Um sistema de interação baseado em gestos manuais tridimensionais para ambientes virtuais

Silvia E. Ghirotti e Carlos H. Morimoto
Departamento de Ciência da Computação – IME/USP
Rua do Matão 1010, Butantã, CEP 05508-090
{silviaeg, hitoshi} @ ime.usp.br

ABSTRACT
With the steady price reduction of 3D visualization devices such as Head Mounted Displays (HMDs) and more recently 3D TVs, we can now foresee the dissemination of applications that was only possible within very expensive virtual reality environments. However, interaction within virtual 3D environments requires more natural modes than those provided by the ubiquitous mouse and keyboard. In this paper we introduce a novel low cost 3D hand gesture based interaction system. We have developed a real-time stereo computer vision system and a hybrid interface that combines natural and symbolic gestures for navigation and manipulation of 3D objects in virtual environments. Results from a pilot study reveals that the hybrid interface presents great power and flexibility without significant increase in the complexity of the user interaction.

RESUMO
A constante redução dos custos de dispositivos de visualização tridimensionais (3D) como Head Mounted Displays (HMDs) e mais recentemente TV's 3D, permite a popularização de aplicativos antes possíveis apenas em ambientes de realidade virtual de custo muito elevado. No entanto, a interação com ambientes virtuais 3D exige modos de interação mais naturais que o mouse e o teclado. Neste artigo nós descrevemos um sistema de interação baseado em gestos manuais tridimensionais de baixo custo, baseado em um sistema de visão computacional estéreo de tempo real, e propomos uma interface híbrida, que combina gestos naturais e simbólicos, para a navegação e manipulação de objetos no ambiente virtual. Resultados de um estudo piloto indicam que a interface híbrida combina de forma eficaz a liberdade de movimentos dos gestos naturais à praticidade dos gestos simbólicos, e sem exigir um longo período de treinamento e adaptação do usuário.

Keywords
Gesture interface, 3D gesture recognition, interaction in 3D virtual environments.

IHC 2010 – IX Simpósio de Fatores Humanos em Sistemas Computacionais. October 5-8, 2010, Belo Horizonte, MG, Brazil.
 ISSN 2178-7697

1. INTRODUÇÃO

A comunicação humana é uma atividade complexa que vai muito além das palavras. Uma das principais ferramentas de suporte à comunicação são os gestos. A família de movimentos que chamamos de gestos é bastante ampla, incluindo desde expressões faciais, passando por movimentos realizados com as mãos, até atividades realizadas com o corpo inteiro. Neste trabalho, abordamos apenas o estudo dos gestos realizados com as mãos (gestos manuais). Bhuiyan e Picking [5] fornecem um histórico recente da utilização de gestos para interfaces com computadores. Outras formas de interação 3D são descritas por Bowman e outros em [4].

Gestos associados à comunicação são denominados gestos semióticos, podendo ser classificados de acordo com a sua dependência com o discurso falado. Interfaces computacionais puramente baseadas em gestos utilizam, principalmente, uma subcategoria de gestos semióticos, denominados simbólicos, como vocabulário de interação. Uma outra subcategoria bastante utilizada são os gestos denominados naturais, que descartam a necessidade de aprendizado e treinamento, mas possuem uma área de aplicação bastante restrita.

Computacionalmente, a estrutura de uma interface baseada em gestos pode ser dividida em três partes: segmentação, reconhecimento e interpretação. A segmentação realiza a detecção da pessoa na cena e a transformação da sua pose num dado momento ou com o passar do tempo. O reconhecimento interpreta as informações obtidas na etapa de segmentação e identifica o gesto sendo realizado. A interpretação contextualiza o gesto ou uma sequência de gestos e extrai desta uma atividade de interação.

Neste trabalho apresentamos um sistema de interação baseado em gestos manuais tridimensionais (3D) para ambientes virtuais. Uma primeira contribuição do trabalho foi o desenvolvimento de um sistema de visão computacional estéreo para o reconhecimento de gestos 3D. Uma segunda e maior contribuição do trabalho, com relação a área de IHC, foi a construção de uma interface baseadas em gestos 3D híbrida, que combina gestos naturais e simbólicos para interação em ambientes virtuais. Uma terceira contribuição foi a elaboração e condução de um estudo piloto que, embora apresente resultados ainda bastante preliminares, ilustra o bom potencial da interface.

O restante do artigo está organizado da seguinte maneira. A próxima seção fornece uma breve descrição de trabalhos

correlatos, com enfoque na definição de conceitos e de uma taxonomia de gestos para interação. Em seguida, definimos os modos de interação utilizados no sistema, utilizando como exemplo um jogo 3D tipo "quebra-cabeça". O sistema de reconhecimento de gestos, utilizando um sistema de visão estéreo é descrito em seguida. Um estudo piloto foi realizado para avaliar o sistema e as formas de interação sugeridas neste trabalho. Os resultados do estudo piloto são apresentados, juntamente com o protocolo experimental e discussões, seguidos dos comentários finais.

2. INTERAÇÃO BASEADA EM GESTOS 3D

Muitos trabalhos recentes exploram interfaces baseadas em gestos bidimensionais, por exemplo, para dispositivos sensíveis a toque, ou gestos de mouse [5]. Neste trabalho, faremos uso da seguinte definição proposta por Mitra e Acharya [14]:

"*gestos são movimentos corporais expressivos e significativos realizados através da movimentação dos dedos, mãos, braços, cabeça, face ou corpo com o objetivo de: 1) transmitir uma informação ou 2) interagir com o ambiente*".

Ou seja, pegar um objeto, correr ou indicar uma direção são atividades consideradas gestos, pois seus modos de execução possuem um papel importante na realização da atividade. Digitar um texto, no entanto, não corresponde a um gesto, pois o sinal referente a um caractere será o mesmo não importando como a tecla seja pressionada.

Os gestos utilizados variam de acordo com aspectos contextuais e culturais segundo Kita [12] e, ainda assim, estão intimamente ligados à comunicação. Por exemplo, pessoas falando ao telefone gesticulam normalmente, mesmo que seu interlocutor não seja capaz de os ver.

Gestos podem possuir significados isolados -- acenar, aplaudir e apontar numa direção -- ou envolvendo objetos externos -- chutar uma bola, pegar e mover um objeto. Podemos, então, classificar os gestos quanto à sua funcionalidade. Cadoz [6] propõe uma classificação em três grupos:

- **semióticos**: utilizados para comunicar uma informação significativa;
- **ergóticos**: utilizados para manipular o mundo físico e criar artefatos;
- **epistêmicos**: utilizados para aprender a partir do meio através da exploração tátil ou háptica.

Neste trabalho, estamos particularmente interessados em como os gestos podem ser utilizados na comunicação com sistemas computacionais, por isso, daremos um maior enfoque nos gestos semióticos que não utilizam dispositivos físicos em contato com o usuário para interação (como luvas e bastões com acelerômetros).

Rimé e Schiaratura [16] propõem a seguinte sub-classificação dos gestos semióticos quanto à sua funcionalidade:

- **simbólicos**: gestos cujo significado é único dentro de uma mesma cultura. Como, por exemplo,o gesto de aprovação feito ao se exibir a mão fechada apenas com o polegar voltado para cima. Linguagens de sinais também se enquadram nesta categoria;
- **deícticos**: gestos mais comumente utilizados em Interação Humano Computador (IHC), pois são aqueles utilizados para apontar ou direcionar a atenção a um determinado evento ou objeto.
- **icônicos**: estes são os gestos utilizados para transmitir informações quanto ao tamanho, forma ou orientação de um objeto em questão. Quando um pescador diz: "Eu pesquei um bagre **deste** tamanho", ao esticar seus braços lateralmente o máximo possível, ele está realizando um gesto semiótico icônico.
- **pantomímicos**: estes são os gestos realizados ao utilizarmos um instrumento ou objeto "invisível", como num jogo de mímica.

A execução de um gesto possui um alto grau de liberdade, o que torna a comunicação através de gestos bastante rica e complexa. Por isso, sistemas de reconhecimento de gestos contam com uma grande variedade de dispositivos utilizados na identificação e rastreamento das partes do corpo relevantes à comunicação. A escolha do dispositivo interfere diretamente na complexidade e qualidade dos gestos analisados. Por exemplo, gestos realizados com dispositivos de ponto único (*single point devices*), como o mouse, limitam o vocabulário do usuário a um conjunto de símbolos planos, compostos por um ou mais traços. Enquanto que dispositivos de rastreamento 3D permitem que o usuário realize gestos mais amplos e naturais expandindo consideravelmente seu vocabulário.

Apesar de possuir uma taxonomia bastante rica, gestos ainda são pouco utilizados como interfaces de sistemas computacionais. As interfaces de gestos mais avançadas utilizam apenas gestos simbólicos ou deícticos. No entanto, podemos apontar duas principais razões para a utilização de gestos como interface de interação:

- Pessoas utilizam normalmente um grande vocabulário de gestos no seu dia-a-dia e aprendem novos gestos fácil e rapidamente, simplesmente observando sua realização por outras pessoas;
- Interfaces baseadas em gestos permitem a utilização natural de frases gestuais, que segmentam o diálogo em trechos com significados simples e fáceis de serem aprendidos e interpretados por sistemas computacionais. Por exemplo, a ação de mover um objeto pode ser segmentada em trechos como segurar o objeto, transladá-lo e soltá-lo.

3. DEFINIÇÃO DOS MODOS DE INTERAÇÃO

Grande parte dos trabalhos apresentados na literatura que propõem soluções para o problema de navegação e interação utilizam apenas um tipo de gesto (simbólico, natural e outros) ou apenas uma única técnica (filtro de partículas, *hidden Markov models* e outras) para realizar o reconhecimento dos gestos e construir a interface.

Neste trabalho, visamos a construção de uma interface híbrida baseada em gestos que utiliza tanto gestos naturais quanto gestos simbólicos, uma vez que gestos naturais se mostraram bastante adequados como interfaces de navegação [13] e gestos simbólicos constituem boas ferramentas para a construção de interfaces baseadas em comandos [10][17]. A interface de gestos para esta aplicação deve satisfazer os seguintes requisitos de interação:

- **Navegação**: movimentação do usuário em três dimensões pelo ambiente virtual;
- **Manipulação**: movimentação e rotação de objetos em três dimensões.

Para satisfazer a esses requisitos, propomos uma interface na qual as tarefas de movimentação (do usuário e dos objetos) são definidas pelos movimentos da mão direita, enquanto que a seleção e rotação de objetos é definida pelos da mão esquerda. Uma outra característica fundamental a sistemas interativos é o desempenho em tempo real, que é alcançado por meio de refinamentos dos algoritmos de visão computacional implementados.

Para melhor contextualizar e explorar as idéias descritas neste trabalho, utilizaremos como exemplo um jogo 3D tipo “quebra-cabeças”, cujo objetivo é encontrar peças em um ambiente virtual e montá-las corretamente para montar um objeto 3D. Essa aplicação permite explorar tanto a navegação no ambiente virtual quanto a manipulação de objetos.

(a) (b)

Figura 1: a) Pose associada à intenção de interação e (b) à posição de descanso.

3.1. Intenção de interação e posição de descanso

Um ponto importante para o desenvolvimento de interfaces baseadas em gestos, como proposto por Sturman e Zeltzer [18], é a definição de uma pose que indica a intenção de interação pelo usuário e outra pose de descanso. Adotamos como pose que indica a intenção de interação, que é definida no momento de inicialização do sistema, aquela ilustrada na figura 1a, marcada como “Inicial”. Nessa pose, o usuário deve se posicionar de frente para o sistema de captura, com as mãos ao lado do tronco, próximas ao peito, com as palmas da mãos voltadas para a frente.

A pose de descanso é importante para o estabelecimento de uma situação onde não há interação com o sistema. Esta posição de descanso é definida de modo que o usuário possa estabelecer momentos de descanso durante a interação, evitando seu cansaço num curto período de utilização do sistema. Para tal, definimos a posição de descanso como sendo aquela na qual o usuário posiciona as mãos relaxadas ao lado do corpo (Figura 1b).

3.2. Movimentação no Ambiente Virtual

Adotamos que a movimentação do usuário e dos objetos é regida pela movimentação da mão direita e que a manipulação é regida pela mão esquerda.

Uma vez estabelecida a intenção de interação, mover a mão direita à frente a partir da posição inicial ocasiona um movimento para a frente no ambiente virtual. Mover a mão para o lado ocasiona uma rotação para o mesmo lado no ambiente virtual. Mover a mão para cima, faz com que ocorra uma rotação para cima e mover a mão para baixo gera uma rotação para baixo. A combinação destes gestos também é possível, ou seja, mover a mão para frente e para o lado ocasiona um movimento para a frente e para a direita.

A velocidade com que o movimento no mundo virtual é realizado depende da distância da mão à posição inicial. Posicionar a mão ligeiramente à frente faz com que o movimento à frente ocorra lentamente, enquanto que esticar completamente a mão, faz com que o movimento se torne mais rápido.

Caso um objeto esteja selecionado, a movimentação deste ocorre juntamente com a movimentação do usuário, como se o usuário “carregasse" o objeto.

3.3. Seleção e Manipulação de Objetos

Ao navegar pelo ambiente virtual o usuário irá se deparar com uma ou mais peças do quebra-cabeças (pequenos cubos). Caso esteja suficientemente próximo do objeto e este se encontre próximo ao centro da imagem, o objeto é destacado dos demais, tendo sua aparência alterada, adquirindo um tom avermelhado, indicando que este pode ser selecionado. Para selecionar o objeto em destaque, o usuário deve levar sua mão esquerda à posição inicial (Figura 2) a fim de indicar sua intenção de interação e, então, esticar sua mão esquerda totalmente à frente (como um clique do mouse para seleção). O objeto em destaque será selecionado e passará a se mover junto com o usuário. A desseleção de um objeto é feita repetindo-se o movimento de seleção, isto é, esticando-se a mão esquerda totalmente à frente.

Para rotacionar um objeto, é necessário que este esteja selecionado. Novamente, o usuário deve levar a mão esquerda à posição inicial e, ao movê-la para a esquerda, o

objeto realiza uma rotação no sentido horário sobre o eixo Y (para cima). Mover a mão para a direita faz com que o objeto seja rotacionado no sentido oposto. Mover a mão para cima rotaciona o objeto no sentido horário sobre o eixo X (para a direita), enquanto que um movimento para baixo o rotaciona no sentido anti-horário sobre o mesmo eixo. A tabela 1 resume as formas de interação para a mão esquerda.

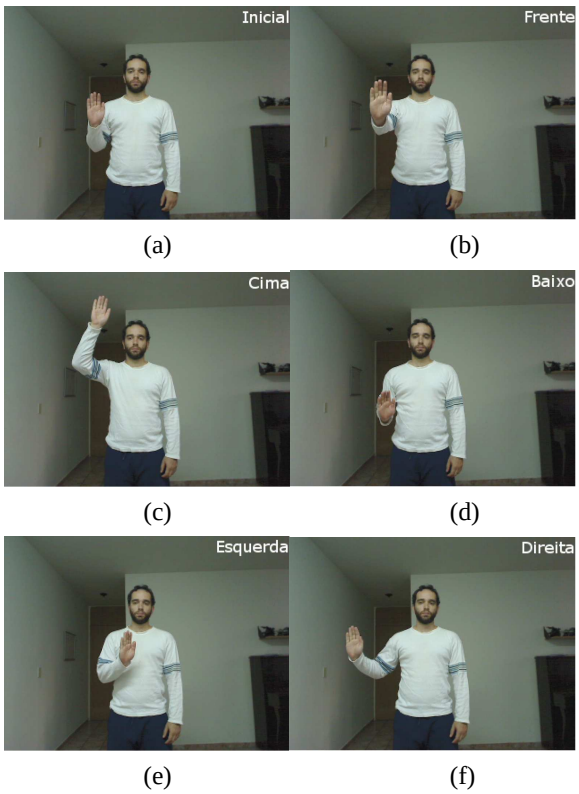


Figura 2: Associação entre poses e interações para a mão direita. A figura (a) ilustra a posição inicial, (b) para frente, (c) para cima, (d) para baixo), (e) para a esquerda e (f) para a direita.

A posição de descanso da mão esquerda, na qual nenhuma manipulação ocorre, é definida de forma semelhante à da mão direita, ao lado do corpo.

4. SISTEMA DE RECONHECIMENTO DE GESTOS

O sistema de reconhecimento de gestos para permitir os modos de interação definidos na seção anterior é constituído por duas partes independentes: o sistema da mão direita, baseado em gestos naturais, e o da mão esquerda, baseado na transição entre poses descritas numa máquina de estados.

Ação	Interação
Para frente	Seleciona/desseleciona objetos
Para cima	Roda no sentido horário em X
Para baixo	Roda no sentido anti-horário em X
Para direita	Roda no sentido horário em Y
Para esquerda	Roda no sentido anti-horário em Y

Tabela 1: mapa de interações para a mão esquerda.

4.1. Sistema de reconhecimento de gestos simbólicos

Gestos simbólicos são apropriados para a execução de comandos discretos [10][17], como rotacionar 90º para a esquerda ou para a direita. Portanto, para controlar a manipulação dos objetos, estabelecemos uma interface baseada em gestos simbólicos representados por transições entre poses estáticas descritas por uma máquina de estados finitos. Uma vantagem na utilização de máquinas de estado ao invés de outras técnicas usadas na literatura, como filtros de partículas e *hidden Markov models*, é a não necessidade de treinamento ou aprendizado. O modelo de poses é relativo ao sistema de coordenadas adotado e, desta forma, ele é automaticamente adaptado de usuário para usuário. A definição das poses pode ser feita manualmente, estabelecendo a posição das mãos dentro do espaço de interação, ou tendo como base as poses reais de um usuário.

4.2. Sistema de reconhecimento de gestos naturais

Gestos simbólicos já foram utilizados na construção de interfaces de navegação [19]. Esta abordagem, no entanto, limita o controle de navegação, pois um vocabulário de gestos simbólicos deve ser limitado e preferivelmente não muito extenso. Desta forma, definir alguns poucos gestos como ``mover à frente", ``mover para a direita" e ``mover para a esquerda", pode constituir uma solução de navegação válida, mas restringe bastante os movimentos do usuário. Gestos naturais já foram utilizados com sucesso na navegação em meio a ambiente virtuais [13] de forma a permitir uma navegação mais livre e controlada. Mover a mão para a frente, para cima ou para o lado a fim de controlar sua movimentação pelo ambiente virtual pode ser considerada uma proposta de interface de gestos naturais, pois a movimentação da mão é diretamente mapeada para a movimentação pelo mundo virtual. O mapeamento é feito gerando-se vetores de movimentação que partem da posição inicial até a posição da mão. A direção e sentido destes vetores indicam a direção e sentido do movimento pelo mundo virtual, enquanto que sua norma define a velocidade.

Para cessar a movimentação, o usuário deve retornar a mão à posição inicial. Quando o usuário deseja permanecer parado, para evitar problemas de precisão, estabelecemos uma região de tolerância ao redor da posição inicial. É necessário ultrapassar o limite desta região para que alguma movimentação tenha início.

4.3. Máquina de Estados

Uma máquina de estados finitos (MEF), ou simplesmente máquina de estados, é um modelo comportamental abstrato composto por um conjunto finito de estados, um conjunto de transições e ações associadas a estes estados. De acordo com Hopcroft *et al.* [8], uma máquina de estados finitos determinística, como as utilizadas neste trabalho, podem ser definidas formalmente por um tupla de cinco elementos $S = \{Q, \Sigma, \delta, q_0, F\}$, sendo que:

- Q é um conjunto de estados finito;
- Σ é um conjunto finito de símbolos de entrada;
- δ é uma função de transição que toma como parâmetros de entrada um símbolo de Σ em um estado de Q, retornando outro estado de Q;
- q_0 é o estado inicial e
- F é um conjunto de estados finais.

MEFs já foram utilizadas anteriormente por Bobick e Wilson [3] e, mais recentemente, por Okkonen *et.al.* [15] para realizar o reconhecimento de gestos em duas dimensões. Neste trabalho, propomos a construção de uma MEF para realizar o reconhecimento de gestos 3D da seguinte forma: construímos um conjunto Σ com **r** símbolos e dividimos o espaço 3D em **r** regiões de forma que a cada região seja atribuído um único símbolo. Um símbolo é emitido se, e somente se, o objeto associado à MEF permaneça em uma mesma região dentro de um intervalo de tempo pré-definido. Caso o objeto atravesse rapidamente uma região, o símbolo referente a ela não será emitido.

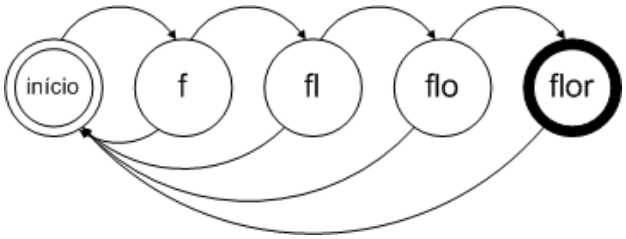


Figura 3: Estados e suas transições para o reconhecimento da palavra *flor*. O conjunto de símbolos é determinado pela letras {*f,l,o,r*} e os estados são as subsequências apresentadas.

Os estados que compõem o conjunto Q representam subsequências válidas de símbolos, conforme ilustra a Figura 3, na qual os símbolos são representados pelas letras da palavra *flor*. Cada máquina de estados possui apenas uma única sequência de estados válida, com um único estado inicial e um único estado final. O comportamento das funções de transição ocorre da seguinte forma: ao receber um símbolo emitido, se este for igual ao último símbolo da subsequência do estado atual, nenhuma mudança de estado ocorre. Caso o símbolo corresponda ao final da subsequência do próximo estado, ocorre uma mudança de estado e, se o próximo estado corresponder ao estado final, a ação associada àquela MEF é executada e todas as MEFs são reiniciadas. Caso o símbolo emitido não corresponda nem ao final da subsequência atual nem ao do próximo estado, a MEF retorna ao seu estado inicial. O comportamento das funções de transição δ é ilustrado de forma resumida na Figura 4.

Neste trabalho, construímos cinco máquinas de estados, cada uma com apenas dois estados, o inicial e o final. Seguindo as propostas de desenvolvimento de interface de Baudel e Beaudouin-Lafon [2], definimos o estado inicial de todas as máquinas como sendo o mesmo e igual à posição inicial. Os estados finais são definidos pelo deslocamento da mão à partir deste ponto, conforme discutido anteriormente. Uma vez que um gesto é reconhecido, todas as máquinas são reiniciadas e, uma vez que todas possuem o estado inicial correspondente à mesma pose, um novo gesto só passa a ser reconhecido quando a mão volta à posição inicial. Observe que gestos mais complexos compostos por sequências de posições podem ser facilmente representados por MEFs com mais estados. No entanto, o aprendizado destes gestos seria também mais complexo.

5. SISTEMA DE VISÃO COMPUTACIONAL

O sistema de visão computacional é responsável pelo rastreamento das mãos e da cabeça do usuário durante a interação. Ele é baseado no trabalho de Keskin *et al.* [9] e Azad *et al.* [1] nos quais se utilizam dispositivos não invasivos e, a partir dos dados visuais capturados por duas câmeras (webcams) em estéreo, obtém-se informações 3D sobre o posicionamento dos objetos de interesse. Uma vez obtidas as três regiões de interesse nas duas imagens, calcula-se o centróide de cada região e, baseando-se na disparidade dos centróides em cada imagem, obtém-se suas posições 3D.

Para tornar o algoritmo mais robusto e eficiente, a imagem de cada câmera é segmentada inicialmente utilizando o algoritmo Codebook [11]. A saída deste algoritmo é uma imagem binária (máscara) contendo os pixels que são diferentes do fundo, como mostra a figura 5.

Para reduzir ainda mais o custo computacional do algoritmo de visão estéreo, a disparidade é calculada apenas para os pixels com cor-de-pele. A cor-de-pele é treinada a partir de pixels selecionados manualmente das mãos e do rosto, e modelada por uma função Gaussiana no espaço de cor TSL. O resultado da segmentação de cor-de-pele pode ser vista na figura 7.

O algoritmo de visão estéreo utilizado é baseado em [7]. Para isso, as câmeras são calibradas e retificadas. Caso a cabeça e as mão sejam detectadas, o sistema calcula o centróide de cada objeto e os rastreia em 3D utilizando um filtro de Kalman independente para cada um. Os filtros de

Kalman são utilizados para evitar problemas de oclusão de curta duração e melhorar a qualidade do rastreamento.

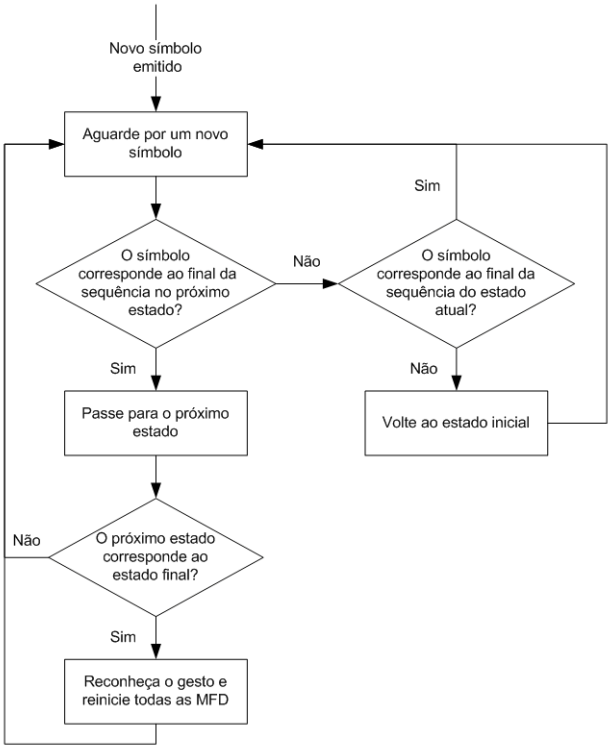
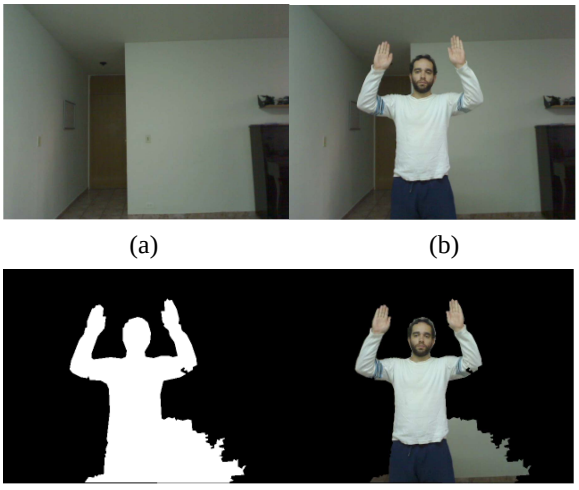


Figura 4: Comportamento da função de transição δ: diagrama de decisão.

6. PROTOCOLO EXPERIMENTAL

Para avaliação da viabilidade de utilização da interface de gestos proposta como ferramenta de navegação e manipulação de objetos no cenário do jogo proposto neste trabalho, foram realizados 3 experimentos pilotos com os seguintes propósitos:

- Avaliação da utilização da interface de gestos naturais como ferramenta de navegação;
- Avaliação da utilização da interface de gestos simbólicos como ferramenta de interação;
- Avaliação da utilização de uma interface híbrida, formada pela integração das interfaces de gestos naturais com a de gestos simbólicos, como solução de interface para o quebra-cabeças 3D.



(a) (b)

(c) (d)

Figura 5: Exemplo de segmentação de fundo utilizando o algoritmo Codebook. A imagem (a) mostra uma imagem do fundo da cena, e (b) uma imagem da cena com o usuário. A imagem em (c) mostra o resultado da segmentação (máscara), e (d) mostra a imagem do usuário combinada com a máscara.

6.1 Preparação do sistema

Antes do início de cada seção de testes, o ambiente foi preparado da seguinte forma: as câmeras do sistema de visão estéreo foram fixadas sobre um tripé com uma distância de 12,0cm entre elas, como ilustrado na Figura 6. As câmeras foram posicionadas a uma altura de 1,35m. Essa configuração permite que uma pessoa de 1.75m se posicione de braços abertos a cerca de 1.80m das câmeras e seja vista por completo dos joelhos para cima.

Fixadas as câmeras, sua calibração foi feita utilizando o método descrito em [7], por meio de 10 imagens distintas de um padrão quadriculado. Foram realizadas 5 sequências de calibrações e aquela que apresentou o menor erro de reprojeção dos pontos correspondentes (cerca de 0.17 pixel), foi utilizada no experimento.

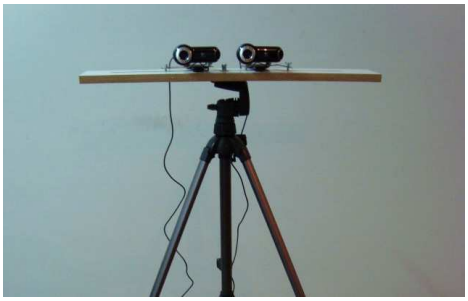


Figura 6: sistema de captura de imagens em estéreo.

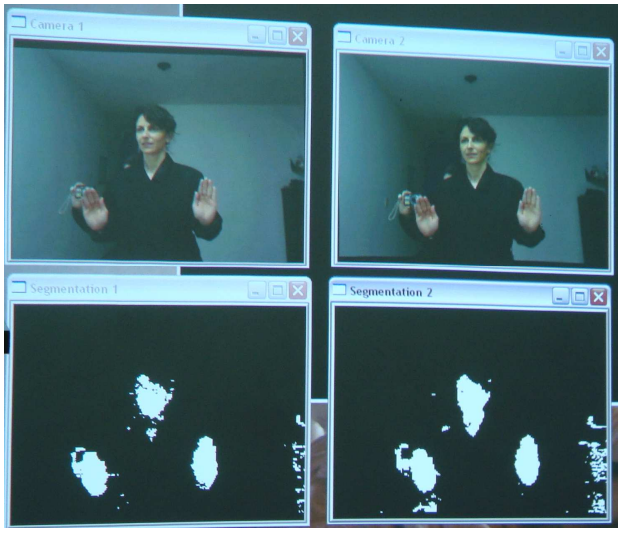


Figura 7: resultado da segmentação por cor-de-pele, para as 2 câmeras estéreo.

No passo seguinte, o modelo de fundo da cena foi construído. Para tal, desligaram-se as opções de ajuste automático de brilho, contraste e cor das câmeras para que estes não variassem durante as demais fases de preparação e de interação com o sistema. Para construir o modelo de fundo foi utilizado o algoritmo CodeBook [11], treinado utilizando uma sequência de 10 segundos de vídeo a 15 quadros por segundo, num total de 150 quadros.

Uma vez construído o modelo de fundo, fizemos a aquisição das amostras de cor-de-pele. Pedimos a cada um dos usuários para se posicionar de frente para o sistema de visão com as palmas das mão voltadas para a frente. As regiões da testa e das palmas são então selecionadas e as informações de cor desses pixels são utilizadas na construção do modelo, como ilustra a figura 7.

Feita a calibração do sistema estéreo e a inicialização dos filtros de segmentação, os objetos de interesse (mãos e cabeça) passam a ser rastreados em 2 e 3 dimensões. Cada usuário estabelece seu sistema de coordenadas local pela reprodução de quatro poses pré-definidas:

- posicionando as mãos de frente para o sistema de captura, com as palmas das mãos voltadas para a frente, à altura do peito. As posições das mãos nesta configuração são definidas como sendo o centro dos seus sistemas de coordenadas;
- movendo as mãos para a frente, definindo os eixos e limites frontais do sistema;
- movendo as mãos para cima, definindo os eixos e limites superiores do sistema;
- movendo as mãos para as laterais, definindo os eixos e limites laterais do sistema.

Definido o sistema de coordenadas local, o sistema está pronto para reconhecer gestos.

De um modo geral, não é necessário repetir as atividades de calibração e construção do modelos de fundo aos se trocar de usuário. Já as atividades de construção do modelo de cor-de-pele e definição do sistema de coordenadas local devem ser refeitas sempre que haja a troca de usuário. No sistema computacional utilizado, o rastreamento dos objetos foi feita a uma taxa de 14 quadros por segundo.

6.2 Perfil dos usuários de teste

Participaram da avaliação piloto do sistema de gestos naturais 4 usuários, 2 do sexo feminino e 2 do sexo masculino, com idades entre 25 e 55 anos e altura de 1,54m a 1,73m. Dois dos usuários tinham alguma experiência no uso de dispositivos 3D, como luva de dados e o Wii Remote, enquanto os demais não possuíam experiência prévia com sistemas tridimensionais virtuais, ou interfaces baseadas em gestos.

6.3. Avaliação do subsistema de gestos naturais

A avaliação da interface de gestos naturais foi feita medindo-se o desempenho de cada usuário para completar uma tarefa simples de navegação no ambiente virtual 3D. A movimentação é feita de modo a se alcançar cinco posições de controle distintas no espaço 3D demarcados por cubos. Dois "caminhos" diferentes foram criados, sendo que cada usuário repetiu a tarefa de navegação 3 vezes para cada caminho. Ao alcançar uma posição do caminho, o cubo marcador da posição é realçado (muda de cor), e o usuário pode então prosseguir para a próxima posição. Para isso, uma estratégia possível é se orientar no espaço, de forma a se alinhar com a próxima posição, e estender a mão direita em sua direção. Outra estratégia possível é combinar movimentos de translação e orientação. A tabela 2 mostra os tempos de execução de cada navegação, em segundos, para cada usuário.

A tabela 2 mostra os tempos médios de cada usuário, para cada caminho, e também os tempos médios de cada tentativa, para o grupo de usuários. Observe que todos os usuários foram capazes de utilizar o sistema de gestos naturais para percorrer os caminhos 1 e 2, com um pouco mais de dificuldade para percorrer o caminho 2. Os usuários mais experientes (vamos chamá-los de grupo I) apresentaram um tempo médio de execução do caminho 1 cerca de 16s menor que os usuários menos experientes (do grupo II) e, para o caminho 2, cerca de 10s menor. Este resultado é uma evidência de que uma maior familiaridade com a tarefa de navegação em ambientes tridimensionais e a utilização prévia de interfaces baseadas em gestos facilita o uso da interface proposta. Os usuários também revelaram que se sentiram mais confortáveis na terceira tentativa, o que pode justificar a redução de tempo na última tentativa com relação a primeira, que pode ser observada no desempenho da maior parte dos usuários.

Figura 8: Diferentes cenários de interação para a avaliação da interface de gestos.

Usua.	1.1	1.2	1.3	Med	2.1	2.2	2.3	Méd
1	52	47	39	46	58	51	56	55
2	41	37	33	37	53	59	48	53
3	57	51	55	54	68	72	63	68
4	59	55	64	59	61	59	63	61
Méd	52	48	48	49	60	60	58	59

Tabela 2: tempos dos testes de navegação utilizando o sistema de gestos naturais. As colunas 1.x mostram os tempos para o caminho 1, e as colunas 2.x mostram os tempos para o Caminho 2.

6.4. Avaliação do subsistema de gestos simbólicos

Os quatro usuários também participaram da avaliação do sistema de gestos simbólicos. Antes de dar início aos testes, os usuários foram apresentados ao vocabulário de gestos e foram concedidos cinco minutos de prática. As tarefas de avaliação deste subsistema consistiram na execução de cinco sequências de dez ações com a mão esquerda, que correspondem aos gestos simbólicos como definidos na tabela 1. Para cada sequência, foram contabilizados o número total de ações realizadas pelo usuário, apresentados na tabela 3. O número mínimo de ações em cada tarefa (e considerado ideal) é 10. Valores acima deste estão associados à imprecisão do sistema e erros do usuário.

Us.	1	2	3	4	5	Média
1	15	17	16	15	12	15
2	18	16	12	10	11	13
3	15	18	16	13	16	16
4	16	14	14	14	15	15

Tabela 3: contagem dos gestos durante os testes de interação com o subsistema de gestos simbólicos.

Para a interface de gestos simbólicos, ambos os grupos apresentaram um desempenho semelhante, o que aponta para o fato de uma experiência prévia com interfaces de gestos e ambientes virtuais não favorecer seu desempenho durante a utilização.

Durante a execução das tarefas, percebemos que a maior parte dos erros cometidos pelos usuários eram devido a uma confusão entre pares de gestos complementares, como ``Rotacionar para a direita" e ``Rotacionar para a esquerda". Ao ser solicitado que executasse um determinado gesto, o usuário acabava por executar o seu oposto.

Demais erros ocorridos estão associados com a precisão do sistema. Durante os experimentos, o gesto que apresentou maior número de erros foi o de ``Rotacionar para baixo". Para executar esse gesto, o usuário deve posicionar a mão esquerda abaixo da posição inicial, porém, acima da posição de descanso. No entanto, muitas vezes a mão não permanecia por tempo suficiente na região do espaço referente ao estado ``para baixo", indo direto à região de descanso, o que ocasionava a mudança indesejada do estado do sistema para ``descanso". Isso indica que o sistema precisa fornecer uma indicação ao usuário (feedback) sobre o seu estado, que neste protótipo ainda não foi implementado.

No entanto, como o conjunto de gestos é bastante simples e incluem ações complementares, a correção dos erros se mostrou simples a todos os usuários, permitindo que cada um completasse todas as tarefas (seqüência de ações) com sucesso.

6.5. Avaliação da interface de gestos híbrida

Participaram da avaliação da interface de gestos híbrida apenas os usuários do grupo I. A tarefa realizada durante a avaliação consiste em montar o quebra-cabeça 3D utilizando as duas interfaces simultaneamente. Para simplificar a tarefa de montagem do quebra-cabeças e facilitar a visualização do ambiente e dos objetos, cada tarefa consistia na montagem de apenas uma peça. Assim, para cada teste, apenas uma peça do cubo foi retirada e posicionada numa orientação aleatória. O espaço também foi "quantizado" de forma que cada peça era ajustada automaticamente ao quebra-cabeças quando esta era colocada suficientemente perto de sua posição correta. Foram realizadas três seções com cada usuário e, em cada seção, uma peça diferente foi removida. A figura 8 ilustra a configuração inicial das peças em cada seção. Foram

medidos os desempenhos dos usuários para a conclusão de cada tarefa (em segundos) e contados o número de comandos executados, como apresentados nas tabelas 4 e 5.

Cen	Usuário 1			M1	Usuário 2			M2
1	55	44	40	46	79	51	46	59
2	48	19	25	31	23	32	31	29
3	62	38	53	51	40	62	29	44

Tabela 4: Tempos, em segundos, obtidos pelos usuários durante os testes da interface de gestos híbrida.

Cen	Usuário 1			M1	Usuário 2			M2
1	4	16	4	8	26	9	4	13
2	2	2	2	2	2	2	2	2
3	4	4	8	5	6	4	4	5

Tabela 5: Número de comandos emitidos pelos usuários durante os testes da interface de gestos híbrida.

Conforme podemos observar pelos resultados das tabelas 4 e 5, o cenário 1 da figura 8 foi o que apresentou uma dificuldade maior em ser completado. Em seguida está o cenário 3 e, por fim, o cenário 2. Nos cenários 1 e 2 a peça faltante encontrava-se não apenas fora da sua posição correta, mas também fora de sua orientação ideal. Já no cenário 3, ela estava apenas fora de lugar.

Observe que o usuário 2 teve alguma dificuldade durante sua primeira interação com o cenário 1, como pode ser observado pelo longo tempo que levou para completar a tarefa, assim como o elevado número de ações realizadas. No entanto, as demais seções transcorreram sem maiores dificuldades.

Considerando que o número mínimo de comandos para colocar as peças em sua posição correta é 2, 2, 0, para os cenários 1, 2 e 3, respectivamente, podemos observar que os erros estão aleatoriamente distribuídos, visto que o usuário 1 cometeu mais erros na segunda e terceira seções, enquanto o usuário 2 cometeu mais erros nas primeiras seções. No entanto, considerando os tempos da 1ª e 3ª seções, pode-se notar que ocorreu uma melhora significativa de desempenho. Devido a repetição da tarefa, é provável que, uma vez descoberta as ações para colocar a peça corretamente no cubo, os usuários foram capazes de se aproximar ao número ideal de comandos.

Apesar destes resultados ainda bastante preliminares, eles indicam que a interface híbrida combina de forma eficaz a liberdade de movimentos dos gestos naturais à praticidade dos gestos simbólicos, e sem exigir um longo período de treinamento e adaptação do usuário.

7. COMENTÁRIOS FINAIS

Neste trabalho apresentamos um sistema de interação em ambientes virtuais tridimensionais utilizando uma combinação de gestos naturais, apropriados para a navegação no ambiente, e gestos simbólicos, apropriados para a seleção e manipulação de objetos na cena.

O sistema utiliza um sistema de visão computacional de baixo custo e de tempo real, que usa câmeras estéreo para o rastreamento dos gestos. O sistema é eficiente pois as mãos e a cabeça são previamente segmentados usando o algoritmo Codebook para segmentação de fundo e também um classificador de cor-de-pele para reduzir o espaço de procura. Uma limitação ainda séria do sistema de visão é o seu tempo de preparação, dada a necessidade de calibração das câmeras e do usuário.

Os resultados do estudo piloto com a interface baseada em gestos mostrou que a interface é bastante simples de aprender e utilizar, mesmo por usuários sem experiência. Em trabalhos futuros, pretendemos comparar o desempenho da interface com outros dispositivos de apontamento 3D, para diferenciar erros devido a complexidade da interface, de erros devido a dificuldade da tarefa de manipulação de objetos em 3D.

REFERÊNCIAS

1. P. Azad, A. Ude, T. Asfour, and R. Dillmann. *Stereo-based markerless human motion capture for humanoid robot systems*. Robotics and Automation, 2007 IEEE International Conference on, pages 3951-3956, April 2007.
2. T. Baudel and M. Beaudouin-Lafon. *Charade: remote control of objects using free-hand gestures*. Communications of the ACM, 36(7):28-35, 1993.
3. A.F. Bobick and A.D. Wilson. *A state-based approach to the representation and recognition of gesture*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(12):1325-1337, Dec 1997.
4. D.A. Bowman, E.Kruijff, J.J. LaViola and I. Poupyrev. *3D User Interfaces: Theory and Practice*. Addison-Wesley Professional; 1st edition, 2004.
5. M. Bhuiyan and R. Picking. *Gesture-controlled user interfaces, what have we done and what's next?*, In: Proceedings of the Fifth Collaborative Research Symposium on Security, E-Learning, Internet and Networking (SEIN 2009), Darmstadt, Germany, 26-27 November 2009, pp59-60.
6. C. Cadoz. *Le geste canal de communication homme/machine: la communication _instrumentale* TSI. Technique et science informatiques, 13(1):31-61, 1994.
7. R. Hartley and A. Zisserman. *Multiple View Geometry* (2nd Edition), chapter 9 - 11. Cambridge University Press, Cambridge, CB2 8RU, UK, 2008.

8. J.E. Hopcroft, J.D. Ullman, and R. Motwani. *Automatos Finitos determinísticos*, pages 48-56. Editora Campus, 2a edition, 2002.
9. C. Keskin, O. Aran, and L. Akarun. *Real time gestural interface for generic applications*. In: European Signal Processing Conference, 2005.
10. J.H. Kim. *An HMM-based threshold model approach for gesture recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(1010):961-973, 1999.
11. K. Kim, T.H. Chalidabhongse, D. Harwood, and L. Davis. *Real-time foreground background segmentation using codebook model*. Real-Time Imaging, 11(Video Object Processing):172-185, July 2005.
12. S. Kita. *Cross-cultural variation of speech-accompanying gesture: A review*. Language and Cognitive Processes, 24(2):145-167, 2009.
13. C. Manders, F. Farbiz, T.K. Yin, Y. Miaolong, B. Chong, and C.G. Guan. *Interacting with 3D objects in a virtual environment using an intuitive gesture system*, volume 1. ACM New York, NY, USA, 2008.
14. S. Mitra and T. Acharya. *Gesture recognition: A survey*. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 37(3):311-324, May 2007.
15. M. Okkonen, V. Kellokumpu, M. Pietikainen, and J. Heikkila. *A visual system for hand gesture recognition in human-computer interaction*. In: Image Analysis, SCIA 2007. Proceedings, Lecture Notes in Computer Science 4522, 709-718, 2007.
16. B. Rime and L. Schiaratura. *Gesture and speech*, pages 239-281. Editions de la Maison des Sciences de l'Homme, 1991.
17. A. Seth, S.S. Smith, M. Shelley, and Q. Jiang. *A low cost virtual reality human computer interface for cad model manipulation*. The Engineering Design Graphics Journal, 69(2):31-38, 2005.
18. D.J. Sturman and David Zeltzer. *A design method for "whole-hand" human-computer interaction*. ACM Transactions on Information Systems (TOIS), 11(3):219-238, 1993.
19. J. Yamato, J. Ohya, and K. Ishii. *Recognizing human action in time-sequential images using hidden Markov model*, page 379-385. 1992.