2
Interfaces Baseadas em Visão

Neste capítulo, numa primeira parte, será feita a revisão de alguns conceitos básicos, importantes para explicar e contextualizar as interfaces baseadas em Visão. São discutidos alguns requisitos funcionais e não-funcionais dos sistemas de Visão em tempo real. Também é apresentada uma justificativa de por que consideramos a Visão Computacional importante na interação baseada em gestos. Na segunda parte deste capítulo são apresentados alguns trabalhos relacionados com o tema de pesquisa aqui exposto.

2.1.
Conceitos Básicos

Muitos dos sistemas de interação baseados em gestos tradicionalmente foram implementados utilizando diversos dispositivos de rastreamento (*Tracking*) ligados ao computador e ao corpo dos usuários. Os sistemas de interação baseados em Visão não utilizam dispositivos de rastreamento explícitos. Eles utilizam apenas câmeras para a captura das imagens, e técnicas de processamento de imagens e reconhecimento de padrões para o rastreamento dos objetos [11]. Não existem restrições para as características das câmeras. Conforme o fazem algumas aplicações [27], podem ser utilizadas câmeras infravermelhas, sensíveis à temperatura, baseadas em distância, etc. Desde que a câmera seja a única fonte de captura de informação, dizemos que o mecanismo de interação é baseado em Visão Computacional.

No contexto da interação Humano-Computador o termo “tempo real” é freqüentemente substituído pelo termo “fortemente acoplado” [7]. Fitzmaurice [7] descreve essa expressão como: “Os sistemas fortemente acoplados possuem uma perfeita sincronização entre suas representações física e virtual, os objetos físicos são detectados continuamente em tempo real”. O termo “perfeitamente sincronizado” também requer uma definição. Em aplicações reais sempre há uma latência (*delay*) entre a modificação do mundo físico e a adaptação da representação virtual no computador.

O que tem mudado nestes últimos 20 anos na maneira como os usuários interagem com o computador? A resposta é simples: a interação tem mudado do teclado para o mouse, o teclado é utilizado em muitos casos unicamente para a digitação de texto. Hoje é difícil pensar em um computador sem mouse, pois esse dispositivo permite uma interação mais intuitiva com objetos gráficos como botões, janelas, menus, barras de rolagem, etc., os quais antigamente precisavam do uso do teclado. Assim como o teclado foi substituído pelo mouse, recentemente e com ajuda da Visão Computacional têm aparecido novas interfaces com a finalidade de descartar os dispositivos físicos de interação. No trabalho de Bérard [1] é feita uma classificação que ajuda a entender melhor a natureza dessas interfaces (Figura 1).

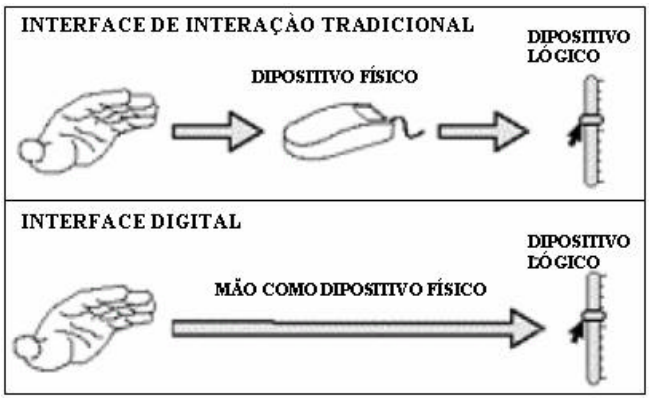

Figura 1 - Interfaces tradicionais e interfaces digitais. Extraído de Bérard [1].

Nas interfaces de interação tradicionais, a presença de um dispositivo físico é importante para traduzir os comandos do usuário para o computador. Se o usuário manipula o dispositivo (por exemplo, muda sua posição) o computador recebe o sinal da mudança e manipula o objeto lógico ligado com esse dispositivo. Como acontece quando o mouse muda de posição, o computador mede essa mudança e modifica a posição do ponteiro. Joystick, Space Ball, Cyber Gloves, etc. são considerados no escopo desse tipo de interface, chamados também por Bérard [1] de *Graspable Interfaces,* ou interfaces que o usuário precisa agarrar para poder interagir. Nessas interfaces, dependendo do dispositivo, podem ser integradas diferentes funcionalidades além dos cliques do mouse (p. ex., posicionamento no espaço e rotações).

Ao invés de usar os dispositivos físicos como intermediários, as interfaces digitais eliminam a presença desses dispositivos e permitem uma interação mais direta utilizando partes do corpo para manipular os objetos lógicos no computador. Por exemplo, os dedos da mão poderiam ser utilizados para

movimentar o ponteiro do mouse. Uma das interfaces digitais mais conhecidas são as telas sensíveis ao tato.

2.2. Requisitos Funcionais

Os requisitos funcionais podem ser definidos como sendo um conjunto de serviços que o sistema deve fornecer. Em sistemas de software existem diferentes níveis de abstração em que esses serviços podem ser desenvolvidos. Aqui, serão considerados apenas os mais básicos. Bérard [1] identifica três serviços que os sistemas de interação Humano-Computador baseados em Visão devem fornecer, estes são: detecção, identificação e rastreamento.

A detecção determina a presença ou ausência de uma determinada classe de objetos na imagem. Tais classes de objetos poderiam ser partes do corpo, mãos, braços, etc. Tendo como referência a imagem inteira, o processo de detecção deve ser capaz de detectar na imagem a classe de objeto que se está procurando. Uma forma de facilitar o processo de detecção é limitar o número de objetos que podem estar presentes na cena em um determinado momento. As técnicas de detecção mais conhecidas são as baseadas em cor ou movimento. O processo de detecção deste trabalho será apresentado nos capítulos 3 e 4.

A identificação determina qual objeto, dentre um conjunto conhecido de objetos, está presente na cena. Perante a presença de objetos compostos, por exemplo uma mão com os dedos, a identificação deve permitir determinar partes desses objetos, tais como os dedos. Outros exemplos são a identificação de símbolos escritos [35], de palavras na linguagem de signos [36] ou de palavras para o reconhecimento de voz. No nosso caso, o processo de detecção encontra os dedos presentes na mão. Para obter mais informações do gesto, o processo de identificação está orientado a identificar os dedos polegar e indicador. Esse processo será apresentado no capítulo 5.

Em muitos casos, os objetos de interesse não permanecem no mesmo lugar ao longo do tempo, o processo de rastreamento utiliza as informações dos dois processos anteriores para manter o foco nos objetos de interesse. No caso deste trabalho o rastreamento se refere à captura das posições de interesse nos dedos.

2.3. Requisitos Não-Funcionais

Muitos dos sistemas de interação podem cumprir com todos os requisitos funcionais, mas demorar horas para fazer isto. Portanto, é indispensável definir alguns requisitos não-funcionais que estabeleçam a qualidade mínima com que os serviços devem ser implementados. Os requisitos não-funcionais considerados são latência, resolução e estabilidade.

A latência é o tempo transcorrido entre a ação do usuário e a resposta do sistema. A latência é uma característica inerente a todo sistema, não existem sistemas sem latência. Frente a este problema, poderíamos perguntar: qual é a latência máxima que um sistema de tempo real pode ter? Essa pergunta é difícil de se responder já que é difícil determinar um valor aceitável, pois os valores dependem da tarefa que o sistema deve desempenhar. Seguindo a definição de interação em tempo real, tentamos conseguir uma interação sem uma latência perceptível para o usuário. Mas qual é o máximo valor de uma latência não perceptível? No trabalho de Handenberg [11] é discutido um experimento que visa encontrar um valor máximo para essa latência, o qual é determinado como sendo próximo a 50 ms ou 20Hz.

Nos sistemas de interação em tempo real, a natureza contínua do mundo físico, onde são feitos os gestos, deve ser capturada e representada no mundo discreto ou mundo perceptível do sistema (Resolução Temporal). A ilusão desse mundo em movimento, como no cinema, pode ser criada capturando imagens a uma taxa maior que 20Hz, valor diretamente relacionado com a latência do sistema. A resolução espacial (número de pixels nas imagens) deve permitir uma representação adequada do ambiente capturado. Idealmente, esse número deveria ser igual ao número de pixels existente nos monitores, mas os sistemas de captura (câmeras) não possuem ainda essa resolução e estão limitados a resoluções menores.

A estabilidade dos sistemas refere-se às flutuações significativas nos valores capturados (gestos ou posições dos dedos, neste caso). Por exemplo, um sistema pode ser considerado estável se ante um padrão (ex. ponta do dedo) imóvel os valores capturados não mudam significativamente. As possíveis causas de instabilidade são principalmente as flutuações nas fontes de iluminação e o ruído inerente aos dispositivos de captura.

2.4. Visão Computacional na Interação Humano-Computador em Tempo Real

No mercado existem diferentes dispositivos que permitem aos usuários usar as mãos para interagir com o computador. Alguns exemplos são teclado, mouse, TrackBall, Track-Pad, Joystick e controles remotos. Outros, mais sofisticados, incluem Cyber-Gloves, 3D-mice (ex. Labtec SpaceBall) e dispositivos de rastreamento magnético (ex. Polhemus Isotrack) ou mecânico. Muitos desses dispositivos são mais baratos, confiáveis e fáceis de fazê-los funcionar do que os atuais sistemas baseados em Visão Computacional. A evolução dos sistemas de Visão Computacional, entretanto, promete resultados melhores num futuro próximo.

2.4.1. Vantagens dos Sistemas Baseados em Visão

Primeiramente, a Visão Computacional é uma tecnologia com um grande potencial para integração em micro-circuitos digitais. Portanto a produção em massa é muito mais fácil de se realizar do que outros dispositivos com partes mecânicas, como os Cyber Gloves, por exemplo. Considerando o crescimento na velocidade de processamento existente, os custos de processamento das imagens poderão ser descartados.

Outra vantagem, muito importante, é a versatilidade. Enquanto outros dispositivos tais como mouse, Joystick e Track-Pad são limitados a funções específicas, a Visão Computacional oferece uma ampla gama de possíveis aplicações, não somente na área da interação humano-computador, mas também em áreas tais como a identificação de usuários e vídeo-conferências. Essas aplicações tornam interessante a inclusão de câmeras em produtos tais como monitores, notebooks, telefones celulares, televisores, projetores e videogames (ex. *EyeToy*).

Na nossa opinião, a principal vantagem da Visão Computacional é a não-intrusividade. Assim como os microfones ou câmeras, os dispositivos baseados em Visão são abertos e não precisam do contato físico com o usuário para interagir. Os usuários podem interagir com o computador livremente, sem cabos e sem manipular dispositivos intermediários. Por esta razão, procuraremos desenvolver algoritmos de Visão Computacional que não requeiram equipamentos colados ao corpo (tais como marcadores ou luvas coloridas).

Esses marcadores poderiam simplificar vários problemas dos algoritmos de Visão, mas destroem a principal vantagem da Visão Computacional que é a não-intrusividade.

Idealmente, a parte técnica do sistema de Visão deveria estar escondida do usuário, o qual, com gestos e movimentos do corpo, poderia se comunicar com o computador. Em conjunção com o reconhecimento de voz, pode-se facilmente imaginar um computador que permita uma interação muito mais natural e intuitiva do que os atuais dispositivos. Claramente este objetivo é mais fácil de se imaginar do que de se fazer. Apesar do avanço das pesquisas recentes em Visão Computacional, ainda não dispomos de produtos comercias bem sucedidos.

2.4.2. Desafios

Muitos dos problemas da Visão Computacional, como por exemplo o de detectar uma mão em movimento sobre um fundo relativamente constante, que parecem simples à primeira vista, são na realidade produto de um complexo processo realizado em nosso cérebro. Reproduzir estes processos no computador é o desafio da Visão Computacional.

Um dos grandes problemas é a quantidade de informação de entrada disponível. A retina tem aproximadamente 125 milhões de células receptivas para capturar informações do ambiente [11]. Mesmo se as câmeras (resolução máxima de captura em tempo real das *WebCam*: 480x680) tivessem a mesma capacidade de captura de informação, os computadores não possuem o enorme poder de processamento em paralelo que possui o nosso cérebro. Portanto, somente alguns processos básicos são implementados na maioria dos sistemas em tempo real baseados em Visão.

Outro problema da Visão é a baixa confiabilidade e a instabilidade, ocasionadas, entre outras coisas, por mudanças de iluminação, oclusão, movimento e ruído nos equipamentos de captura. O sistema de visão humano integra várias características que são observadas em paralelo (ex. cor, movimento, contornos) junto ao seu conhecimento do mundo para lidar com esses problemas. Conseguir isto num computador não é uma tarefa fácil.

Nossas capacidades são o fruto da integração de nosso conhecimento do mundo (nossa experiência) ao longo de nossas vidas e é importante considerar isto nos algoritmos de Visão se quisermos resultados mais robustos.

Por estas razões é difícil construir um sistema de Visão Computacional de propósito geral que seja capaz de trabalhar com toda classe de objetos e em todos os ambientes. É preciso restringir o campo de ação e construir sistemas mais específicos.

2.5. Trabalhos Relacionados

Nos últimos anos, muitos trabalhos de pesquisa relacionados à interação Humano-Computador baseada em Visão têm sido desenvolvidos. No que diz respeito à interação utilizando a mão, muitos dos sistemas desenvolvidos têm se dedicado ao reconhecimento de gestos e posturas, o que significa que a interação com o computador é baseada no reconhecimento de diferentes posturas e gestos da mão. Uma interessante diversidade de abordagens tem sido apresentada, mas não existe uma que seja amplamente utilizada para o reconhecimento de gestos.

Por outro lado os sistemas que, além de reconhecer alguns gestos básicos, interagem utilizando os dedos da mão têm captado pouco interesse. Não obstante, os sistemas existentes são de grande interesse para este trabalho porque na maioria dos casos se depararam com os mesmos problemas de segmentação e rastreamento que encontramos em nossa pesquisa.

2.5.1. Sistemas Baseados no Reconhecimento de Gestos

Segundo Pavlovic [28], os sistemas de reconhecimento de gestos da mão podem ser classificados em dois grupos: os baseados no modelo 3D e os baseados na aparência da imagem 2D.

Nos sistemas baseados no modelo 3D, o modelo da mão é definido através de um conjunto de parâmetros que descrevem todos os graus de liberdade da mão. Geralmente, o mecanismo consiste em procurar um conjunto de parâmetros que melhor ajustem o modelo 3D e sua projeção 2D na imagem (Figura 2). Alguns trabalhos que aplicam esta técnica são os de Kuck & Huang [18], Lee & Kunii [24], Regh & Kanade [30], Stenger [37] e Wu [46].

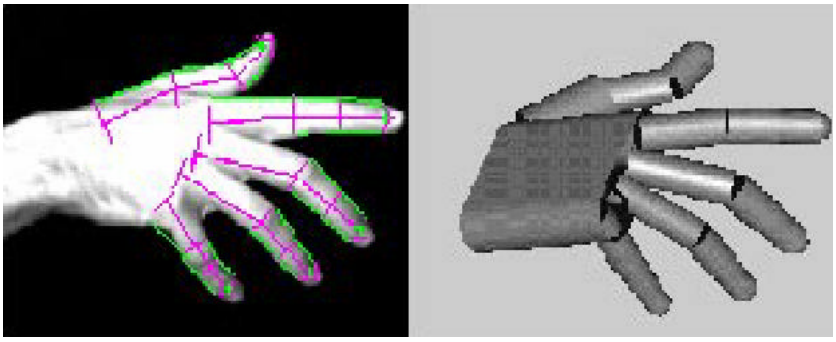

Figura 2. Resultados do trabalho de Regh & Kanade [30]. O esqueleto é sobreposto na imagem original da mão. No lado direito, o modelo 3D é calculado a partir dos parâmetros encontrados no esqueleto. Extraída de Regh & Kanade [30].

Nos sistemas baseados na aparência da imagem 2D, os parâmetros são extraídos diretamente das imagens. Parâmetros como a posição dos dedos na imagem, contornos e linhas, cor da mão, momentos e transformadas de espaço, entre muitos outros, são explorados.

No trabalho de Freeman [8] são descritas algumas técnicas simples para a interação através de visão. Um desses exemplos mostra como a orientação da mão, que é o conjunto de pixels diferentes do fundo, é utilizada para dirigir um robô (Figura 3.a); outro exemplo mostra como a diferença de imagens pode ser utilizada para inferir direção e movimento (Figura 3.b); e outro exemplo mostra como os chamados histogramas de orientação dos pixels são utilizados para reconhecer diferentes posturas da mão (Figura 3.c). No trabalho de Staner [36] é utilizada uma segmentação simples da cor da mão sobre fundo preto e são empregados modelos de Markov para reconhecer algumas posturas da mão. No trabalho de Sato [31] os parâmetros de uma câmera infravermelha são ajustados para capturar a mão; depois, na imagem gerada, são reconhecidas algumas posturas simples da mão.

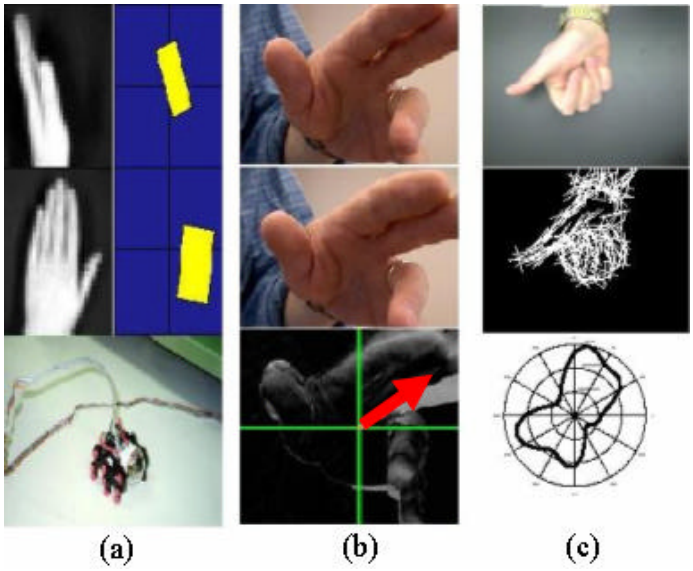

Figura 3 – Métodos simples para interagir através da Visão apresentados no trabalho de Freemam [8]. (a) Orientação da mão para dirigir um robô. (b) Subtração de imagens para inferir direção do movimento. (c) Histogramas de orientação.

Outros sistemas utilizam feições locais das mãos e dos dedos para reconhecer gestos mais complexos na interação. Existem diferentes abordagens, com destaque para as baseadas na silhueta. Tais sistemas utilizam diversos algoritmos de Visão para conseguir extrair a silhueta da mão. Nos trabalhos de Heap [12] (Figura 4.a) e Hall [10], a silhueta da mão é modelada como sendo uma curva e o reconhecimento é feito através de medidas de correlação (semelhança). No trabalho de Segen [33] é utilizado um ambiente controlado para a extração da silhueta; depois, utilizando algumas heurísticas, são reconhecidas algumas posturas da mão. Essas posturas são utilizadas para controlar um jogo e para navegar sobre um terreno (Figura 4.b).

Outras abordagens para o reconhecimento de gestos constituem os sistemas que utilizam imagens de treinamento para modelar os gestos a serem reconhecidos. O trabalho de Laptev [22] utiliza funções gaussianas (*blobs*) para representar o conjunto de imagens de treinamento que contém as posturas a serem reconhecidas (Figura 4.c). Dentro deste grupo podemos considerar também os trabalhos baseados em HMM (*Hidden Markov Models*), tais como os apresentados por Lee & Kim [23], Schlenzing [32], Vogler & Metaxas [42] e Wilson & Bobick [44]. No trabalho apresentado por Viveck [41] é utilizada a transformada de Fourier-Mellin junto com Redes Neurais para o reconhecimento de posturas da mão.

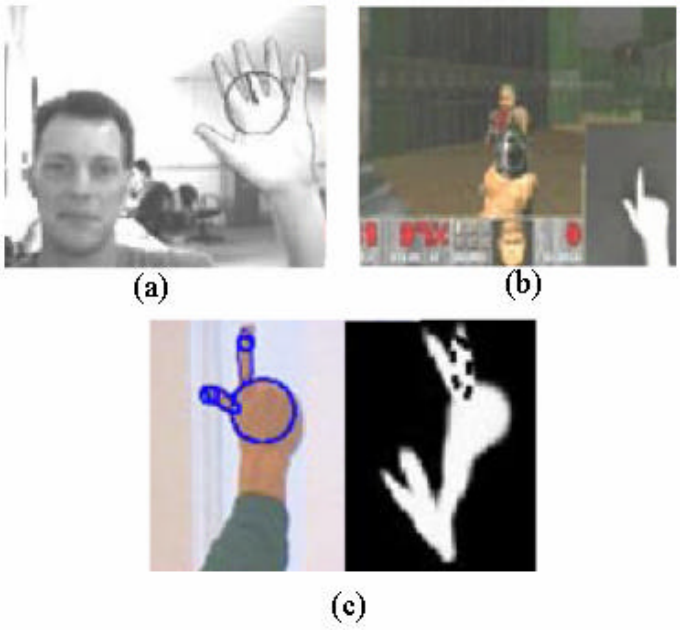

Figura 4 – (a) Extração do contorno da mão com o modelo de curva rígida. Extraído de Heap [12]. (b) Posturas da mão para controlar funções básicas de um jogo de computador. Extraído de Segen [33]. (c) Funções de Gauss (*blobs*) para detectar a mão e os dedos. Extraído de Laptev [22].

2.5.2. Sistemas Baseados na Detecção dos Dedos

Tradicionalmente, os sistemas baseados na detecção dos dedos são utilizados para construir interfaces que simulem funções básicas do mouse ou do teclado. Muitos desses sistemas podem ser considerados como sendo uma especialização dos sistemas de reconhecimento de gestos, pois a princípio eles utilizam as técnicas descritas na seção anterior, mas a detecção dos dedos é importante para a interação. Muitos desses sistemas, para poder reconhecer os dedos, utilizam diversas abordagens que exploram características da mão e os dedos, assim como das imagens.

No trabalho de Queck [29] é apresentado um sistema chamado "*FingerMouse*", que explora a cor da mão para segmentá-la do fundo; a ponta do dedo é sempre o ponto com maior coordenada *y*. Esse sistema é utilizado para movimentar o ponteiro do mouse com um dedo, e o evento clique do mouse é gerado ao se apertar uma tecla no teclado (Figura 5). Outro trabalho que explora a cor da mão é o de Kurata [21], que utiliza o dedo da mão para selecionar menus na imagem. No trabalho de Kulessa [19] é mostrado que é muito difícil construir um modelo de cor da mão que seja invariante a flutuações de

iluminação. Por isso, os sistemas que utilizam unicamente informações da cor da mão não têm bom desempenho e são muito limitados.

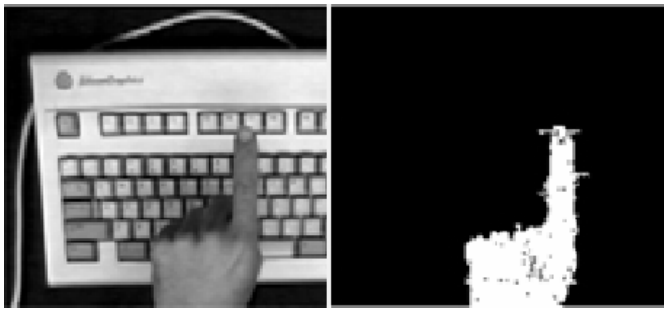

Figura 5 - Sistema “FingerMouse”. Extraído de Queck [29].

Muitos dos sistemas de detecção tentam explorar diferentes características da forma da mão e dos dedos. Para isto, estes utilizam medidas de correlação entre um padrão a ser reconhecido e os dados de entrada da imagem. Na maioria dos casos as imagens de entrada são imagens pré-processadas e a medida de similaridade (correlação) é feita nessas imagens. Devido à procura de um padrão específico na imagem, o número de gestos e a variação destes são muito limitados. No trabalho de Freeman [9] é apresentado um sistema de correlação que detecta uma mão estendida e fechada e utiliza a posição do dedo indicador para interagir com um programa que controla algumas funções básicas de uma televisão. A medida de correlação é feita numa área definida da imagem e a segmentação da mão é feita por cor. Crowley [4] desenvolve um sistema, com um plano de fundo especial, que explora as características da forma do dedo na correlação, permitindo desenhar linhas utilizando a ponta do dedo. No trabalho de Handenberg [11] é apresentado um sistema que utiliza a diferença de imagens na segmentação e então, na correlação, utiliza círculos para procurar a posição dos dedos. Alguns gestos como o clique padrão e o clique direito do mouse são implementados (Figura 6.a).

Outros sistemas facilitam a etapa de segmentação utilizando câmeras infravermelhas, o que permite uma segmentação mais acurada das mãos. No trabalho de Ukita [39] é mostrado um sistema que utiliza o dedo para desenhar linhas em que os dedos são modelados como sendo semicírculos, os quais são utilizados no processo de correlação. No trabalho de Oka [27] as pontas dos dedos são modeladas como sendo círculos no processo de correlação. O movimento dos dedos detectados é utilizado para reconhecer algumas figuras geométricas, as quais são utilizadas como comandos na interação (Figura 6.b).

Figura 6 – (a) Interação com o computador — posicionar a ponta do dedo num mesmo lugar por um determinado tempo ocasiona um clique. Extraído de Handenberg [11]. (b) Detecção dos dedos no trabalho de Oka [27]. Extraído de Oka [27].

Em vez de procurar diretamente pelas características dos dedos, alguns trabalhos procuram pelo contorno ou silhueta da mão e depois, na silhueta, procuram os pontos que representam as pontas dos dedos. Normalmente a modelagem do contorno é feita através de curvas *B-Splines* ou contornos ativos (*Snakes)* e o seguimento dos pontos de controle é feito através de algoritmos como *Kalman* ou *Condensation*. Nos trabalhos de Hall [10] e MacKormick [26] (Figura 7) são utilizadas essas técnicas para a detecção e o seguimento dos contornos, onde depois são detectados os dedos.

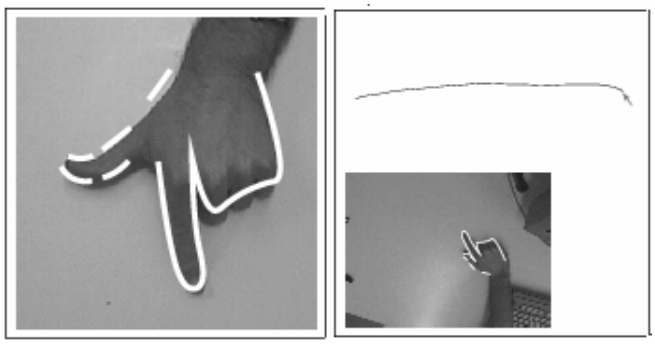

Figura 7 – Detecção do contorno da mão apresentado no trabalho de MacKormick [26]. A ponta do dedo detectada é utilizada para desenhar. Extraído de MacKormick [26].

2.6. Sistema de Interação Proposto

Muitos dos sistemas aqui expostos utilizam os gestos reconhecidos e os dedos detectados em aplicações e ambientes específicos. Muitas vezes esses ambientes e aplicações condicionam os processos de detecção dos dedos e reconhecimento de gestos. Diferentemente desses sistemas, no presente trabalho é proposto um sistema base para o reconhecimento de alguns gestos simples da mão que possam ser utilizados para reconhecer gestos mais

complexos em ambientes de trabalho convencionais. Este sistema se apresenta como sendo uma camada de baixo nível que garante o reconhecimento de um conjunto de gestos básicos para outras aplicações, as quais podem utilizar esses gestos de acordo as suas necessidades.

O sistema proposto neste trabalho possui as características dos sistemas baseados na detecção dos dedos. O objetivo principal é a detecção e o rastreamento dos dedos da mão. Com base nos dedos detectados e em suas informações (número, posição, direção, etc.) é reconhecido também um conjunto básico de gestos. Assim, na interação, tanto os gestos reconhecidos quanto as informações dos dedos são utilizados.

Para poder detectar os dedos da mão é empregada uma abordagem baseada na silhueta. Nesta abordagem é necessária a extração da silhueta da mão para detectar os dedos. A detecção da silhueta é baseada numa abordagem de segmentação que considera as características de cor e da iluminação do ambiente onde a mão está se movimentando. Nessa etapa de segmentação é introduzida uma abordagem de segmentação de fundo que, além de segmentar o objeto de interesse, procura diminuir as restrições do ambiente e a influência da iluminação na modelagem do fundo. Ao contrário de muitos trabalhos que se concentram exclusivamente nas características dos objetos para a sua segmentação e reconhecimento, neste trabalho são exploradas em conjunto as características do objeto e as características do ambiente.

No próximo capítulo é apresentada a etapa de segmentação desse sistema, na qual uma abordagem de subtração do fundo é utilizada para segmentar a mão do fundo da cena.