Multi-Modal Painter: Pintar Usando as Mãos e Fala

Mariana Cerejo José Santos Alfredo Ferreira Manuel J. Fonseca Joaquim A. Jorge

Grupo de Interfaces Multi-Modais Inteligentes
Departamento de Engenharia Informática
INESC-ID/IST/ Universidade Técnica de Lisboa

http://immi.inesc-id.pt

Sumário

Nos últimos anos várias equipas de investigação têm vindo a desenvolver um novo paradigma de interacção entre humanos e computadores, procurando reduzir ou eliminar totalmente a necessidade de manuseamento de dispositivos físicos para controlar a máquina. O trabalho aqui apresentado foca a integração de reconhecimento de gestos e fala para controlo de aplicações. Para validarmos esta integração desenvolvemos uma aplicação multi-modal que consiste numa ferramenta de desenho 2D. Com esta aplicação o utilizador pode desenhar e pintar utilizando apenas gestos e comandos de fala.

Palavra-chave

Interfaces Multi-Modais, Reconhecimento de Gestos, Reconhecimento de Fala

1. INTRODUÇÃO

Para além da evolução tecnológica, a necessidade de dar uma resposta às exigências dos actuais utilizadores, serviu de impulso para um novo conceito de interacção pessoa-máquina. O antigo conceito de WYSIWYG ("What You See Is What You Get"), que colocava o utilizador como agente externo e manipulador, está prestes a deixar de ser a única verdade absoluta neste assunto. Assim, o utilizador deixa de ser a entidade que manipula os dispositivos de introdução de informação, tornando-se ele próprio o "veículo da informação".

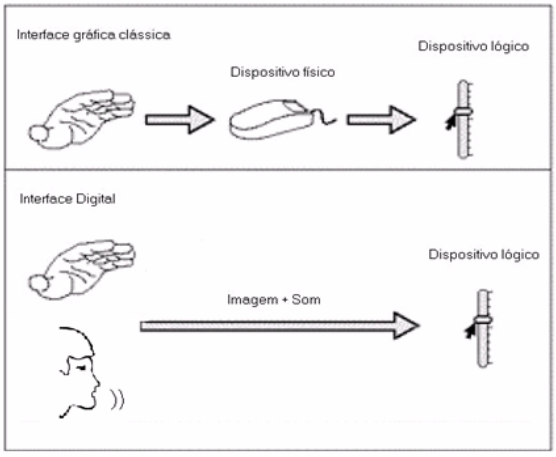

Figura 1 – Mudança de paradigma de Interface

Seguindo esta linha de desenvolvimento, apresentamos um paradigma baseado na utilização de gestos (com as mãos) e comandos de fala para controlar o sistema.

Para realizar os gestos, o utilizador coloca dedais de cores diferentes nos dedos. Utilizando uma câmara captam-se os gestos do utilizador, enquanto que com um microfone, é feito exactamente o mesmo, mas para a fala. Ambos os sinais de entrada são reconhecidos e posteriormente processados e interpretados, de acordo com o contexto corrente.

2. ARQUITECTURA

A Figura 2 representa uma arquitectura geral do sistema. Como se pode constatar, existe uma estruturação em camadas que definem cada módulo. Assim, existe um módulo que trata o reconhecimento gestual, um outro que trata o reconhecimento de fala e o principal que consiste na integração multi-modal dos dois. Além destes, existe ainda a própria aplicação, que faz a fusão multi-modal das duas técnicas de interacção.

Esta arquitectura permite-nos ter um sistema distribuído, podendo ter cada componente em computadores separados, ou seja, se desejarmos podemos ter uma máquina responsável pelas câmaras (componente gestual), outra pelos microfones (componente de fala) e outra responsável pela aplicação.

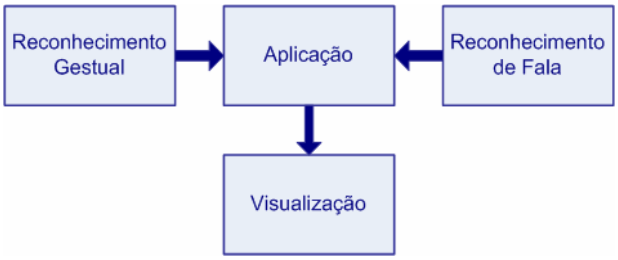

Figura 2 – Arquitectura geral do *Multi-Modal Painter*

Acrescentamos ainda um componente de calibração ao módulo de reconhecimento gestual, para minimizar os efeitos da variabilidade da luminosidade do ambiente na taxa de reconhecimento.

3. RECONHECIMENTO GESTUAL

Uma das mais valias do nosso sistema está no facto de suportar qualquer tipo de fundo, dispensando assim a necessidade de um ambiente controlado para obter boas taxas de reconhecimento. No entanto, é necessário proceder a uma fase de calibração das cores das marcas, para que estas possam ser reconhecidas.

3.1 Calibração

Para o bom funcionamento do sistema é necessária uma calibração prévia sempre que mudamos de ambiente. Para haver um seguimento de marcas robusto foi necessário ter em conta as variações de luminosidade nas diferentes áreas da imagem capturada pela câmara. Assim, a interface utilizador apresenta cinco quadrados (quatro nos cantos da imagem e um no centro) em que o utilizador terá de colocar cada marca, captando toda a informação necessária para se realizar a calibração (ver Figura 3).

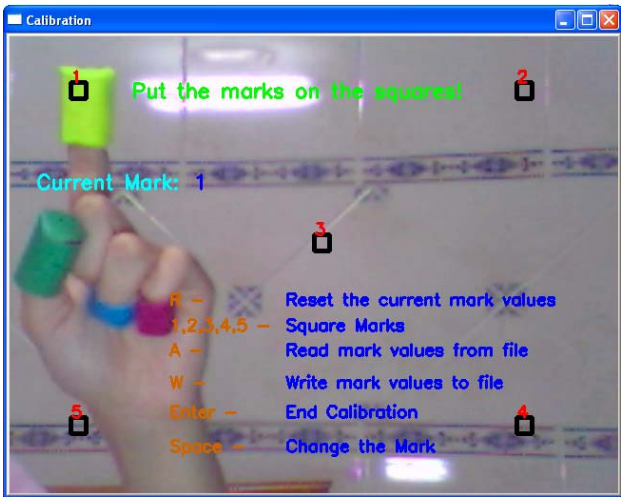

Figura 3 – Calibração inicial das marcas

Esta informação fica guardada num ficheiro, que será posteriormente utilizada pela aplicação, sendo apenas necessário executar a calibração uma única vez, desde que não haja alterações de luminosidade no ambiente, ou movimentos bruscos na câmara.

3.2 Detecção de gestos

A identificação dos gestos realizados pelo utilizador é composta por três passos: primeiro fazemos detecção das marcas que se encontram nos dedos do utilizador; segundo, construímos polígonos ligando os centróides de cada marca (sempre pela mesma ordem); finalmente, identificamos o polígono resultante de modo a reconhecer a pose realizada.

A detecção das marcas durante a interacção é baseada em técnicas de visão por computador. Através dos valores obtidos na calibração, aplicamos filtros passa-banda para a tonalidade e passa-alto para a saturação, obtendo-se desta forma a posição das marcas. Em conjunto com a aplicação dos filtros, utilizamos também algoritmos para eliminação de ruído. Identificada a localização das marcas, calculamos os seus centróides, e com base nestes o polígono correspondente.

Para classificar os polígonos resultantes, utilizamos uma biblioteca de reconhecimento de formas geométricas, chamada CALI [Fonseca02]. Esta biblioteca usa um conjunto de características baseadas em áreas e perímetros dos maiores e menores triângulos/rectângulos, criados a partir do polígono original, conseguindo-se uma boa precisão na detecção dos gestos (ver Figura 4).

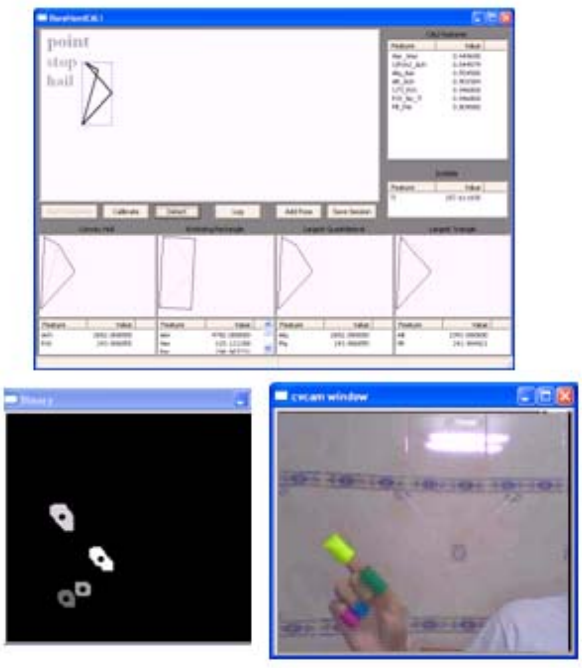

Figura 4 – Reconhecimento de gestos com o CALI

De modo a identificar os valores de cada característica para cada tipo de pose, realizamos testes com oito utilizadores. Começamos por pedir a cada utilizador que colocasse os dedais (marcas) nos dedos e que efectuasse todos os gestos que se encontram na Figura 5, por ordem. Depois deste passo, voltamos a pedir a cada utilizador que voltasse a repetir as poses, mas agora com o objectivo de medir a taxa de reconhecimento. A taxa de reconhecimento situou-se nos 93%. Estudos posteriores revelaram que a utilização de uma árvore de decisão, para situações em que uma das marcas fica oculta, em combinação com a biblioteca CALI apresenta melhores resultados.

No estudo experimental, procuramos ainda verificar e validar a semântica associada a cada um dos gestos.

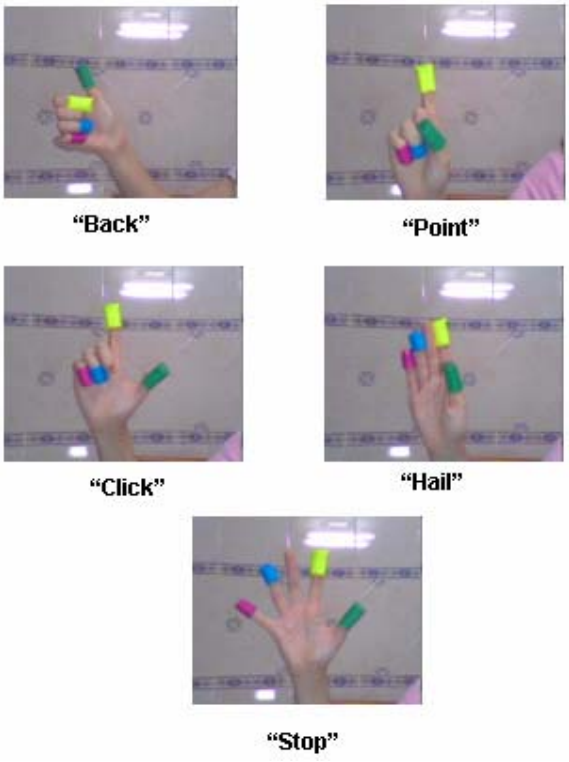

Figura 5 – Gestos Reconhecidos

3.3 Semântica gestual

Para cada gesto estudamos qual seria o melhor significado que lhe poderíamos atribuir, no contexto da nossa aplicação de desenho. Decidimos então atribuir a cada gesto um único evento, conforme descrito na Tabela 1. Para simplificar a aplicação e utilizar o menor número de gestos (para ser mais simples para o utilizador), cada gesto pode ter uma semântica diferente consoante o seu contexto, como podemos ver pelo exemplo do Click. Neste caso, o gesto Click servirá não só para simular o botão esquerdo do rato (para escolha de opções ou de determinados comandos) mas também para efectuar a própria pintura na aplicação.

Tabela 1 – Emparelhamento Gesto-Evento

Gesto	***Evento***
Back	Botão direito do rato
Point	Controlo do cursor
Click	Botão esquerdo do rato
Hail	Apagar
Stop	*Blur*

A escolha dos gestos e respectivas características de reconhecimento foram objecto de estudo conforme se pode analisar em [Ferreira06].

4. RECONHECIMENTO DE FALA

A outra técnica de interacção abordada consiste no reconhecimento de fala que, possibilita ao utilizador proferir comandos específicos.

Assim, a contribuição essencial do reconhecimento de fala consiste em permitir ao utilizador dar comandos, mais rapidamente do que usando gestos ou outro modo de interacção mais tradicional. Por exemplo, para efectuar a acção cortar/colar seria necessário efectuar a opção de "seleccionar", "cortar" mover para o sítio desejado e ainda seleccionar novamente a opção "colar" no sistema de menus. Com a fala esta interacção torna-se muito mais eficaz pois não se fica preso ao sistema de menus. Outro exemplo, um utilizador encontra-se a efectuar um desenho com a cor vermelha e pretende mudar para a cor verde. Tem duas opções:

- Através de gestos, selecciona no menu correspondente, a cor verde;
- Profere a palavra "Green" e a cor é automaticamente mudada.

A tecnologia de reconheciemnto de fala adoptada foi o Microsoft Speech SDK, pois a sua taxa de reconhecimento é alta, mesmo sem treino prévio. Além disso, consegue não só reconhecer palavras isoladas mas também frases completas. No nosso trabalho pretendemos apenas identificar comandos isolados.

5. APLICAÇÃO

Para validar o paradigma proposto, criámos uma aplicação multimodal de desenho em 2D que é controlada por gestos e fala (ver Figura 6). De notar que, apesar da nossa escolha para a validação ter recaído sobre esta opção, a nossa abordagem é suficientemente genérica para ser aplicada em alternativas distintas, igualmente válida para o efeito, tais como "browsers" ou modelação.

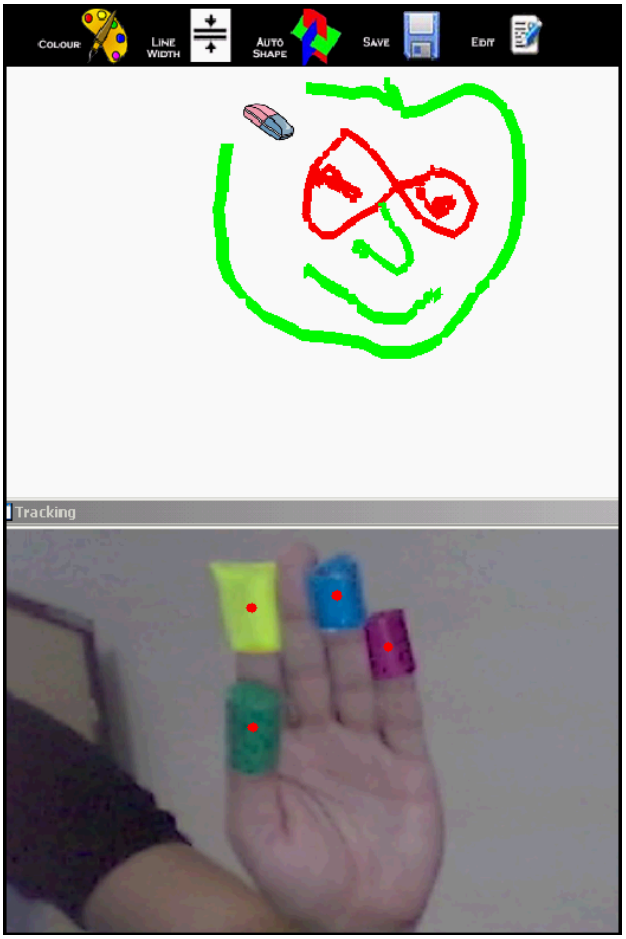

Figura 6 – Multi-Modal Painter

A aplicação desenvolvida permite efectuar um conjunto de acções, quer utilizando apenas gestos, quer usando as duas técnicas. Combinando ambas as técnicas, as limitações de uma são complementadas pela outra e vice-versa. Assim pode-se desenhar, apagar, escolher a cor e a grossura do traço, gravar e abrir o desenho, escolher uma forma predefinida (por exemplo, quadrado ou triângulo), copiar, cortar e colar uma área, preencher, mover e seleccionar um objecto, apagar tudo, fazer *blurring* ou recorrer à ajuda, usando gestos e/ou fala.

6. CONCLUSÕES E TRABALHO FUTURO

O objectivo inicial, que consistia em demonstrar que existem novos paradigmas de interacção que podem substituir os convencionais, torna-se facilmente concreto pois neste caso o utilizador tornou-se veículo de informação não sendo necessário recorrer a periféricos externos para o efeito.

Apesar de ainda não se terem realizado testes formais com utilizadores, os resultados obtidos foram bastante promissores pois consegue-se fazer o controlo da aplicação utilizando apenas gestos e fala sem recorrer aos dispositivos habituais.

Em suma, os resultados obtidos foram bons mas futuramente iremos elaborar uma sessão de testes com utilizadores para obter uma avaliação de usabilidade válida.

Para além da avaliação, pretendemos ainda estender a nossa abordagem para aplicações distintas da aqui apresentada. Nomeadamente, na sua adaptação com vista a obter um sistema de gestão de documentos semelhante ao BumpTop 3D Desktop [Agarawala96].

7. REFERÊNCIAS

[Agarawala96] Anand Agarawala, Ravin Balakrishnan, Keepin' it Real: Pushing the Desktop Metaphor with Physics, Piles and the Pen. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2006)*.

[Ferreira06] Alfredo Ferreira, Ricardo Jota, Mariana Cerejo, José Santos, Manuel J. Fonseca, Joaquim A. Jorge. Recognizing Hand Gestures with CALI , aceite para publicação em *Ibero-American Symposium on Computer Graphics (SIACG 2006)*

[Fonseca02] Manuel J. Fonseca, César Pimentel and Joaquim A. Jorge, CALI: An Online Scribble Recognizer for Calligraphic Interfaces, *Proceedings of the 2002 AAAI Spring Symposium - Sketch Understanding*, pages 51-58, Palo Alto, USA, Mar 2002