [23] Terrence Fong, Francois Conti, Sebastien Grange, and Charles Baur. Novel interfaces for remote driving: gesture, haptic and pda. *SPIE Telemanipulator and Telepresence Technologies VII*, November 2000.
- [24] Joseph Gabbard and Deborah Hix. *A taxonomy of Usability characteristics of virtual environments*. PhD thesis, Office of Naval Research, Virginia Polytechnic Institute, 1997.

- [25] Bruno Gnecco, Marcelo Zuffo, Paulo Bressan, and Roseli Lopes. Dixelib: A real time synchronization library for multiprojection virtual reality distributed environments. In *SVR 2001*. SBC, October 2001.
- [26] Vicente Goes, Liliane Machado, Marcio Cabral, and Marcelo Zuffo. Interactive stereoscopic full-color direct volume visualization for virtual reality in medicine. In *MMVR 2001 Proceedings*. Aligned Management Associates, Inc., January 2001.
- [27] Yves Guiard. Asymmetric division of labor in human skilled bimanual action: the kinematic chain as a model. In *The Journal of Motor Behavior*, pages 486–517, 1987.
- [28] Marcelo Guimarães, Bruno Gnecco, Marcio Cabral, and Marcelo Zuffo. Synchronization and data sharing library for pc clusters. In *IEEE VR - VRCluster'03*. IEEE, March 2003.
- [29] Greg Hager and Peter Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence - PAMI*, 20(10):1025–1039, 1998.
- [30] Pengyu Hong, Matthew Turk, and Thomas Huang. Gesture modeling and recognition using finite state machines. *Proceedings of the Fourth IEEE International Conference on Gesture Recognition*, 2000.
- [31] Andrew Senior Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil Jain. Face detection in color images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1046–1049, May 2002.
- [32] Giancarlo Iannizzotto. Real-time object tracking with movels and affine transformations. *The IEEE International Conference on Image Processing - ICIP*, 1:316–318, 2000.

- [33] Point Grey Research Inc. Dragon fly firewire camera specifications - <http://www.ptgrey.com/products/dragonfly/specifications.html>, 2003.
- [34] Yang Jie and Waibel Alex. A real-time face tracker. In *Proceedings of WACV'96*, pages 142–147, 1996.
- [35] Yang Jie, Lu Weier, and Waibel Alex. Skin-color modeling and adaptation. In *Computer Vision - ACCV'98*, pages 687–694. Universität Karlsruhe; Institut für Logik, Komplexität und Deduktionssysteme, R. Chin. Vol. 2., 1997.
- [36] Rudolph Kalman. A new approach to linear filtering and prediction problems. *Transaction of the ASME Journal of Basic Engineering*, pages 35–45, march 1960.
- [37] K Kaurs. *Designing Virtual Environments for Usability*. PhD thesis, Centre for HCI Design, London, 1998.
- [38] Roland Kehl and Luc Van Gool. Real-time pointing gesture recognition for an immersive environment. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.
- [39] Adam Kendom. *Gesture and Speech: How They Interact - Nonverbal Interaction*. Beverly Hills: Sage Publications, 1983.
- [40] Adam Kendom. *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press, 1990.
- [41] Frederick Kjeldsen. Visual interpretation of hand gestures as a practical interface modality. *Phd Thesis - Columbia University*, 1997.

- [42] Christian Leubner, Christian Brockmann, and Heinrich Müller. Computer-vision-based human-computer interaction with a back projection wall using arm gestures. *Euromicro Conference*, 2001.
- [43] Victor Lombardi. Virtual reality vs. immersive environments: Choosing the right computer interface for the future, Acesso 10 de Fevereiro de 2005.
- [44] Vicon Motion Systems Ltd. URL: [http:// www.vicon.com](http://www.vicon.com), Acesso em Janeiro de 2005.
- [45] Paulo Maciel and Peter Shirley. Visual navigation of large environments using textured clusters. In *Symposium on Interactive 3D Graphics*, pages 95–102, 211, 1995.
- [46] Scott MacKenzie, Abigail Sellen, and William Buxton. A comparison of input devices in element pointing and dragging tasks. *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 161–166, 1991.
- [47] Loretta Malandro, Larry Barker, and Deborah Ann Barker. *Nonverbal Communication*. Addison-Wesley, 1989.
- [48] Birgitta Martinkauppi. *Face colour under varying illumination - analysis and applications*. PhD thesis, Acta Universitatis Ouluensis, 2002.
- [49] Birgitta Martinkauppi, Mika Laaksonen, and Maricor Soriano. Behavior of skin color under varying illumination seen by different cameras at different color spaces. In *Machine Vision Applications in Industrial Inspection IX*, pages 102–112. Proceedings of SPIE Vol. 4301, Martin Hunt, 2001.
- [50] Peter Maybeck. *Stochastic Models, Estimation, and Control*. Academic Press, Inc., 1979.

- [51] Ivana Miki, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision - IJCV*, 53(3):199–223, 2003.
- [52] Thomas Moeslund, Moritz Störring, and Erik Granum. Vision-based user interface for interacting with a virtual environment. In *Proceedings of DANKOMB - 4th International Workshop on Gesture and Sign Language based Human Computer Interaction*. Aalborg University, 2000.
- [53] Carlos Hitoshi Morimoto. Laboratório de tecnologias de interação - URL: [http:// latin.ime.usp.br](http://latin.ime.usp.br), Acesso 10 de Janeiro de 2005.
- [54] Jakob Nielsen. *Usability Engineering*. Cambridge, MA: AP Professional, 1993.
- [55] Jenny Preece, Yvonne Rogers, and Helen Sharp. *Interaction design: beyond human-computer interaction*. John Wiley & Sons Inc., 2002.
- [56] James Rehg, Maria Loughlin, and Keith Waters. Vision for a smart kiosk. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 690–696, 1997.
- [57] Miguel Ribo, Axel Pinz, and Anton Fuhrmann. A new optical tracking system for virtual and augmented reality applications. In *IEEE Instrumentation and Measurement Technology Conference - IMTC*. IEEE, 2001.
- [58] Ben Schneiderman. *Designing the user interface*. Addison-Wesley, 1997.
- [59] Atid Shamaie, Wu Hai, and Alistair Sutherland. Hand-gesture recognition for hci, 2003.
- [60] Atid Shamaie and Alistair Sutherland. Graph-based matching of occluded hand gestures. In *IEEE Applied Imagery Pattern Recognition Workshop - AIPR*. IEEE, October 2001.

- [61] Christian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3d body tracking. *Proceedings of the 2001 Conference on Computer Vision and Pattern Recognition - CVPR*, 2001.
- [62] Luciano Soares, Marcio Cabral, Paulo Bressan, Roseli Lopes, and Marcelo Zuffo. Powering multiprojection immersive environments with clusters of commodity computers. In *1st Ibero-American Symposium in Computer Graphics - Proceedings SIACG 2002*, July 2002.
- [63] IEEE Standards. Ieee 1394 - standard for a high performance serial bus description.
- [64] IEEE Standards. Ieee 1394b - standard 1394b - <http://grouper.ieee.org/groups/1394/b/>, 2003.
- [65] Moritz Störing, Hans Andersen, and Erik Granum. Skin colour detection under changing lighting conditions. *7th Symposium on Intelligent Robotics Systems*, pages 187–195, July 2001.
- [66] Ivan Sutherland. The ultimate display. In *Proceedings of IFIPS Congress*, pages 506–508. IFIPS, May 1965.
- [67] Ivan Sutherland. A head mounted three dimensional display. In *Proceedings of the AFIPS Fall Joint Computer Conferences*, pages 757–764. Thompson Books, 1968.
- [68] Sabine Süsstrunk. Color demosaicing, 2003.
- [69] T.D.Grove, K.D.Baker, and T. N. TAN. Color based object tracking. *14th International Conference on Pattern Recognition - ICPR*, 1998.
- [70] Image Guided Technologies. URL: [http:// www.imageguided.com](http://www.imageguided.com), Acesso em Setembro de 2004.

- [71] Jean-Christophe Terrillon and Shigeru Akamatsu. Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. *IEEE International Conference on Face and Gesture Recognition*, pages 54–61, 2000.
- [72] Matthew Turk and Alan Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- [73] Matthew Turk and Alan Pentland. Face recognition using eigenfaces. *Proceedings of the 1991 Conference on Computer Vision and Pattern Recognition - CVPR*, 1991.
- [74] Jonas Gomes; Luiz Velho. *Image Processing for Computer Graphics*. Springer, 1997.
- [75] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, University of North Carolina - Chappel Hill, 1995.
- [76] Greg Welch, Gary Bishop, Leandra Vicci, Stephen Brumback, and Kurtis Keller. High-performance wide-area optical tracking: The hiball tracking system. In *Presence: Teleoperators and Virtual Environments*, 10(1), 2001.
- [77] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfindex: Real-time tracking of the human body. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 780–785. IEEE, July 1997.
- [78] Christine Youngblut, Rob Johnson, Sarah Nash, Ruth Wienclaw, and Craig Will. Review of virtual environment interface technology - unclassified. Technical report, INSTITUTE FOR DEFENSE ANALYSES - IDA, 1996.

- [79] Hanning Zhou and Thomas Huang. A bayesian framework for real-time 3d hand tracking in highly cluttered background. In *10th International Conference on Human-Computer Interaction*, June 2003.
- [80] Thomas Zimmerman, Jaron Lanier, Chuck Blanchard, Steve Bryson, and Young Harvill. A hand gesture interface device. *ACM SIGCHI Bulletin , Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface*, 17, 1986.
- [81] João Zuffo and Marcelo Zuffo. Caverna digital - sistema de multiprojecao estereoscopico baseado em aglomerados de pcs para aplicacoes imersivas em realidade virtual. In *4th SBC Simposio de Realidade Virtual*, pages 139–147. Sociedade Brasileira de Computacao, Outubro 2001.