Universidade de São Paulo
Instituto de Física de São Carlos

Elias Ximenes do Prado Neto

Detecção de gestos manuais utilizando câmeras de profundidade

São Carlos

2014

Elias Ximenes do Prado Neto

Detecção de gestos manuais utilizando câmeras de profundidade

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de mestre em Ciências.

Área de Concentração: Física Aplicada Opção Computacional

Orientador: Prof. Dr. Odemir Matinez Bruno

Versão Corrigida
(versão original disponível na Unidade que aloja o Programa)

São Carlos

2014

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação do IFSC,
com os dados fornecidos pelo(a) autor(a)

Ximenes do Prado Neto, Elias
Detecção de gestos manuais utilizando câmeras de profundidade / Elias Ximenes do Prado Neto; orientador Odemir Martinez Bruno - versão corrigida -- São Carlos, 2014.
153 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Computacional) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2014.

1. Interface gestual. 2. Gestos manuais. 3. Kinect. 4. Câmera de profundidade. 5. Base de dados. I. Martinez Bruno, Odemir, orient. II. Título.

Dedico este trabalho ao meu filho Rafael Ferreira Prado, com grandes expectativas de que venha a contribuir com o seu desenvolvimento.

AGRADECIMENTOS

Ao Prof. Dr. Odemir Martinez Bruno, que nos anos de convivência, muito me ensinou, contribuindo para meu crescimento científico e intelectual.

Ao Instituto de Física de São Carlos, pela oportunidade de realização do curso de mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de mestrado e pelo apoio financeiro para a realização desta pesquisa.

À minha mãe Vera Lúcia Mascarenhas Leite, por sempre estar presente nas horas em que mais precisei.

Ao meu Pai Elias Ximenes do Prado Júnior, por todo aprendizado lógico, político e artístico e por sempre me incentivar a ser uma pessoa melhor.

À Fernanda Costa Ferreira, pelo seu companheirismo e compreensão diante de todas as dificuldades enfrentadas.

Ao meu irmão Danilo Mascarenhas Prado, pelo auxílio com os métodos do sistema descrito neste trabalho e especialmente pela nossa amizade incomensurável.

À minha irmã Danna Mascarenhas Prado, por seu carinho incondicional, por todas palavras de apoio e pela revisão dos textos desse trabalho.

Aos meus colegas e amigos Laurindo de Sousa Britto Neto, Hilário Seibel Júnior, Daniel Henriques Moreira e Felipe Leonel Grijalva Arévalo, pelo auxílio com os métodos de classificação e gravação da base de poses manuais.

Aos meus colegas e amigos Núbia Rosa da Silva, Maurício Falvo, João Batista Florindo, Dalcimar Casanova e Anderson Gonçalves Marco, por suas colaborações nas elaborações e testes dos métodos de extração de características, e pelos bons momentos de convivência durante essa jornada.

"E se o mundo não corresponde em todos os aspectos a nossos desejos, é culpa da ciência ou dos que querem impor seus desejos ao mundo?"

Carl Sagan

RESUMO

PRADO NETO, E. X. *Detecção de gestos manuais utilizando câmeras de profundidade*. 2014. 153 p. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2014.

É descrito o projeto de um sistema baseado em visão computacional, para o reconhecimento de poses manuais distintas, além da discriminação e rastreamento de seus membros. Entre os requisitos prioritários deste software estão a eficácia e a eficiência para essas tarefas, de forma a possibilitar o controle em tempo real de sistemas computacionais, por meio de gestos de mãos. Além desses fatores, a portabilidade para outros dispositivos e plataformas computacionais, e a possibilidade de extensão da quantidade de poses iniciais, também consiste em condições importantes para a sua funcionalidade. Essas características tendem a promover a popularização da interface proposta, possibilitando a sua aplicação para diversas finalidades e situações; contribuindo dessa forma para a difusão deste tipo de tecnologia e o desenvolvimento das áreas de interfaces gestuais e visão computacional. Vários métodos foram desenvolvidos e pesquisados com base na metodologia de extração de características, utilizando algoritmos de processamento de imagens, análise de vídeo, e visão computacional, além de softwares de aprendizado de máquina para classificação de imagens. Como dispositivo de captura, foi selecionada uma câmera de profundidade, visando obter informações auxiliares aos vários processos associados, reduzindo assim os custos computacionais inerentes e possibilitando a manipulação de sistemas eletrônicos em espaços virtuais tridimensionais. Por meio desse dispositivo, foram filmados alguns voluntários, realizando as poses manuais propostas, de forma a validar os algoritmos desenvolvidos e possibilitar o treinamento dos classificadores utilizados. Esse registro foi necessário, já que não foram encontradas bases de dados disponíveis contendo imagens com informações adequadas para os métodos pesquisados. Por fim, foi desenvolvido um conjunto de métodos capaz de atingir esses objetivos, através de sua combinação para adequação a diferentes dispositivos e tarefas, abrangendo assim todos os requisitos identificados inicialmente. Além do sistema implementado, a publicação da base de imagens de poses de mãos produzida também consiste em uma contribuição para as áreas do conhecimento associadas a este trabalho. Uma vez que as pesquisas realizadas indicam que esta base corresponde ao primeiro conjunto de dados disponibilizado, compatíveis com vários métodos de detecção de gestos manuais por visão computacional, acredita-se que esta venha

a auxiliar ao desenvolvimento de softwares com finalidades semelhantes, além possibilitar uma comparação adequada entre o desempenho desses, por meio de sua utilização.

PALAVRAS-CHAVE: Interface gestual. Gestos manuais. Kinect. Câmera de profundidade. Base de dados.

ABSTRACT

PRADO NETO, E. X. *Detection of hand gestures using depth cameras*. 2014. 153 p. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2014.

A project of a computer vision based system is described here, for the recognition of different kinds of hand poses, in addition to the discrimination and tracking of its members. Among the software requirements priority, were the efficiency and effectiveness in these tasks, in order to enable the real time control of computer systems by hand gestures. Besides these features, the portability to various devices and computational platforms, and the extension possibility of initial pose number, are also importants conditions for its functionality. Several methods have been developed and researched, based on the methodology of feature extraction, using image processing, video analysis, and computer vision algorithms; in addition to machine learning software for image classification. As capture device, was selected a depth camera, in order to obtain helper information to several associated processes, so reducing the computational costs involved, and enabling handling electronic systems in three-dimensional virtual spaces. Through this device, some volunteers were recorded, performing the proposed hand poses, in order to validate the developed algorithms and to allow the used classifiers training. This record was required, since available databases containing images with relevant information for researched methods was not found. Finally, were developed a set of methods able to achieve these goals, through its combination for adaptation to different devices and tasks, thus covering all requirements initially identified. Besides the developed system, the publication of the hand poses image database produced, is also an contribution to the field of knowledge related with this work. Since the researches carried out indicated that this database is the first set of available data, compatible with different computer vision detection methods for hand gestures, it's believed that this will assist in developing software with similar purposes, besides permit a proper comparison of the performances, by means of its use.

KEYWORDS: Gestural interface. Hand gestures. Kinect. Depth camera. Database.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Introdução

Desde a criação do primeiro sistema computacional até os dias atuais, a evolução dos métodos de interação entre pessoas e computadores está diretamente relacionada ao desenvolvimento científico e tecnológico. Enquanto novas interfaces nos auxiliam a pesquisar e desenvolver de forma mais rápida, segura e intuitiva; o progresso trazido por estas, permite a criação de procedimentos e ferramentas mais acessíveis a cada época, que possibilitam a criação e popularização de novas interfaces ainda melhores que as utilizadas anteriormente (1).

Ao observar que as gerações de computadores existentes, foram profundamente marcadas pelas diferentes interfaces de usuários criadas para suas utilizações, se torna notório que a popularização e o desenvolvimento dos computadores se devem, entre outros fatores, a sua progressiva facilidade de utilização, não apenas entre os profissionais ligados diretamente a área, como também para leigos em informática (2) e portadores de necessidades especiais (3, 4). Alguns exemplos mais promissores de interfaces de usuários para a manipulação de computares foram: controle direto dos circuitos, entrada e saída com cartões perfurados, linha de comando com teclado e monitor, e interface gráfica de janelas e mouse (5).

No entanto, acredita-se que com o avanço geral da capacidade de processamento, comunicação e amostragem de dados; os meios utilizados para a transmissão de informações por parte dos usuários possa ser um gargalo para o funcionamento de sistemas futuros (6). Além disso, uma vez que múltiplas formas de interação de computadores com seres humanos tem se constituído como necessárias a inúmeras atividades diárias, estas já são consideradas como indispensáveis ao modo de vida atual da sociedade como um todo. Esses fatos evidenciam a importância do desenvolvimento de novos paradigmas de interação entre pessoas e máquinas, como meio de aprimorar nossos métodos de comunicação, trabalho, aprendizagem e lazer.

Por essas razões, muitos estudos são realizados sobre as interfaces de usuário, sua influência na vida de seus utilizadores e os impactos que essas representam em nível social (7). Dessa forma, alguns novos métodos estão sendo estudados e utilizados para este fim, entre estes estão em destaque a interpretação de comando de voz (8), leitura de impulsos nervosos (9, 10) e reconhecimento de gestos (11, 12). Apesar das propostas para os futuros tipos de interface

se distinguirem quanto ao modo de entrada dos dados por parte dos usuários, acredita-se que as interfaces vindouras serão utilizadas através de múltiplos meios de comunicação, o que reduz a possibilidade de interpretações errôneas dos comandos por meio dessa redundância nas informações, e amplia a acessibilidade para leigos em informática, assim como para pessoas com deficiências físicas e mentais.

Afim de contribuir para a solução desses problemas, propõe-se neste trabalho uma interface gestual baseada em sensores de profundidade de baixo custo, capaz de obter dados acurados sobre movimentos de um conjunto abrangente de poses de mãos produzidos pelos usuários, de forma discreta, acessível, natural e com o capacidade de utilização em diversas plataformas computacionais. Uma vez que verificou-se a possibilidade de identificação eficaz das coordenadas das mãos, por meio de softwares auxiliares voltados para esse fim (13, 14), concluiu-se que seria possível identificar poses de mãos mais complexas do que as desejadas inicialmente, dispondo de menores custos computacionais inerentes. Dessa forma, a proposta inicial desse software sofreu várias reformulações de escopo, afim de contemplar um conjunto de funcionalidades que ultrapassassem o estado da arte dos métodos pesquisados.

A escolha dessa linha de pesquisa se deve ao fato de que o reconhecimento de gestos apresenta fortes sinais de já ser a forma de interface mais popular para a utilização de computadores, e está sendo cada vez mais implantado em dispositivos portáveis, consoles de jogos, televisões e plataformas computacionais diversificadas (3, 15, 16). Fora este fato, esta opção também considera o potencial dessa área em um futuro próximo, visto que durante o estudo deste tema ocorreram avanços tecnológicos significativos nos sensores ópticos populares, assim como a diversificação das interfaces sem toque (11, 17, 18) dos mais diversos tipos. Ademais, a utilização de gestos também é comumente introduzida para ampliar a acessibilidade a sistemas computacionais, já que muitas pessoas desprovidas de sua capacidade de comunicação oral, como: doentes, idosos e crianças, são capazes de realizar e compreender gestos simples (15).

Afim de implementar a interface proposta, muitos conhecimentos recém desenvolvidos e diversas técnicas clássicas de análise de vídeo (19–21) serão pesquisados, analisados e comparados. Além da literatura referente a interfaces usuário computador (1), também serão investigadas soluções que atendam ao problema proposto, nas áreas de visão computacional (22–24) e processamento de imagens (25).

Capítulo 2

Conceitos relacionados

Conforme visto no Capítulo 1, a utilização de interfaces gestuais já é comum em vários dispositivos eletrônicos. No entanto, esse conceito é sujeito a diversas interpretações, conforme o maquinário utilizado em questão.

Uma vez que o reconhecimento de gestos pode ser descrito como a análise dos movimentos de seus interlocutores em busca de um significado relativo ao contexto em que se encontram (26, 27), podem ser considerados como métodos mais simples de reconhecimento de gestos a detecção da movimentação de dedos em telas sensíveis ao toque e superfícies digitalizadoras, ou mesmo, movimentos mais elaborados produzidos por acelerômetros, *mouses*, *trackballs* e *joysticks*.

Já que são bastante diversificadas, esse capítulo é dedicado ao detalhamento e exemplificação dos tipos de interfaces gestuais existentes, com foco nas interfaces por visão computacional baseadas em gestos manuais.

Nas próximas seções esse tipo de interface é analisado quanto às tecnologias utilizadas para o seu funcionamento, as categorias de gestos em que se propõe a detectar, o tipo de informação que pretende computar, o grupo de membros do corpo propostos para análise, assim como uma visão geral das técnicas utilizadas para esse fim.

Além das várias características que diferenciam uma interface desse tipo, são apresentados resumos de métodos científicos capazes de reconhecer gestos manuais para fins variados com base exclusivamente no tipo de tecnologia utilizada.

2.1 Categorias de gestos

Uma das maiores diferenças entre as interfaces gestuais é a categoria de gestos em que está em seu escopo, já que elas podem se propor a identificar desde simples padrões de movimentos, realizados pelo usuário (28), até elaboradas linguagens de comunicação inteiramente baseadas em gestos (29, 30). A seguir são descritas algumas categorias de gestos frequentemente

utilizadas para a comunicação cotidiana (26, 27).

2.1.1 Emblemas

Expressam através de gestos simples, mensagens comuns e com significados amplamente conhecidos de forma independente da linguagem falada, como confirmações, negações e saudações, Figura 2.1. Este tipo de gesto, é também utilizado para transmitir mensagens previamente resumidas, o que o torna propício para ser amplamente utilizado na execução de comandos rotineiros em interfaces gestuais.

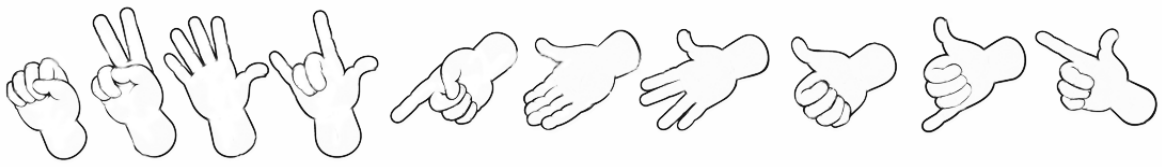


Figura 2.1 – Exemplos de emblemas
Fonte: Elaborada pelo autor

2.1.2 Gesticulações

As gesticulações consistem no tipo de gesto mais comum entre seres humanos, sendo geralmente expresso de forma espontânea e inconsciente. Costumam ser usadas inclusive em situações onde seria dispensável, como por exemplo ao falar ao telefone, ou sem que o interlocutor esteja vendo. Costumam ser aplicadas para enfatizar parte do discurso e coordenar a comunicação, podendo ser introduzidas em conjunto com interfaces baseadas na interpretação de linguagem falada, ou ainda como comandos para diferentes modos de interação, Figura 2.2 (31).

2.1.3 Pantomimas

Onde o usuário utiliza mímicas para narrar eventos e demonstrar elementos do mundo real, Figura 2.3 (32). Esta categoria de gestos é amplamente utilizada em comunicações interpessoais e é uma forma de arte cênicas de difícil execução, onde além de atenção, também

Figura 2.2 – Exemplos de gesticulações
Fonte: CASSANO, A (31)

é necessária relativa capacidade de interpretação e imaginação por parte de seus interlocutores.

Figura 2.3 – Exemplo de pantomimas
Fonte: CUSTÓDIO, L. N. et al. (32)

2.1.4 Idiomas gestuais

Comumente utilizados por pessoas desprovidas de sua capacidade de comunicação oral, muitas linguagens baseadas em gestos existem pelo mundo de forma independente da língua falada em cada pais, como a Língua Brasileira de Sinais (33), Língua Gestual Portuguesa (34), Língua de Sinais Americana (35), Língua de Sinais Japonesa (36) e Gestuno (37). Afim de ilustrar a sua variedade, alguns exemplos de sinais em LIBRAS podem ser visualizados na Figura 2.4 (38), e é referenciado um dicionário virtual dessa língua, disponibilizado pela Sociedade Acessibilidade Brasil (39).

Muitas pesquisas são realizadas para interpretação automatizadas dessas e de outras linguagens de sinais, onde a complexidade de análise dos sistemas propostos pode variar desde o simples reconhecimento de seu alfabeto datilográfico, até a utilização de todos os outros tipos de gestos vistos nesta seção (de acordo com o padrões estabelecidos para a língua em

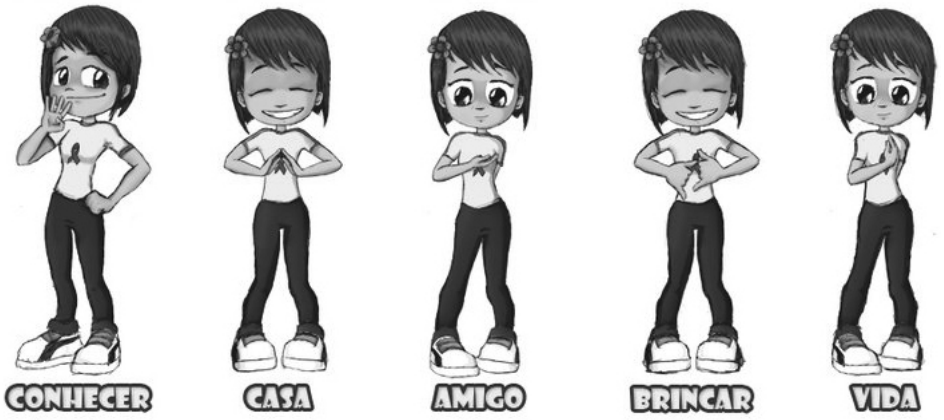


Figura 2.4 – Exemplos de sinais da linguá Libras
Fonte: RAID, J. (38)

questão), além da interpretação de construções linguísticas especialmente elaboradas para a comunicação gestual.

2.2 Informações a serem transmitidas

Dependendo da categoria de gestos escolhida, a interface gestual correspondente terá que detectar diferentes tipos de informações a partir dos gestos realizados por seus usuários. De acordo com a pesquisa divulgada no livro *The silent language* (40), 35% da comunicação humana é baseada em gestos utilizados para apontar para pessoas e coisas, expressar os nossos sentimentos e auxiliar a comunicação oral. No entanto os gestos podem ser utilizados para substituir completamente a comunicação oral, apresentando vantagens sobre esta para vários tipos de informações, como localização espacial, descrição de movimentos e formas e expressão de emoções. Segue abaixo uma lista de dados variados que podem ser extraídos a partir de gestos humanos (27).

Emoções: transmissão de sentimentos ou sensações do usuário para o sistema.

Localização espacial: local onde o gesto ocorre, direção que o usuário está indicando.

Variação no espaço: caminhos percorridos, ou indicados pelos membros do corpo de um usuário, responsáveis pela realização desses mesmos gestos.

Variação no tempo: dependendo da capacidade de discriminação da interface, os gestos podem ser divididos ainda em poses e gestos dinâmicos:

Poses: Expressões mapeadas para arranjos corporais estáticos como: poses assumidas pelas mãos, expressões faciais, e posturas corporais diversificadas. A sua análise é bastante utilizada para interpretar emblemas, localizações espaciais e emoções dos usuários.

Gestos dinâmicos: Sequências de movimentos predefinidos, com conjuntos poses representados seu inicio, meio e fim. Devem ser utilizados no intuito de reconhecer pantomimas, gesticulações e comunicações mais elaboradas.

2.3 Membros do corpo selecionados para interação

Interfaces gestuais para finalidades distintas, geralmente se propõem a interpretações com detalhamento adequado aos grupos de membros do corpo que participam ativamente do seu escopo de detecção de gestos. As interfaces costumam ser classificadas dessa forma quanto a três grupos principais: cabeça e membros da face, análise corporal completa, e gestos manuais. Segue uma descrição das informações que podem ser obtidas por cada um desses grupos e exemplos de utilizações em interfaces de usuário.

2.3.1 Cabeça e membros da face

Além de movimentos simples realizados pela cabeça do usuário, que podem expressar uma negação ou uma confirmação a uma solicitação realizada a ele, existem interfaces inteiramente desenvolvidas com base na detecção da direção dos olhos e de diferentes padrões de piscadelas para a execução de seus comandos (41, 42).

Este tipo de interface além de poder ser utilizada por pessoas com deficiências motoras mais severas, tem como vantagem respostas rápidas e precisas às solicitações dos usuários, onde apenas através do seu olhar é possível controlar um sistema de forma similar à utilização de um mouse em uma interface gráfica de janelas.

Sua maior desvantagem consiste em lidar com movimentos de olhos e pálpebras realizados involuntariamente, que podem incorrer na execução de comandos indesejáveis. Além disso, a acurácia na detecção da direção do olhar depende de métodos para rastreamento da face do usuário, ou de câmeras acopladas à sua cabeça, de forma a capturar os movimentos a distâncias e direções pré-determinadas, Figura 2.5 (43). Interfaces gestuais baseadas em movimentos

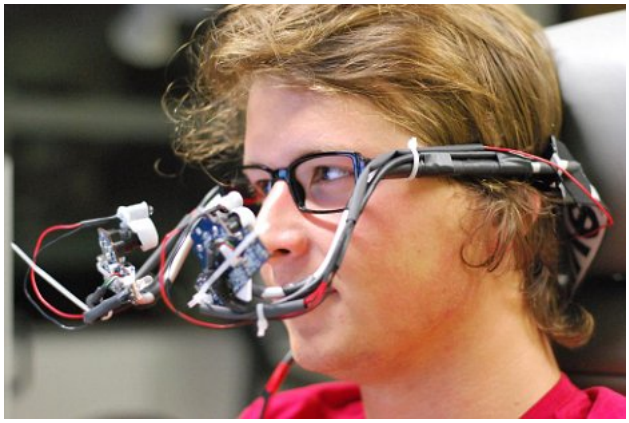


Figura 2.5 – Dispositivo para rastreamento da direção do olhar
Fonte: MCGUINNESS, R. (43)

da face humana podem analisar as formas assumidas pela boca, nariz e sobrancelhas de seus usuários para inferir seus sentimentos e sensações atuais (44), e para um controle personalizado de avatares em *chats* e jogos eletrônicos (45).

2.3.2 Análise corporal

As interfaces que se propõe a detecção de gestos corporais costumam abstrair movimentos menores, como os que são realizados pelas mãos e membros da face de seus usuários, para se concentram na detecção das posições e orientações do corpo como um todo, onde ele é comumente representado através poucas juntas e extremidades, Figura 2.6, para a representação completa de suas poses. Entre as aplicações que se utilizam comumente deste tipo de interfaces estão o rastreamento de pessoas (46), auxílio à reabilitação física (4), avaliação e treinamento simulado de atividades corporais (47), sistemas de realidade virtual (48) e jogos de ultima geração. Acredita-se que em um futuro próximo serão também populares sistemas de captura de movimentos mais completos, contendo informações sobre a face e as poses assumidas pelas mãos dos usuários, o que deve contribuir para a criação de jogos mais imersivos e interfaces virtuais de qualidade superior.

2.3.3 Gestos manuais

Como consequência de sua praticidade, naturalidade e generalidade, muitos estudos vêm sendo realizados para criar interfaces baseadas em reconhecimento de poses de mãos para

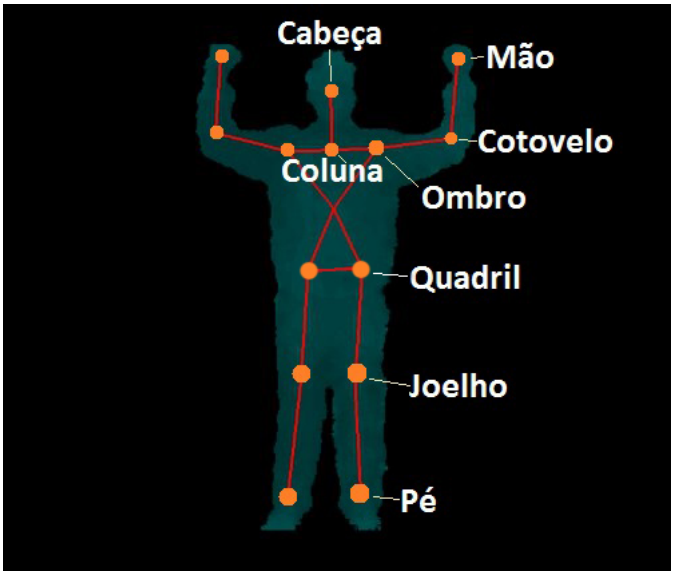


Figura 2.6 – Esqueletização da *OpenNI*
Fonte: Elaborada pelo autor

múltiplas funções, onde em apenas um artigo (6) foram enumeradas e classificadas mais de duzentos e cinquenta recentes publicações em grandes conferências e revistas. O aumento da atividade nesta área se deve tanto a popularização dos dispositivos descritos no Capítulo 3, quanto ao lançamento de sistemas para interação por gestos para plataformas de jogos populares (49–51), além da crescente demanda por aplicações que se beneficiem desta tecnologia; já que as interfaces por gestos manuais possibilitam a interação com computadores de forma atraente e intuitiva, e incentivam a utilização do sistema por usuários com pouca habilidade com interfaces de computadores convencionais, possibilitando uma manipulação direta através de suas próprias mãos. Nas palavras de Chaudhary et al.:

> *"Hand gestures recognition (HGR) is one of the main areas of research for the engineers, scientists and bioinformatics. HGR is the natural way of Human Machine interaction and today many researchers in the academia and industry are working on different application to make interactions more easy, natural and convenient without wearing any extra device"* (15).

Apesar da utilização de computadores através de gestos manuais ser simples de realizar e adequada a inúmeras aplicações, a sua interpretação automatizada inclui severas dificuldades, já que a mão humana é extremamente deformável e com capacidade se arranjar de formas complexas espacialmente (52, 53). As interfaces por gestos de mãos podem ser analisadas de forma semelhante à utilizada para a classificação de gestos em geral, no entanto entre as características mais analisadas para esses sistemas estão: os tipos de sensores utilizados para a captura dos dados, as dimensões espaciais analisadas, sua sensibilidade a variações

temporais, a metodologia utilizada para representação dos gestos, as modalidades de gestos e as aplicações propostas para a interface em questão (6, 15, 26).

2.4 Técnicas para classificação de gestos

Para que um gesto registrado seja corretamente classificado, geralmente são necessárias diversas etapas de processamento dos dados que lhes representa, onde suas informações são refinadas sucessivamente até que seja possível atribui-lo um rótulo específico. As diferentes abordagens e técnicas para este fim variam bastante quanto aos métodos utilizados, assim como a eficácia, o custo computacional e a capacidade de diferenciação de um número elevado de poses distintas.

2.4.1 Abordagens

Várias formas de representações e tipos de modelos foram propostos e implementados para o reconhecimento de gestos manuais por visão computacional, no entanto a divisão desses métodos em dois grandes grupos é largamente aceita e divulgada na literatura, sendo estes: a abordagem baseada no rastreamento de modelos geométricos, e a abordagem baseada em análise de aparência; ou como essa última também é chamada, abordagem baseada em extração de características (15, 26).

2.4.1.1 Abordagem baseada em Modelos

As técnicas baseadas em modelos tridimensionais geralmente se utilizam de ferramentas de desenho assistido por computador, onde modelos virtuais são criados ou reconstruídos para a representação das possíveis poses assumidas pelos usuários, afim de que sejam relacionadas às imagens com registro de poses reais, para estimar a posição e orientação tridimensional de suas partes constituintes. E, desta forma, classificar os seus gestos (6).

Essa abordagem é frequentemente aplicada para a detecção e reconhecimento de gestos em geral, e apresenta um funcionamento similar às técnicas para o reconhecimento de gestos que utilizam sensores portáveis, Seção 2.6.1, de forma que a posição e a orientação de cada

membro são obtidas pelo sistema, e a partir dessas são computadas as poses e gestos realizados pelos usuários. Apesar de utilizarem diferentes tipos de características das imagens para que os membros sejam rastreados, a maior parte do processamento desses métodos é concentrada nas simulações de possibilidades de poses e cálculo de suas correspondências para com as imagens fornecidas, de modo que essa abordagem tem como principal vantagem a capacidade de descrever e reproduzir naturalmente os gestos detectados, podendo ser aplicada diretamente em interações com objetos virtuais.

Com o propósito de processar e relacionar esses modelos com as imagens em tempo real, é comum a utilização de técnicas de programação de auto desempenho (54). Já que um dos problemas com este tipo de abordagem é o número de variações que uma aplicação é capaz de tratar, a geração de um número mais limitado de possibilidades nos modelos pode simplificar a solução a custo de uma menor generalidade de poses possíveis. Como os gestos manuais dos usuários apresentam muitos graus de liberdade, além das mãos serem deformáveis e articuladas, a extração de características destes pode ser complexa e se tornar um obstáculo somada aos outros problemas existentes no cálculo de posses possíveis (55).

Muitos tipos de modelos já foram propostos para a representação dos gestos das mãos, juntamente com suas respectivas formas de rastreamento e algoritmos para a extração de características. Os tipos de modelos utilizados variam desde elaboradas estruturas geométricas texturizadas e animadas, até simples relações de probabilidade entre a posição, o tamanho e a orientação de cada uma das partes que compõem as mãos dos usuários. Modelos mais robustos para representar a posição e orientação dos possíveis gestos, podem amenizar erros causados por ruídos e relacionamentos incorretos entre as simulações e as imagens. As estimativas absolutas obtidas para as posições individuais de cada uma das partes não costuma ser confiável para usos independentes, sendo comumente ocorridas interpretações com variações bruscas entre as poses geradas de forma correta e incorreta (6).

Modelos volumétricos, texturizados e deformáveis São os que suportam a maior quantidade de informação, sobre a forma, a aparência da pele e os movimentos das mãos, Figura 2.7 (56). Costumam ser mais restritos quanto a variações de poses específicas, onde muitas delas são geradas por interpolações lineares entre geometrias representando poses bases.

Modelos Geométricos Tridimensionais Carregam informações de menor precisão, em textura e aparência, do que os modelos texturizados flexíveis, no entanto podem ser mais generalistas quanto às orientações das partes que compõem a mão, uma vez que são formados apenas pela associação geométrica dessas partes, juntamente com os graus de liberdade

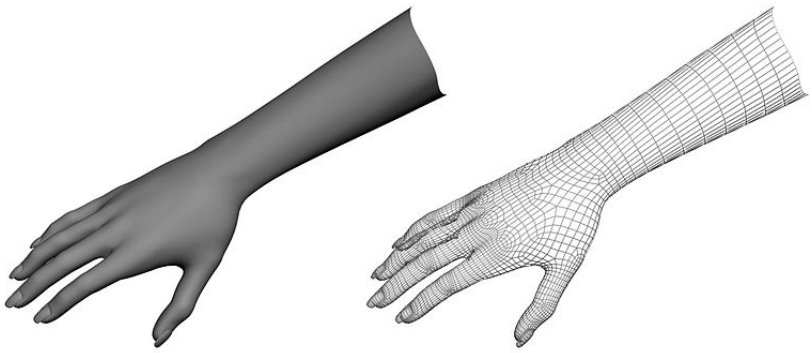


Figura 2.7 – Modelo deformável
Fonte: GESTURE RECOGNITION (56)

disponíveis para cada uma delas, Figura 2.8 (56).

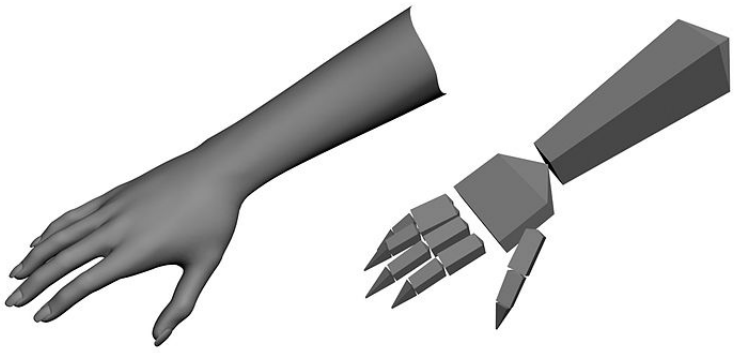


Figura 2.8 – Modelo articulado
Fonte: GESTURE RECOGNITION (56)

Modelos de Esqueleto Não suportam informações volumétricas bem definidas como os modelos geométricos tridimensionais e por isto podem ser ainda mais generalistas, quanto as possíveis poses assumidas pelas mãos, Figura 2.9 (57). Esses métodos apresentam similaridades com técnicas de extração de características, como a detecção de pontas de dedos e esqueletizações, onde estas características podem ser utilizadas para auxiliar as etapas de rastreamento.

2.4.1.2 Abordagem baseada em análise de aparência

A abordagem baseada em análise de aparência é comumente aplicada à detecção de gestos manuais e possui métodos heterogêneos, já que se utiliza das características mais

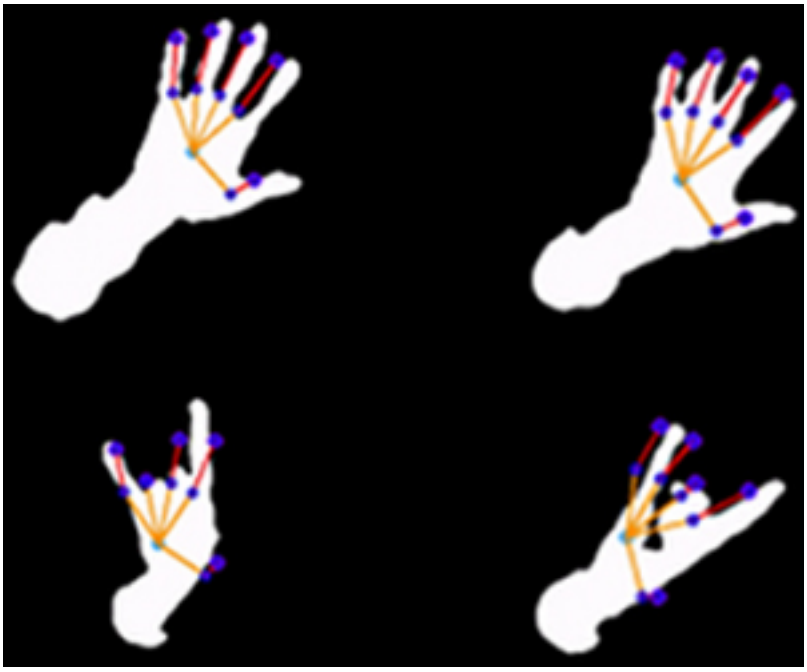


Figura 2.9 – Modelo de esqueleto
Fonte: HAND GESTURE RECOGNITION VIA MODEL FITTING IN ENERGY MINIMIZATION W/OPENCV (57)

diversas encontradas em imagens de tipos variados. Sua principal vantagem consiste em uma melhor adaptação a grupos de gestos que possam ser representados pelas características selecionadas para o seu funcionamento. Existem propostas para uma divisão desta abordagem em outras duas categorias, correspondendo à abordagem baseada em características de baixo nível e a abordagem baseada em aparência em geral (26), onde a primeira abrange apenas a utilização de características que não contém informações relacionadas à reconstrução da pose ou do movimento das mãos (29, 58), e a segunda explora características mais diretamente relacionadas a esses fatores incluindo casamento de formas (59), detecção das pontas dos dedos (60) e esqueletização das imagens de mãos (61).

Os métodos de reconhecimento de gestos baseados em aparência analisam múltiplas informações de imagens relativas a gestos de mãos, e então classificam estas conforme sua adequação a conjuntos de padrões de características registrados previamente. Dependendo do tipo de classificador aplicado para a detecção de gestos em tempo real, o cálculo da pertinência dos gestos a serem identificados, com base em grupos de gestos utilizados para aprendizagem do sistema, pode ser mais simples do que os algoritmos de rastreamento de modelos geométricos.

Muitas pesquisas tem explorado a velocidade e flexibilidade para detecção de gestos dessa abordagem em tarefas que não necessitam da orientação ou da posição das partes das mãos, mas sim de um conjunto abrangente de gestos que possa ser detectado rapidamente (17, 22).

Os tipos de representações de gestos baseados em aparência incluem modelos baseados em cores, análise de silhuetas, gabaritos deformáveis, esqueletizações de imagens, modelos baseados em movimento e quaisquer outras características diferentes do rastreamento de modelos geométricos que possam ser utilizadas para representar gestos manuais em imagens. Segue uma breve descrição de alguns tipos de características usualmente selecionadas.

1. **Cores:** Bastante abrangente utilizando diversos descritores texturais e técnicas de segmentação (55, 62).

2. **Silhueta:** Utiliza várias características da silhueta, como detecção do centroide, cálculo de perímetro, orientação e casamento entre formas (59).

3. **Movimento:** Se baseiam em técnicas de detecção de movimento como fluxo óptico e rastreamento de pontos de interesse (63).

4. **Identificação de membros:** Aplicam técnicas de morfologia matemática e geometria computacional para inferir as poses assumidas pelas mãos (64), ou a posição das pontas dos dedos (25, 60).

2.4.2 Etapas para o reconhecimento de gestos

Na busca por uma divisão das fases de execução que seja intrínseca às diversas interfaces gestuais existentes, são encontradas na literatura propostas distintas e com definições conflitantes entre si. No trabalho *Vision based hand gesture recognition for human computer interaction: a survey* (6) uma perspectiva voltada para os métodos que utilizam rastreamento de objetos geométricos é apresentada descrevendo as etapas de detecção, rastreamento e reconhecimento como inerentes para este fim. No entanto uma divisão diferente foi publicada no artigo *Survey on Gesture Recognition for Hand Image Postures*, onde é considerada uma sequencia de fases bastante comum em métodos que utilizam modelagem de gestos baseados em aparência, sendo estas: a segmentação das imagens, detecção e extração de características, e o reconhecimento de gestos, respectivamente (26). Na falta de uma divisão mais genérica, nesse trabalho serão considerados e apresentados métodos para o reconhecimento de gestos sobre três fases distintas, denominadas para este fim como seleção de Identificação de dados, extração de características e classificação de gestos.

2.4.2.1 Identificação de dados

A partir de imagens coloridas ou de profundidade, uma infinidade de informações podem ser adquiridas por meio de algoritmos de processamento de imagens para a aplicação de métodos de visão computacional. No entanto as diferentes interfaces sem toque baseadas em gestos manuais costumam ser guiadas por grupos reduzidos de tipos simples de dados, onde a sua obtenção consiste no primeiro passo para que os gestos realizados possam ser reconhecidos. Métodos comumente utilizados nessas fase incluem a segmentação das imagens de mãos (17), extração de contornos (59), localização de pontos de interesse (65), obtenção de descritores texturais (29), sacola de características (66), e extração de fluxo óptico (63).

2.4.2.2 Extração de características

Nesta etapa as características obtidas anteriormente são relacionadas de forma a agregar consistência e significado, possibilitando assim a sua utilização para o reconhecimento de gestos. Para este fim os sistemas podem realizar diversos tipos de aprendizados, como a identificação de regiões específicas, métricas em tempo real da estrutura da mão dos usuários, e adaptação a diferentes tons de pele. Muitos são os tipos de relacionamentos inferidos, dependendo da categoria de dados adquiridos, entre os mais comuns estão: análise de casco convexo (20), esqueletização de imagens (61), reconstrução tridimensional (11) e rastreamento com base em modelos geométricos (54); sendo o último caso diferenciado dos demais pela utilização de informações temporais, com base na sequência de imagens fornecida.

2.4.2.3 Classificação de gestos

Com base nas informações extraídas e correlacionadas os gestos realizados pelas mãos podem então ser classificados, os métodos utilizados para este fim são divididos em duas abordagens:

Regras preestabelecidas: Costumam ser métodos bastante simples, onde as informações obtidas são analisadas e a pose ou o tipo de gesto realizado é deduzido através de uma regra desenvolvida a priori. São bastante utilizados em interfaces de tempo real, uma vez que podem

ser baseados em regras de baixo custo computacional, como a análise dos ângulos formados pelos dedos, ou comparações entre modelos pré-computados. No entanto estes métodos tendem a ter uma acurácia limitada em comparação a sistemas elaborados de decodificação de regras.

Aprendizado de máquina: Bastante utilizados em trabalhos científicos, avaliam os gestos através de classificadores e outros processos estocásticos. Já que muitas desses métodos apresentam custos proibitivos para interfaces em tempo real, costumam ser gerados cálculos pré-computados na fase de aprendizado do sistema para a sua utilização na fase de reconhecimento.

Para o reconhecimento de gestos com base em aprendizado de máquina são utilizados desde as técnicas mais simples, como: métricas de distâncias euclidianas das características e distribuição de Gauss, até algoritmos ainda considerados como de baixo custo computacional, como lógica difusa, computação evolutiva e raciocínio probabilístico, além de técnicas de modelagens estatísticas, como: Análise de componentes principais e Máquina de estados finitos. Abordagens de mais alto nível incluem: Modelo ocultos de Markov, Filtro de Kalmam e as Redes neurais artificiais (15).

Definição: O aprendizado de máquina é um ramo da inteligência artificial, que tem como função básica construir e estudar sistemas computacionais capazes de aprender a partir de exemplos. Uma definição generalista desta área consiste em: "*Field of study that gives computers the ability to learn without being explicitly programmed*" (67). No entanto, os sistemas desenvolvidos por esse campo de pequisa podem ser descrito em termos operacionais como: "*A computer program is said to learn from experience* E *with respect to some class of tasks* T *and performance measure* P*, if its performance at tasks in* T*, as measured by* P*, improves with experience* E" (68).

Características: Um sistema de aprendizagem de máquina lida basicamente com representação e generalização dos dados utilizados como fonte de experiência. Deste modo, a capacidade de representar os dados fornecidos em instâncias utilizáveis, assim como o porte de um conjunto de funções para avaliação das mesmas, são parte inerente a qualquer método deste tipo. Por sua vez, a habilidade de generalização contribui para a correta avaliação de dados não utilizadas na fase de treinamento (69), evitando que o sistema cometa erros provenientes do fenômeno de sobreajuste. Os algoritmos de aprendizado de máquina costumam ser categorizados com base no método de classificação e tipo de treinamento utilizado:

Tipo de treinamento: O aprendizado de máquina pode ser norteado através de informações prévias acerca dos dados de treinamento, ou utilizado com a finalidade de diferenciar estes mesmos dados e encontrar relações entre eles:

Aprendizado supervisionado: Utiliza dados rotulados na fase de treinamento, de modo que a saída esperada para cada exemplo fornecido é conhecida previamente. Desta forma, esta abordagem é mais adequada para o treinamento de sistemas capazes de discriminar um conjunto de classes pré-determinado.

Aprendizado não supervisionado: Seu treinamento é baseado em dados não rotulados, sendo responsabilidade do método descobrir similaridades entre eles e agrupá-los de forma a expressar estas semelhanças adequadamente.

Aprendizagem semi-supervisionada: Consiste na combinação das abordagens anteriores onde são utilizadas informações rotuladas e não rotuladas para fins de treinamento. Esta abordagem pode ser utilizada quando se tem um conjunto significativo de dados não rotulados disponíveis e deseja-se orientar o resultado da classificação através de exemplos.

Método de classificação: Difere esse tipo de sistema quanto a sua estratégia de discriminação dos dados fornecidos:

Transdução: Classifica os dados utilizados na fase de testes com base em suas similaridades com uma ou mais instâncias dos dados fornecidos na fase de treinamento. Deste modo estes métodos não se utilizam de generalizações para produzir funções de classificação. No entanto, é necessário o armazenamento de um conjunto significativo de amostras para o seu funcionamento.

Indução: Considera os exemplos de treinamento como provenientes uma distribuição de probabilidade desconhecida, e através destes constrói um modelo capaz de generalizá-los, permitindo assim prever novos casos. Embora esta abordagem seja menos vulnerável ao sobreajuste e não necessite manter instâncias de treinamento em memória, ela tende a ser menos robusta, já que não é utilizável em conjuntos de testes que produzam previsões mutuamente inconsistentes.

2.5 Tecnologias envolvidas

Para o funcionamento de uma interface gestual qualquer é necessário a utilização de dispositivos especialmente desenvolvidos para este fim, como telas sensíveis ao toque (70), controles hápticos (71), acelerômetros (28), trajes com múltiplos tipos de sensores (72, 73) e câmeras digitais de tecnologias variadas (74–76). Além da detecção da posição e orientação de membros dos indivíduos, outras informações podem ser utilizadas para inferir os gestos a medida em que estes forem realizados, como análises estatísticas de relações entre as imagens capturadas com bancos de dados de características visuais (77) e padrões de movimentação de pontos de interesse detectados dinamicamente em cena (66).

Abordagens baseadas na detecção e rastreamento de membros do usuário podem incluir informações extras para as interfaces que às utilizam, como as respectivas posições e rotações durante o tempo, possibilitando a animação de avatares em ambientes de realidade virtual (48) e jogos eletrônicos; ou para controladores de dispositivos do mundo real (76) e simuladores virtuais de ferramentas. Por sua vez, as abordagens baseadas na utilização de outros tipos de características podem suprir a maior parte das necessidades de uma interface identificadora de comandos e serem utilizadas até mesmo para a interpretação de linguagens baseadas em gestos, a custos computacionais inferiores (29, 30).

Os diferentes tipos de tecnologia para a detecção de gestos apresentam grandes variações quanto ao seu custo, precisão, confiabilidade, latência e invasibilidade. Serão enumeradas a seguir algumas vantagens e desvantagens das abordagens utilizadas para este fim.

2.5.1 Interfaces portáveis e vestíveis

A interpretação de gestos humanos pode ser realizada por meio de vários tipos de dispositivos aplicados em seus corpos, com o intuito de rastrear o movimento de seus membros ou computar variações de distâncias entre eles. Esta categoria de detecção de gestos pode exigir certa adaptação de seus utilizadores e costuma necessitar de algum tempo para montagem e calibração dos equipamentos. A Figura 2.10 (78), exibe um dispositivo háptico para a manipulação de realidades virtuais.

Interfaces baseadas em dispositivos hápticos e rastreadores eletrônicos costumam apresentar maior precisão para o reconhecimento de poses e melhor tempo de resposta do que

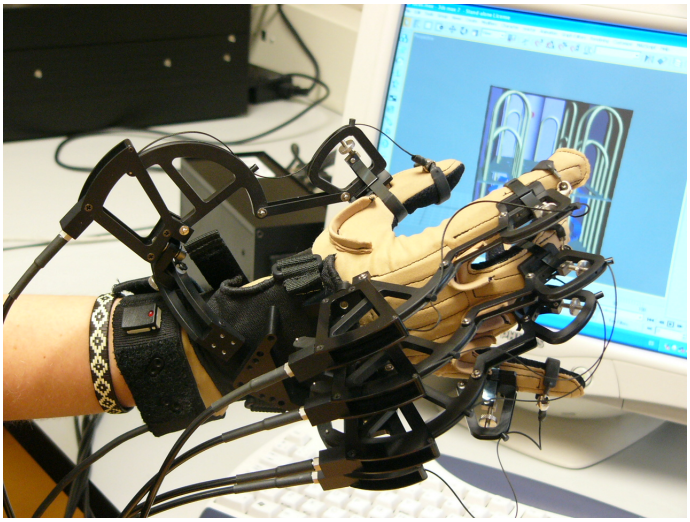


Figura 2.10 – Dispositivo háptico para ambientes de realidades virtuais
Fonte: GUAYARA, E. (78)

métodos baseados em visão computacional (79–81), no entanto essa abordagem pode ser demasiadamente invasiva para uma interface gestual de uso geral, pois é necessário que o usuário porte uma variedade sensores para a localização de uma quantidade razoável de membros de seu corpo (72, 73), prejudicando dessa forma a naturalidade da interação com o meio virtual. A situação é mais grave quando é necessária a conexão física destes instrumentos a computadores portáteis ou até mesmo plataformas fixas, o que pode ser bastante desconfortável e muito pouco prático, necessitando de algum tempo para sua preparação. Além disso, esses aparatos podem ter um peso significativo para alguns usuários, o que tende a restringir ainda mais a sua utilização.

2.5.2 Interfaces sem toque

Embora as interfaces multi toques já estejam presentes em um número considerável de dispositivos (70), um avanço singular está sendo obtido para a interação usuário computador por meio de novas técnicas e ferramentas de visão computacional, este progresso se deve em parte ao recente surgimento de inúmeras interfaces baseadas na detecção de movimentos humanos, obtidos através de câmeras de modelos variados (analisados na Seção 3.2.5 e na Seção 3.2.3) para diversos tipos de consoles e plataformas computacionais. As interfaces sem toque estão se popularizando rapidamente e já podem também ser encontradas em dispositivos móveis (81) e monitores de última geração (16).

Essa popularização se deve prioritariamente às vantagens de utilizar gestos ao ar livre para

se comunicar com computadores, em comparação a utilização de telas sensíveis ao toque, acelerômetros e dispositivos hápticos. Além da quantidade de informações que podem ser inseridas pela interpretação dos movimentos realizados por múltiplos membros do corpo humano simultaneamente, através deste tipo de interface é possível empregar gestos corriqueiramente utilizados para se comunicar com outros seres humanos, a fim de realizar tarefas computacionais correspondentes, e também é mais intuitivo para descrever movimentos realizados por seres humanos ou que possam ser simulados através de seus corpos (12).

Entre as muitas vantagens da abordagem por visão computacional em comparação com diferentes meios para detecção de gestos estão: baixa invasibilidade, baixo custo, alta escalabilidade, diferenciação de comandos realizados por usuários distintos e reconhecimento de gestos de qualquer parte do corpo dos usuários em três dimensões (15). A Figura 2.11 (82) ilustra a cena do filme de ficção científica. *Minority Report*, que é considerado como fonte de inspiração das interfaces sem toque.

Figura 2.11 – Cena do filme *Minority Report*
Fonte: SPIELBERG, S. (82)

Por consequência muitos estudos veem sendo realizados para a correta interpretação dos gestos de usuários utilizando câmeras digitais (58, 65, 74) e câmeras de profundidade (11, 83, 84). Entre as aplicações que estão em foco para se beneficiar deste tipo específico de interface se destacam: Aplicativos para desenvolvimento participativo, entrada de dados em dispositivos móveis (85, 86), modelagem e animação virtual de objetos e personagens tridimensionais (74, 87), navegação em ambiente virtual, e manipulações de maquinários em ambientes específicos, onde o contato com o usuário seja desaconselhável, quer seja para evitar a transmissão de agentes patológicos aos mesmos ou para minimizar possíveis danos físicos aos equipamentos

(88).

Apesar de ser menos invasiva e mais flexível do que as técnicas baseadas em dispositivos portáveis, a detecção de gestos baseadas em visão computacional apresentam problemas quanto à identificação e localização dos membros, e quanto à detecção de partes oclusas dos seus utilizadores (52). O uso de sensores de profundidade ou múltiplas câmeras dispostas estrategicamente sobre o ambiente podem amenizar estes problemas; fornecendo informações precisas das coordenadas tridimensionais de cada pixel, e limitando os problemas de oclusão à inferência de poses em áreas não capturadas por quaisquer câmeras utilizadas.

Outras desvantagens dessa abordagem são a baixa velocidade de captura de movimentos e o tempo necessário para o reconhecimento de gestos individuais ou sequências compostas por múltiplos membros, o que costumam ser especialmente prejudicial para sistemas de rastreamento de movimentos e interfaces para o controle de equipamentos do mundo real (27).

2.6 Exemplos de interfaces gestuais por dispositivo de captura

Além de diferenciar os equipamentos mais populares, esta seção apresenta alguns trabalhos de detecção de gestos manuais recentes, conforme o tipo de dispositivo de sensor empregado para este fim. A organização utilizada tem como finalidade: caracterizar o estado atual do reconhecimento de gestos manuais, exibir aplicações propícias para este tipo de tecnologia, e enfatizar os resultados obtidos com o auxílio de sensores específicos. Para cada tipo de dispositivo são especificados suas vantagens e desvantagens sobre os demais além do tipo de aplicação para o qual este é mais recomendável.

2.6.1 Sensores portáveis

Assim como as abordagens baseadas em contato físico para a detecção de gestos em geral, a detecção de gestos manuais pode se utilizar de muitos tipos de dispositivos portáveis. Entre os mais comumente aplicados estão os acelerômetros, luvas virtuais, leitores de impulsos nervosos musculares, além do rastreamento por dispositivos mecânicos ultrassônicos e magnéticos (6).

Em geral esse tipo de mecanismo tende a apresentar baixa resolução e falhas de localização espacial, além dos problemas de invasibilidade, levantados na Seção 2.5.1. No entanto a sua aplicabilidade a ambientes sem controle de iluminação, em conjunto com a sua precisão para o rastreamento de pequenos movimentos, velocidade de detecção de poses e indiferença a oclusões, justificam as sua utilização em diversos tipos de aplicações.

Modular Multi-finger Haptic Device: Mechanical Design, Controller and Applications: Entre os trabalhos mais recentes encontrados com dispositivos hápticos, este pode ser destacado pela sua originalidade e modularidade. É apresentado um dispositivo háptico, escalável para múltiplos dedos com base em uma configuração modular com graus liberdade redundantes, onde cada um de seus módulos pode ser utilizado por um dedo distinto de cada uma das mãos de seus usuários (89).

Além da descrição desse dispositivo é encontrado neste documento os requisitos para gerar implementações baseadas neste tipo de equipamento e exemplos de aplicações com altos requisitos de precisão, incluindo retroalimentação tátil para os usuários.

Decoding static and dynamic arm and hand gestures from the JPL BioSleeve: Baseado na leitura dos impulsos nervosos emitidos pelo braço de seus usuários, o artigo de Wolf et al. apresenta um método para classificar gestos realizados por braços e mãos, através de sensores de eletromiografia de superfície e uma unidade de medida inercial, embutidos em um equipamento para fixação no antebraço de seus utilizadores (79).

Apesar de se propor ao reconhecimento de uma diversidade de poses relativamente pequena, pela sua eficácia, eficiência e independência de localização (características das abordagens por sistemas embutidos), esse método é bastante indicado para a manipulação sistemas críticos, como: robôs e módulos aeroespaciais tripulados.

Using Postural Synergies to Animate a Low-Dimensional Hand Avatar in Haptic Simulation: Neste trabalho, é explanada uma técnica para animar uma marionete realista com 20 graus de liberdade com base em estudos de biomecânica da mão humana (80).

A abordagem proposta baseia-se no conhecimento de um conjunto de restrições cinemáticas sobre a mão, que permite representar suas posturas usando um número menor de variáveis. As poses da marionete são estimadas através um algoritmo de inversão cinemática, que leva em conta sinergias e estima a cinemática da mão como um todo.

Como pode ser visto nesse artigo a animação baseada na sinergia da marionete envolve

apenas cálculos algébricos simples, sendo adequado para implementação em tempo real, onde o utilizador tem um conhecimento prévio de como utilizar este tipo de dispositivo.

Gestural, Emergent and Expressive: Three Research Themes for Haptic Interaction: Apresentando novas perspectivas para área, este artigo, baseia-se em três estudos de caso nas áreas de design participativo, design de interação e artes eletrônicas (90), onde é realizada uma reflexão sobre suas implicações para as pesquisas em interfaces hápticas, além de uma discussão sobre os princípios de design que compõem este tipo de interface.

2.6.2 Abordagens baseadas em imagens

As abordagens baseadas na análise de imagens capturadas sequencialmente por um sensor, se caracterizam pelo aproveitamento de dados não disponíveis em métodos que utilizam sensores portáveis. Entre os muitos tipos de características que podem ser extraídas de diferentes tipos de imagens estão inclusas: diferenças de profundidade, análise de formas, variações de cores, reconhecimento de texturas, identificação de pontos de interesse e algoritmos elaborados para localização e rastreamento objetos (12, 15).

Já que nesses métodos os sensores se localizam a distâncias variáveis das mãos de seus utilizadores, a quantidade de informação relacionada ao reconhecimento do gesto irá variar de acordo com a sua localização em relação à profundidade do campo de visão das câmeras. A velocidade de captura dos dispositivos também pode ser um empecilho para a detecção de gestos em ritmos mais acelerados. Entre algumas limitações conhecidas nas tecnologias estudadas estão a sua sensibilidade a condições de iluminação variadas e problemas quanto ao tratamento de oclusões totais ou parciais dos gestos realizados (27).

Além desses fatores, inerentes a utilização de visão computacional, ainda existem muitos desafios para o reconhecimento de gestos e poses de mãos pela utilização de sensores luminosos. Uma grande variedade de imagens de gestos pode ser de difícil diferenciação e classificação, assim como a identificação da localização e orientação de todas as partes que compõem uma mão. Grandes variações entre a aparência das mãos de diferentes utilizadores e a correta interpretação do tempo de permanência e mudança entre poses, também podem ser problemas complexos, conforme o contexto de gestos a ser interpretado(6).

Uma vez que essas técnicas tendem a utilizar recursos computacionais consideráveis, os métodos para identificação e classificação de gestos geralmente são limitados, de forma a

Figura 2.12 – Luvas coloridas
Fonte: WANG, R. Y.; POPOVI¢, J. (91)

melhorar o seu desempenho em interfaces com grandes restrições quanto ao tempo de resposta. No entanto em interfaces onde a acurácia seja prioritária a adaptação dos usuários a utilização do sistema a velocidades mais baixas pode ser necessária. As abordagens de detecção de gestos baseados em imagens são divididas quanto ao tipo de equipamento utilizado para o seu funcionamento e os métodos de representação e classificação de gestos desenvolvidos (15).

Como equipamento para detecção de gestos por visão computacional, além de câmeras digitais coloridas é comum a utilização de câmeras de profundidade (Seção 3.2), luvas coloridas e marcadores. Em seguida são apresentados resumos de trabalhos recentes, utilizando os três tipos de tecnologias, incluindo abordagens mistas onde vários dispositivos podem ser utilizados para obter informações gestuais complementares, redundantes ou para utilizações ainda não previstas.

2.6.2.1 Luvas coloridas e sinalizadores de posição

Entre as técnicas para a detecção de gestos manuais é muito comum a utilização de luvas coloridas, Figura 2.12, ou sinalizadores nas pontas dos dedos, uma vez que é mais simples e rápido localizar pequenos objetos padronizados e variações de cores pré-selecionadas, do que diferenciar as múltiplas formas em que as mãos podem ser registradas. Outro artefato frequentemente aplicado para sistemas de visão computacional em geral são os marcadores em uma parte do ambiente, para a definição de sistemas de coordenadas equivalentes, ou para a calibração de câmeras.

Este tipo de tecnologia é bastante utilizado para a captura de movimentos de atores e

objetos, visando a animação de personagens para filmes ou jogos, já que é capaz de aumentar a precisão dos sistemas de forma pouco invasiva e a baixos custos em relação aos sensores portáveis. Os tipos de marcadores podem ser classificados como reflexivos ou luminosos, onde os reflexivos são compostos por pequenos objetos ou materiais mais simples, como pano, adesivos, pinturas, anéis e dedais, já os luminosos utilizam *leds* com padrões de cores variadas e pulsos luminosos coordenados.

Real-time classification of dynamic hand gestures from marker-based position data: Neste trabalho é proposto um sistema de reconhecimento de gestos manuais dinâmicos, com base na identificação de seus movimentos através de marcadores não rotulados (92). A principal contribuição dessa pesquisa é a possibilidade de desenvolvimento de sistemas de captura de movimentos mais práticos e simples já que o classificador desenvolvido é capaz de processar com precisão o posicionamento de partes componentes sem necessidade de correspondências entre marcadores conhecidos.

Basicamente uma luva com marcadores compostos por *leds* infravermelhos é utilizada, onde esses são aplicados em locais que apresentam deformações mais estáveis, como o dorso das mãos, por exemplo, a fim de descrever um padrão assimétrico preestabelecido. Esses marcadores são então rastreados, de forma a calcular suas posições aproximadas no espaço tridimensional e reconstruir os movimentos das mãos utilizando-se de relações espaciais conhecidas inerentes ao padrão aplicado.

O classificador utilizado é uma extensão de trabalhos anteriores para o reconhecimento de gestos manuais estáticos que servirão para formar a base de um vocabulário, já que ao serem combinados permitirão descrições precisas de vários gestos expressivos. Modelos ocultos de Markov e deformações dinâmicas estão sendo cogitadas para concretizar este objetivo.

AR pen and hand gestures: a new tool for pen drawings: Esta pesquisa, explora as possibilidades de interação utilizando a mão não dominante durante a criação de desenhos utilizando a mão dominante (93). Para tanto é idealizada uma ferramenta na forma de uma caneta, capaz criar e manipular realidades aumentadas interativas, ao registrar o ambiente e sobrepor desenhos físicos com imagens projetadas em tempo real.

Como prova de conceito foi construído um sistema composto por uma caneta com um pico projetor e uma câmera que é capaz de registrar marcadores reais e virtuais além de gestos simples de mãos e desenhos feitos no papel, possibilitando aos artistas controlar a imagens projetadas enquanto desenham.

Este dispositivo cria novas possibilidades de interação, onde através de marcadores e gestos da mão não dominante os usuários podem ajustar o tamanho ou a posição de um conteúdo, colorizar temporariamente imagens reais, selecionar e recortar imagens virtuais e compor ilustrações ainda na fase de rascunho de projeto, possibilitando uma validação precoce das estéticas mais adequadas.

User-defined gestures for augmented reality: Este artigo, promove um estudo de usabilidade, com base em gestos manuais para a manipulação de interfaces virtuais, onde apresenta um conjunto de gestos definidos com base em experimentos realizados com usuários, no intuito de orientar projetos de interfaces virtuais baseadas neste modo de interação (94).

Os autores anunciam ainda este trabalho como o primeiro realizado em conjunto com usuários para a definição de gestos utilizados na manipulação de realidades virtuais. A pesquisa foi feita sobre uma área de interação pré-definida, com um marcador em forma de imagem em seu centro. Cada participante foi então registrado interagindo com esta área por múltiplas câmeras coloridas, além de um sensor de profundidade *Asus Xtion Pro Live,* Figura 3.14c, para a reconstrução das mãos do usuário e do conteúdo virtual. Além do conteúdo destas câmeras foi gravado também o vídeo e o áudio transmitidos aos usuários durante a utilização da interface e o rastreamento do marcador central relativo à sua posição.

Apesar de não descreverem um método para a detecção dos gestos de mãos assumidas pelos entrevistados, esses são registrados por recursos distintos, como câmeras coloridas, câmeras de profundidade e marcadores, o que possibilita diversificadas abordagens futuras.

GestAction3D: A Platform For Studying Displacements And Deformations Of 3D Objects Using Hands: Não tão recente, mas não menos interessante é o trabalho de Lingrand at al. (87), onde é proposta uma interface de baixo custo para a interpretação dos gestos dos seus usuários, juntamente com dois softwares para a sua avaliação.

A primeira aplicação se dispõe ao estudo da edição de malhas tridimensionais com base em princípios de interação similares aos utilizados em softwares comerciais de modelagem e animação 3D (95–97). A segunda aplicação se propõe ao suporte a segmentação supervisionada de imagens médicas tridimensionais, com intervenções nas fases de inicialização, convergência e refinamento.

De acordo com os estudos realizados pelos atores, a melhor solução para prover mobilidade e naturalidade na utilização de gestos manuais a baixos custos, foi a utilização de luvas coloridas e câmeras estéreo. As luvas utilizadas contêm cinco *leds* de cores distintas para sinalizar a

localização de cada um dos dedos, onde a aquisição estéreo é feita utilizando duas câmeras digitais alinhadas, de modo a facilitar o cálculo de suas coordenadas tridimensionais.

Apesar da existência de outros dispositivos de manipulação tridimensional, tais como mouses 3D e luvas virtuais, os autores consideram a solução dada com maior capacidade de popularização, por fatores como custo e ergonomia. A interface apresentada é avaliada como intuitiva, já que recupera o movimento tridimensional produzido pelo usuário e reproduz diretamente no software, além de conter outros comandos para mudança de contexto. É relatado que o sistema pode ser utilizado com facilidade até mesmo por leigo e crianças.

2.6.2.2 Câmeras Coloridas

As câmeras que registram informações de luminosidade do ambiente são as mais comumente encontradas, sendo frequentemente embutidas nos dispositivos móveis e portáveis mais atuais, Seção 3.1, e por isso muito indicadas para a utilização em interfaces populares baseadas em visão computacional. Além da baixa invasibilidade em comparação à utilização de sensores portáveis e marcadores, estes sensores também podem ser utilizados sob iluminação natural sem maiores prejuízos quanto ao seu desempenho.

No entanto, além de serem bastante sensíveis a variações de iluminação e mudanças no ambiente registrado, a detecção de gestos costuma ser limitada a um pequeno número de poses, em comparação com outras técnicas, quando é utilizada apenas uma câmera para este fim. Conforme visto na Seção 3.2.1, a utilização de mais de uma câmera pode ser considerada como um sistema baseado em imagens de profundidade dependendo das técnicas aplicadas às imagens capturadas. A segmentação dos gestos é geralmente abordada com base em espaços de cores variados (24, 85), no entanto múltiplos métodos são utilizados com foco no escopo das interfaces propostas. Muitos estudos já foram realizados para a interpretação de gestos através deste tipo de câmera, onde são descritos a seguir alguns dos mais atuais encontrados na literatura.

How can human communicate with robot by hand gesture?: Este artigo relata a implementação de um software de reconhecimento de gestos em um robô pertencente a um museu, possibilitando-o obedecer a comandos transmitidos na forma de gestos e assim fornecer conteúdo interativo sobre o ambiente (58).

A proposta utiliza uma metodologia em dois passos, onde em princípio a imagem da mão

é identificada e segmentada, com base no tom de pele do usuário. Em seguida, usando um classificador em cascata *AdaBoost*, cada grupo de regiões encontrados são classificados como uma postura da mão.

É relatado que a divisão do método de reconhecimento em duas fazes reduziu significativamente o tempo computacional, em comparação com apenas o uso do classificador *AdaBoost*, sem que houvesse uma segmentação prévia. O robô também é integrado com sucesso e validado positivamente no contexto de sua interação como guia de museu.

Hand Gesture Tracking Based on Particle Filtering Aiming at Real-time Performance: *O m*étodo descrito se propõe a resolução de problemas de falha de rastreamento, causadas por deformações parciais das mãos, ou períodos prolongados de oclusão completa durante interações gestuais manuais em tempo real (53).

A solução oferecida é baseada em Filtros de partículas e *Mean-shift*, onde quando a mão está parcialmente obstruída ou movendo-se lentamente, uma otimização local com *Mean-shift* pode ser induzida para aumentar o número de partículas atuantes. É relatado que ao utilizar o *Mean-shift* das partículas como a sua fórmula de propagação, o custo inerente as essas simulações pelo filtro de partículas é removido, de modo que a velocidade e a eficácia do algoritmo de rastreamento foram melhoradas.

Os autores afirmam ainda que as simulações realizadas indicam que este método tem desempenho adequado para ser utilizado em tempo real e apresenta alta robustez quando utilizado em cenas complexas, em comparação com algoritmos tradicionais.

Instant 3D design concept generation and visualization by real-time hand gesture recognition: O sitema descrito neste artigo é voltado para a modelagem tridimensional de objetos e possibilita a criação e edição de formas através do reconhecimento dos gestos de mãos de seus utilizadores em tempo real, ou por tradução livre: é um sistema de conceitualização e visualização de formas através de gestos de mãos em três dimensões (74).

Neste trabalho é apresentado o arcabouço e os respectivos componentes do sistema, além dos menus e ícones inteligentes, onde as suas aparências e funções são alteradas com base em contextos de funcionamento pré-determinados.

O reconhecimento dos gestos foi realizado através de rastreamentos de modelos geométricos e modelos ocultos de Markov, onde esses foram projetados levando em consideração a sua facilidade de uso e aptidão para reconhecimento contínuo em tempo real.

Thai sign language translation using Scale Invariant Feature Transform and Hidden Markov Models: Exibindo grande versatilidade na detecção de poses, este trabalho, demonstra um sistema de tradução automática para a língua de sinais Thai (29).

Com fins em sua criação, foi gerada uma biblioteca de assinaturas a ser utilizada na classificação dos dados relativos aos gestos dos usuários. Estes descritores foram obtidos com base em registros realizados utilizando cinco indivíduos enquanto se comunicavam nesta linguagem, os registros foram realizados durante vários dias e em diferentes momentos do dia. Além desses, vinte outros indivíduos também foram registrados, apenas para fins de testes.

A biblioteca de assinaturas foi processada através de observações dispostas em descritores encontrados em pontos específicos dos vídeos, permitindo assim que fossem identificados em múltiplos quadros chaves. Foram utilizadas transformações de características livres de escala (*SIFT*), para a extração de descritores invariantes a iluminação e diferenças morfológicas entre as mãos. O Modelo Oculto de Markov foi utilizado então para traduzir as sequências de classificações coletadas em um conjunto de palavras.

O sistema obteve uma média de 86% a 95% de acerto, quando utilizado por indivíduos registrados previamente. Uma média de 76.56% de acertos foi obtida baseada em testes com múltiplos utilizadores e cenários auxiliados por parte dos usuários, onde estes vestiam camisas de manga comprida e exibiam as poses de suas mãos em frente a um fundo monocromático. O resultado para cenários sem restrições foi de 74% em média, sendo considerado bastante promissor para métodos baseados inteiramente neste tipo de tecnologia.

2.6.2.3 Câmeras de profundidade

Facilmente encontrados em novos televisores e interfaces computacionais mais avançadas (Seção 3.2) o reconhecimento de gestos manuais utilizando câmeras de profundidade já é uma realidade, e se apresenta como um forte impulsionador de pesquisas nesta área. Muitos trabalhos são encontrados com resultados animadores, onde a efetividade dos sistemas propostos tende a variar de acordo com a generalidade dos gestos em seu escopo.

As principais vantagens deste tipo de sensor são: baixa invasibilidade e alta praticidade, facilidade inerente de segmentação dos gestos capturados, controle de iluminação próprio e independente da luminosidade perceptível aos seus usuários, e a sua capacidade intrínseca para o registro de informações tridimensionais. Apesar dos benefícios supracitados, este tipo de sensor ainda é oferecido sob custos elevados para a sua popularização massiva, além de

ser extremamente vulnerável a iluminação natural, ou qualquer outra fonte de luz infravermelha. Em geral apresentam precisão e velocidade de captura inferiores, em comparação aos dispositivos portáveis e sistemas de captura de movimento por luvas coloridas e marcadores.

Deste modo, este tipo de tecnologia é indicado para interfaces gestuais populares, que sejam utilizadas em ambientes internos, onde se deseja uma maior eficácia na captura dos gestos, sem a necessidade de portar e vestir objetos especiais, ou de quaisquer outras formas de controle da cena.

Real Time Hand Pose Estimation Using Depth Sensors: No livro: *Consumer Depth Cameras for Computer Vision*, é descrito um algoritmo com base em uma abordagem por reconhecimento de objetos para a classificação das formas assumidas pelas mãos (54). Como uma aplicação do sistema, é utilizado um módulo de reconhecimento para dez gestos da língua de sinais americana, onde é obtida uma taxa de reconhecimento de 99.9% em tempo real. O processo é realizado através de um modelo de mão tridimensional realista, utilizado para representar mãos através de 21 partes distintas.

Através de animações desse modelo por meio de um gerador de imagens de profundidade sintéticas, uma árvore de decisão randômica é treinada, onde é classificada para cada pixel a pertinência a uma das partes das mãos. Os resultados da classificação alimentam então um algoritmo de busca local para estimar a localização das juntas dos esqueletos de cada uma das mãos. O sistema descrito pode processar imagens de profundidade capturadas pelo *Kinect* em tempo real independentes de suas variações temporais.

Os autores consideram que embora o reconhecimento de gestos manuais venha sendo um desafio para a visão computacional, a recente disseminação de modelos populares de câmeras de profundidade, detalhados na Seção 3.3, foi um grande impulsionador para as pesquisas envolvendo detecções de poses humanas. É apontada ainda a extração de parâmetros dos esqueleto das mãos, como uma importante tarefa para o reconhecimento de linguagens de sinais.

Real-time Hand Gesture Recognition from Depth Images Using Convex Shape Decomposition Method: Este artigo, apresenta um novo método para o reconhecimento de gestos em tempo real, realizados pelas duas mãos de seus usuários, a partir de imagens de profundidade (64). O método proposto agrega uma coleção de técnicas que possibilitam a detecção, segmentação e a classificação de gestos manuais, onde a detecção e localização das mãos são realizadas usando informações providas por um sensor de profundidade. Deste

modo as mãos são então segmentadas com robustez, mesmo em ambientes movimentados, sem a necessidade de nenhum tipo de marcador.

Utilizando as imagens segmentadas é aplicado um método de decomposição em formas convexas para a identificação das partes componentes das mãos presentes nas imagens. As palmas, as pontas dos dedos e os esqueletos das mãos são então reconhecidas com base na decomposição de suas formas além de outras técnicas auxiliares. Os experimentos realizados demonstram que um reconhecimento preciso de gestos pode ser obtido para aplicações em tempo real utilizando este método.

Real-Time 3D Hand Gesture Recognition from Depth Image: Nesse trabalho é proposto um novo algoritmo para o reconhecimento de gestos de mãos em três dimensões a partir de imagens de profundidade (11). Um sensor de profundidade *Kinect* para *Xbox* obtém imagens de profundidade em tempo real, onde essas são então segmentadas a partir de imagens dos usuários e posteriormente convertidas em nuvens de pontos tridimensionais. Através dessas nuvens de pontos são extraídas características invariantes a momentos tridimensionais para serem utilizadas na fase de reconhecimento. Um classificador baseado em máquinas de suporte vetoriais é então utilizado para rotular a forma das mãos em diferentes categorias. Além do método é criada uma base de gestos manuais, utilizada no desenvolvimento e na verificação do algoritmo proposto, através de resultados experimentais que comprovam a sua robustez.

Real-Time 3D Hand Gesture Recognition from Depth Image: Diferente das abordagens tradicionais, a pesquisa relatada no artigo *Robust Hand Gesture Recognition with Feature Selection and Hierarchical Temporal Self-Similarities,* propõe um método robusto para reconhecer gestos de mãos de acordo com a variação de suas características em múltiplas imagens (83).

Este trabalho utiliza várias características obtidas a partir de imagens RGB e de Profundidade adquiridas pelo *Kinect*, onde são selecionadas as que apresentarem maior robustez para o reconhecimento dos gestos, com base em um novo modelo de descritor proposto. É então utilizada uma matriz hierárquica de auto similaridades para detectar as transformações gestuais, de forma a aumentar ainda mais a robustez do sistema.

De acordo com os autores este método pode ser largamente utilizado para o reconhecimento de gestos onde a sua efetividade é comprovada através de experimentos realizados. É relatado ainda que o número de sistemas de identificação de gestos manuais baseados em

informações de profundidade adquiridas com a câmera de profundidade tem crescido bastante, e é apontado como um dos principais fatores a facilidade de segmentação das imagens de mãos a partir dessas informações, em comparação com a utilização de imagens coloridas para a realização da mesma tarefa.

Capítulo 3

Dispositivos ópticos para detecção de gestos

Tendo como objetivo deste trabalho a criação de uma interface sem toque (Seção 2.5.2) para a manipulação de sistemas computacionais, foram pesquisados diversos dispositivos para a obtenção em tempo real de informações relativas às poses das mãos dos usuários. Muitos destes foram descartados por serem: potencialmente danosos à saúde (98, 99), incompatíveis com a maioria das plataformas computacionais, com custos elevados para o usuário final, carentes de informações adequadas para o reconhecimento de gestos complexos em três dimensões (100), ou por oferecerem baixa qualidade nas informações adquiridas.

Entre muitas possíveis soluções, foram rejeitados pelos motivos supracitados, sensores baseados em lasers: de alta potência, equipamentos de ultrassom e escâneres corporais de alta definição.

Apesar de equipamentos munidos de lasers com amplitudes elevadas obterem imagens com menos ruídos, estes podem cegar irremediavelmente os seus usuários (101); os aparelhos de ultrassom fornecem imagens tridimensionais acuradas (102), mas oferecem riscos elevados se utilizados continuamente (103); por sua vez, os escâneres corporais apresentam riscos variados a saúde, conforme a tecnologia utilizada (104–106), apesar de serem desenvolvidos visando uma intervenção mínima. No entanto uma variedade de sensores ópticos aparece em destaque na literatura (65, 84, 86), onde verifica-se positivamente a sua aplicabilidade para interfaces sem toque. A Tabela 3.1 exibe uma comparação entre estes dispositivos, de acordo com as características desejáveis enumeradas para a interface proposta.

1. Popular: dispositivos encontrados facilmente, e já adquiridos por um número significativo de usuários;
2. Acessível: custos similares a outros periféricos utilizados em computadores pessoais;
3. Tempo real: captura de imagem com velocidade igual ou superior a 30 Hz;
4. Imagem de profundidade: imagens representando a distância para com o dispositivo através de seus *pixels*;

5. Utilizável sob iluminação infravermelha: pode ser utilizado sob incidência significativa de luz infravermelha, como irradiação direta de raios solares e emissões de dispositivos eletrônicos variados;

6. Utilizável sob altas variações de luz ambiente: não apresenta problemas quanto a oscilações da iluminação artificial utilizada, ou variações de luminosidade solar indireta;

7. Softwares gratuitos, eficazes para rastreamento e segmentação: contém softwares que podem ser utilizados sem custos financeiros, capazes separar facilmente elementos da cena registrada e obter acompanhar a sua localização a cada momento.

Tabela 3.1 – Comparação entre os dispositivos pesquisados

Categoria de sensor óptico	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Câmeras digitais monoculares	×	×	×		×		
Câmeras estéreo	×	×	×	×	×		
Escâneres de triangulação				×	×	×	
Câmeras de tempo de voo		×	×	×		×	×
Escâneres holográficos			×	×	×	×	
Câmeras de luz estruturada	×	×	×	×		×	×

Fonte: Elaborada pelo autor

Através das características assinaladas, é possível observar que as câmeras de luz estruturada correspondem a categoria de sensor mais indicada para a tarefa em questão. Segue uma descrição mais detalhada de todos esses tipos de dispositivos, além de diversas informações sobre o modelo selecionado; contendo um histórico resumido, uma lista de seus componentes eletrônicos, seu princípio de funcionamento, além dos motivos específicos que levarão a sua seleção para utilização neste projeto.

3.1 Câmeras digitais monoculares

Este é o tipo de sensor óptico mais comum na atualidade (107), sendo geralmente encontrado em computadores portáteis e dispositivos móveis, e podem variar largamente tanto quanto aos custos, quanto qualidade. O funcionamento básico de uma câmera digital consiste em utilizar um conjunto de lentes, para capturar os fótons refletidos pelo ambiente e então projetá-los sobre uma matriz de fotodetectores, os quais convertem então os fótons capturados a uma taxa de tempo variada, dependendo da quantidade de luz ambiente e do tipo de equipamento utilizado.

Ao final desse período, essa matriz de cargas acumuladas é lida e convertida em uma imagem. Fotodetectores feitos de silício, não são capazes de distinguir a luminosidade capturada por qualquer outro modo além de sua intensidade, sendo sensíveis também a iluminação infravermelha. Visando a aquisição de cores e filtragem de outras frequências luminosas, cada fotodetector de uma câmera digital é coberto por um filtro de cores individual.

Filtros do tipo *RGB** são comumente utilizados, para simular o padrão de cores absorvido pelos olhos humanos. Os filtros colocados sobre os *pixels* formam um padrão que é utilizado para a criação de uma imagem, onde geralmente é utilizada a interpolação via técnicas de processamento de sinais. Os filtros de cores Bayer† são bastante populares em câmeras digitais *RGB*, neste padrão os filtros para a cor verde são utilizados em uma maior quantidade dos que os de outras cores, visando simular a maior quantidade de percepção luminosa do olho humano para com essa cor, Figura 3.1.

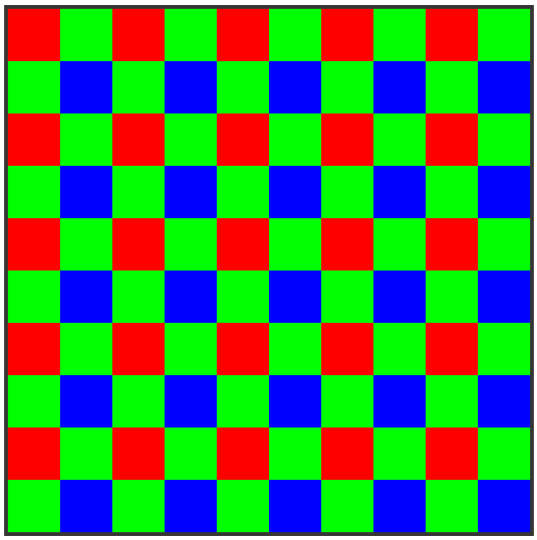


Figura 3.1 – Filtro de cor Bayer
Fonte: Elaborada pelo autor

As matrizes de fotodetectores são construídas em chips eletrônicos do tipo *CCD* (dispositivo de carga acoplada), ou do tipo *CMOS* (semicondutor de óxido metálico complementar). Em um dispositivo *CCD*, a carga é transportada ao longo do chip através de um canto da matriz de fotodetectores, onde então um conversor de analógico para digital registra a carga capturada em cada fotodetector. A Figura 3.2 ilustra como informação de cores se propaga através da matriz de fotossensores deste tipo, até a saída do sinal gerado para outros dispositivos. Nos dispositivos *CMOS*, vários transistores são colocados em cada fotodetector, esses aparelhos são utilizados para amplificam e mover as cargas capturadas pelos transistores

*Corresponde a utilização de filtros para as cores vermelha, verde e azul, e é o mais utilizado atualmente.

†Este é um filtro para as cores aditivas primárias, seu nome é dado em homenagem ao seu inventor Bryce Bayer.

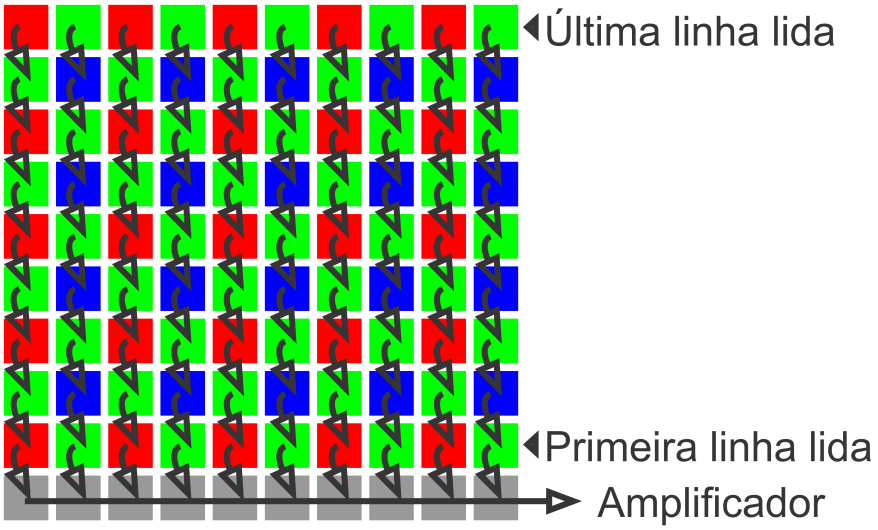


Figura 3.2 – Propagação de informações em dispositivos *CCD*
Fonte: Elaborada pelo autor

utilizando conectores eletrônicos. O *CMOS* permite deste modo que cada *pixel* seja lido individualmente, conforme pode ser visto na Figura 3.3 (108), onde *PD* é referente ao foto diodo (*Photo Diode*) utilizado e os demais sinais são representações dos circuítos digitais utilizados. Além do modo de transmissão de carga existem muitas diferenças entre esses dois tipos de tecnologia, os chips *CCD* eram mais utilizados até alguns anos atrás; atualmente os chips *CMOS* vem se popularizando bastante, e com isto se desenvolvendo tecnologicamente mais rapidamente. Apesar de existirem câmeras digitais monoculares com alta resolução, elevada taxa de

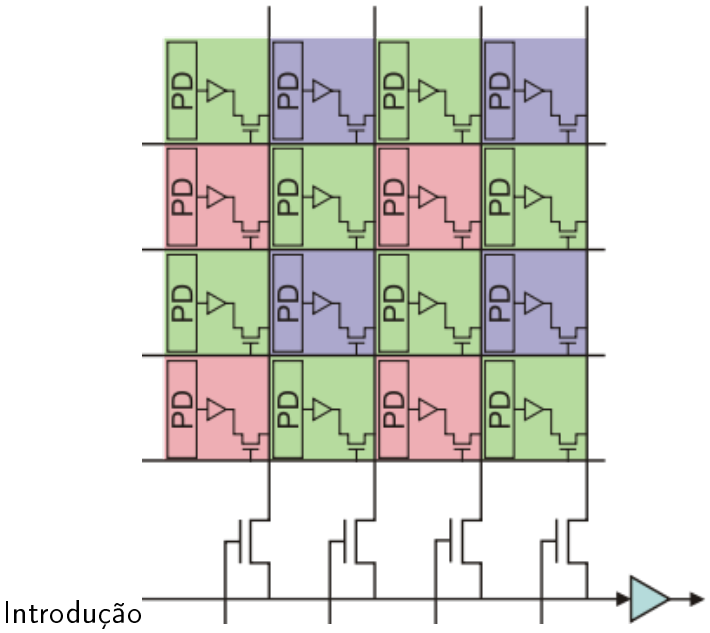


Figura 3.3 – Propagação de informações em dispositivos *CMOS*
Fonte: EXPEED. (108)

atualização, a preços acessíveis e largamente compatíveis com as plataformas computacionais

mais utilizadas, a análise de imagens por câmeras monoculares além de ser extremamente suscetível a iluminação externa e a diversos tipos de ilusões de óptica (109–112), e é ainda falha e onerosa para a extração de informações tridimensionais das cenas capturadas.

A falta desse tipo de informação dificulta a segmentação dos corpos dos usuários, e a correta localização dos gestos durante o seu movimento no espaço. Entre exemplos de ilusões de óptica comumente obtidos pela utilização de câmeras comuns podem ser enumeradas:

1. Estimativas errônea das dimensões de um objeto ou cena registrada, por meio da análise de corpos de tipos conhecidos, mas de tamanho incomum; ou por ambientes que simulam distorções de perspectiva, Figura 3.4a (113).

2. Indeterminações sobre a presença de oclusões e continuidades de formas, uma vez que dois corpos quaisquer podem ter formatos que se encaixem, ou um pode estar obstruindo parcialmente a visão do outro, Figura 3.4b (114).

3. Indeterminação sobre o sentido da rotação de objetos a maiores distâncias, já que a perspectiva é reduzida nessas situações, Figura 3.4c (115).

(a) Distorção de perspectiva
Fonte: AMES ROOM (113)

(b) Falsa continuidade
Fonte: ZIMMERMAN, C. (114)

(c) Rotação indeterminada
Fonte: SPINNING DANCER (115)

Figura 3.4 – Ilusões comuns em imagens coloridas individuais

3.2 Sensores de profundidade

Os sensores de profundidade compartilham várias características com as câmeras digitais comuns, assim como as câmeras digitais, eles têm um campo de visualização limitado e só

podem coletar informações sobre as superfícies que não estejam oclusas a partir de sua perspectiva atual. A diferença básica entre esses dois tipos de dispositivos é que: enquanto as câmeras digitais são utilizadas para coleta informações de cor sobre superfícies, que estejam dentro de seu campo de visão; um sensor de profundidade obtém um conjunto de distâncias até as superfícies que estejam sob sua perspectiva. A imagem produzida por um sensor de profundidade descreve assim esse conjunto de distâncias através de seus *pixels* (116). A Figura 3.5 (117) exibe a projeção no espaço tridimensional dos pixels de uma imagem de profundidade. De forma similar às câmeras de profundidade, um escâner tridimensional é um

Figura 3.5 – Exemplo de imagem de profundidade
Fonte: MCDONALD, K (117)

dispositivo que analisa objetos, ou ambientes reais, com o intuito de construir modelos digitais tridimensionais correspondentes. Já que a utilização de sensores de profundidade é bastante comum na digitalização de objetos tridimensionais (98, 118, 119), a definição dada para escâneres tridimensionais pode ser então extrapolada para as câmeras de profundidade, onde através das imagens de profundidade capturadas por esse dispositivo, cria-se uma nuvem de pontos representando a superfície de um objeto, que pode então ser utilizada para processar a sua forma através de métodos de reconstrução tridimensional.

Na maioria das vezes, uma única imagem de profundidade não será suficiente para produzir um modelo tridimensional completo, de modo que centenas de imagens de profundidade capturadas a partir de direções diferentes podem ser necessárias para obter informações sobre todos os lados do objeto ou cenário em questão. Ainda existe uma grande limitação entre os tipos de objeto que podem ser digitalizados. Os sensores ópticos apresentam problemas para a detecção de superfícies altamente reflexivas e objetos transparentes. Tecnologias variadas po-

dem ser utilizadas para construir dispositivos de digitalização tridimensionais, cada tecnologia vem com suas próprias limitações, vantagens e custos.

Os primeiros sensores de profundidade eram majoritariamente analógicos, esses dispositivos utilizavam fotodiodos de efeito lateral (*LEP*) e câmeras vidicon para converter informações ópticas em sinais eléctricos, que eram então processados para extrair informações de profundidade. Essas tecnologias eram difíceis de calibrar, por consequência de imprecisões e desvios eletrônicos. Durante a década de 1970 e início da década de 1980, o aumento da disponibilidade de dispositivos eletro-ópticos e a popularização dos computadores pessoais, tornou viável o desenvolvimento de sistemas de sensores de profundidade automatizados para aplicações industriais.

Os tipos de sensores profundidade mais comuns atualmente são as câmeras estereoscópicas, escâneres por triangulação laser, câmeras de tempo de voo e câmeras baseadas em luz estruturada. Os sensores de profundidade já são largamente utilizados para diversas aplicações industriais como, análise eletrônica, inspeção de alimentos e criação de modelos tridimensionais para museus, através do escaneamento de objetos e de patrimônios arquiteturais (98). Atualmente muitas pesquisas estão sendo realizadas para a detecção de gestos corporais (4, 12, 46, 120), e para reconhecimento de gestos de mãos e expressões faciais em tempo real por meio deste tipo de dispositivo (45, 84, 121, 122).

3.2.1 Câmeras estereoscópicas

Considera-se para fins deste trabalho como câmeras estereoscópicas, sistemas de visão computacionais que se baseiam na aquisição de pares de imagens, para posterior decodificação da informação tridimensional implicitamente capturada nelas (123, 124). Este tipo de sensor é comumente utilizado para capturar vídeos ou imagens estereoscópicas, que são atualmente populares em cinemas e monitores 3D.

Já que as imagens formadas em cada uma de suas câmeras, apresenta diferenças entre as posições dos objetos, Figura 3.6 (125), que variam de acordo com sua proximidade, é possível então deduzir a sua posição tridimensional com base nessa informação e na disposição das câmeras entre si. Esse processo é chamado de estereofotogrametria, ou seja, a medição de distâncias dos componentes de uma cena a partir de imagens estéreo.

Supondo duas câmeras de configurações idênticas, com seus eixos ópticos em paralelo e separadas perpendicularmente ao longo do eixo x por uma distância b, a posição de um ponto

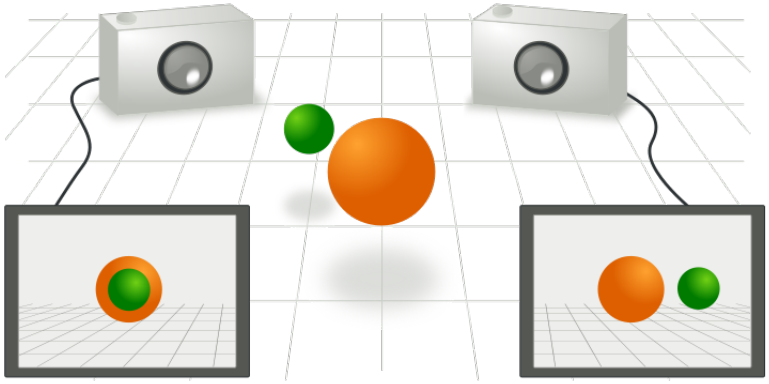


Figura 3.6 – Imagens da mesma cena obtidas por diferentes câmeras
Fonte: EPIPOLAR GEOMETRY (125)

(x, y, z) qualquer pode ser calculada em relação a origem O do sistema, com base na distância focal f das câmeras, e nas coordenadas de suas projeções (x'_l, y'_l) e (x'_r, y'_r), em cada imagem obtida.

Com base na Figura 3.7, é possível inferir por semelhança de triângulos que

$$\frac{x'_l}{f} = \frac{x + b/2}{z} \text{ e } \frac{x'_r}{f} = \frac{x - b/2}{z}$$

, e já que as câmeras se encontram sob a mesma altura

$$\frac{y'_l}{f} = \frac{y'_r}{f} = \frac{y}{z}$$

. Essas três equações podem então ser utilizadas para encontrar as coordenadas x, y e z, onde pela diferença entre as duas primeiras obtêm-se o valor da disparidade entre os pontos na forma:

$$x'_l - x'_r = f \cdot \frac{b}{z}$$

, o que leva as fórmulas

$$x = b\frac{(x'_l + x'_r)/2}{x_{'l} - x'_r},\ y = b\frac{(y'_l + y'_r)/2}{x'_l - x'_r},\ z = b\frac{f}{x'_l - x'_r}$$

, onde observa-se que a distância z é inversamente proporcional ao valor de disparidade. Visando a viabilização dessa tarefa duas informações devem ser então determinadas: um conjunto de pontos correspondentes nas imagens estéreo, e as propriedades geométricas das câmeras utilizadas. No entanto encontrar pontos correspondentes entre duas imagens não

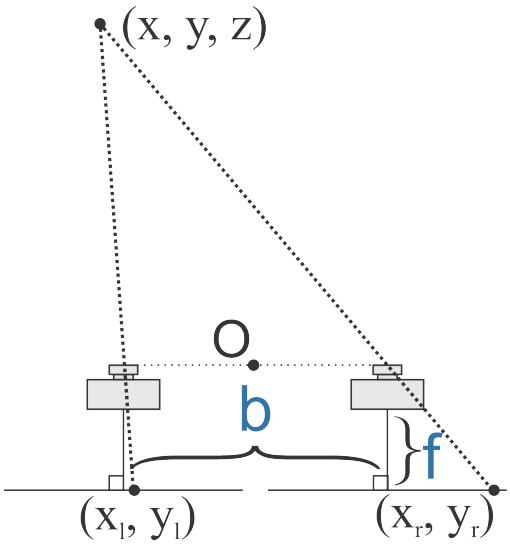


Figura 3.7 – Informações para o cálculo da posição tridimensional
Fonte: Elaborada pelo autor

é uma tarefa simples. Existem muitos fatores que podem inviabilizar esse processo, como oclusões, baixas resoluções de imagens, distorções e ruídos. Por esta razão, considera-se que as correspondências estão sob restrições, já que pode não haver informação suficiente nas imagens para garantir que o conjunto de correspondência encontrado seja único. Como consequência a detecção de pontos característicos em imagens é obtida através de técnicas de análise de imagem bidimensionais, onde procura-se por pontos contidos em áreas com características bem distinguíveis, como por exemplo linhas e cantos, que sejam resistentes a ruídos e distorções.

Por sua vez as propriedades das câmeras são determinadas pela sua geometria epipolar, que descreve o relacionamento entre os pontos observados no mundo, através de seus campos de visão, e as imagens que incidem sobre seus respectivos planos de detecção. A Figura 3.8 (125) representa duas câmeras, com seus respectivos pontos focais O_L e O_R, observando um ponto P. A projeção de P em cada um dos planos de imagem corresponde aos pontos p_L e p_R. Os pontos E_L e E_R são conhecido como epipolos, e correspondem a projeção de O_L e O_R nos planos de projeção das duas câmeras. Os pontos tridimensionais, obtidos por esta relação a partir de pontos de correspondência, são conhecidos como pontos correspondentes ou pontos combinados. Após o cálculo das coordenadas tridimensionais de um conjunto significativo de pontos da cena, recupera-se as informações das superfícies dos objetos através de processos de reconstrução, como visto anteriormente na Seção 3.2, isto pode ser feito através da triangulação dos pontos detectados, obtendo-se então superfícies formadas por malhas poligonais.

Por consequência de sua dependência para com a detecção de pontos de correspondên-

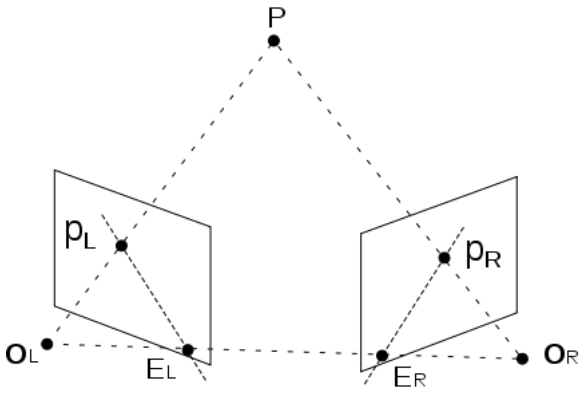


Figura 3.8 – Geometria epipolar
Fonte: EPIPOLAR GEOMETRY (125)

cia as câmeras estereoscópicas apresentam alto ruído e baixa precisão em suas imagens de profundidade, quando comparadas às outras câmeras de profundidade (126). Sua principal vantagem baseia-se em não utilizar qualquer tipo de iluminação especial para interpretar a cena, podendo funcionar em ambientes externos sob iluminação natural. Ainda que limitados a conjuntos de poses e movimentos simples, diversos trabalhos foram encontrados visando a detecção de gestos em tempo real com este tipo de dispositivo, onde as mãos dos usuários puderam ser localizadas com taxas de acerto significativas (20, 85, 127).

3.2.2 Escâneres de triangulação a laser

Desenvolvido inicialmente pelo Conselho Nacional de Pesquisa do Canadá em 1978 (128), a triangulação a laser criou os fundamento dos escâneres baseados em laser sincronizadas, que utilizam espelhos de varrimento para o seu direcionamento, Figura 3.9. A introdução das matrizes *CCD*, descritos anteriormente na seção 3.1, gerou um importante avanço para este tipo de sensor, onde a detecção da posição do laser emitido passou a ser limitado somente por seus limites físicos (129).

A performance de um escâner de triangulação a laser é analisada pelas leis da Física Óptica, onde em um sistema óptico ideal, sujeito a uma taxa de ruídos desprezível, devem ser levadas em conta apenas os limites de distorção e difração. Um cálculo aproximado da distância de uma emissão pode ser obtido por

$$z = \frac{d \cdot f}{p + f \cdot tan(\theta)}$$

, sendo f é a distancia focal da câmera utilizado, θ e o ângulo de emissão, d é a distância

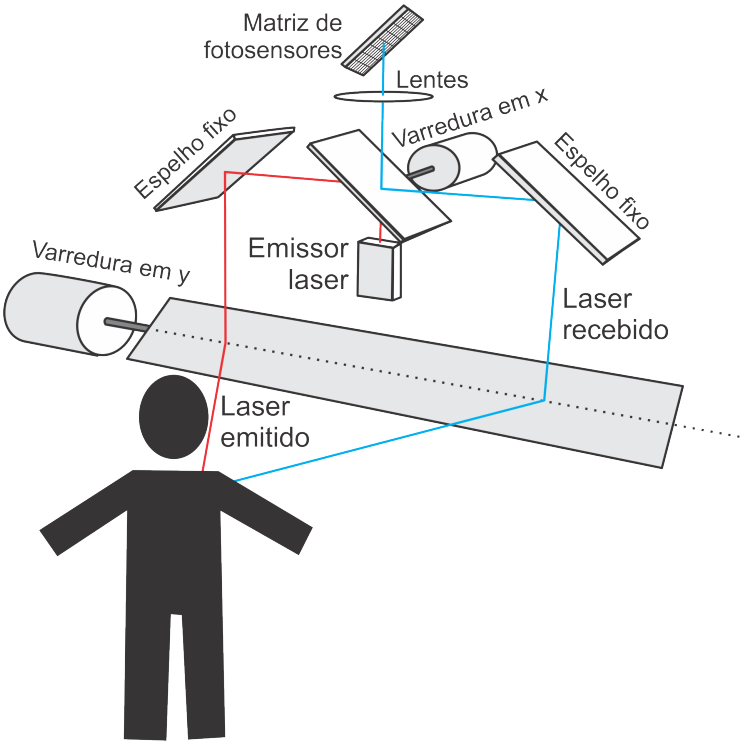


Figura 3.9 – Esquema de uma escâner de triangulação a laser
Fonte: Elaborada pelo autor

entre o plano focal e o emissor e p é a distância entre o eixo de projeção e o ponto luminoso registrado (130).

A precisão dos sistemas de triangulação é dado por

$$M_{3D} = \frac{\delta p}{\delta z} = \frac{f \cdot d}{z^2}$$

, onde δp é o pico/pixel de precisão da posição registrada, limitado pelo ruído de *subpixel* quando lasers são utilizados(129), como demostrado na equação

$$\delta p = \frac{1}{\sqrt{2\pi}} \cdot \lambda \cdot fn$$

. Usando a geometria convencional, o campo de visão do sensor é fornecido pela fórmula

$$\Phi = \tan^{-1}\left(\frac{P}{2 \cdot f}\right)$$

, dado P como a dimensão do *CCD* do dispositivo.

Os escâneres lineares são o tipo de escâner de triangulação mais utilizados, devido especialmente à sua simplicidade óptica e mecânica. Esta variação dos escâneres de triangulação é uma extensão natural dos detector de ponto, já que possibilita a detecção de um perfil completo em um única imagem capturada. No entanto esse tipo de escâner apresenta uma

capacidade inferior em sua relação com o campo de visão e resolução de profundidade, além de uma imunidade inferior a iluminação ambientes externos.

Apesar de sua elevada precisão e imunidade a iluminação ambiente, os escâneres de triangulação baseados em emissões lasers costumam apresentar uma taxa de captura inferior e custos elevados em comparação com outros sensores de profundidade, além disso as imagens de profundidades geradas contém ruídos característicos, que podem prejudicar a caracterização de superfícies suaves, como corpos de seres humanos. Por estas razões este tipo de dispositivo é pouco utilizado para a captura de movimentos, sendo mais comumente aplicado ao registro de objetos para aplicações industriais.

3.2.3 Câmeras de tempo de voo

Desenvolvidas a partir do princípio do tempo de voo[‡] (131), este tipo de sensor já é largamente utilizadas para reconstrução geométrica de objetos e reconhecimento de gestos (121, 132, 133). As câmeras de de tempo de voo (*TOF*) são uma ótima opção para registro de grandes estruturas, já que permitem a medições de intervalos mais longos com precisão relativamente constante durante todo o volume, no entanto, como estes sistemas se baseiam na detecção do tempo de voo da luz através do ar, as medições realizadas podem ser afetados por jatos de ar e variações de corrente nos dispositivos eletrônicos (98). Diferentes métodos são utilizados para este tipo de sensor como emissão de pulsos, modulação de amplitude, modulação de frequência, detecção híbrida, e auto mistura de diodos (134).

O método mais popular se baseia na emissão de pulsos de luz coerente e na respectiva medição do tempo em que esse levará para ser refletido até um detector luminoso, Figura 3.10, onde geralmente são utilizados este fim fotodiodos de avalanche. Uma captura com velocidade de picossegundos requer componentes eletrônicos muito sensíveis, com alta largura de banda, atrasos de grupo constantes, e excelente estabilidade térmica. Para a redução de ruídos, são utilizados múltiplos pulsos, onde então é calculada a média dos valores registrados.

Visando a ampliação do espectro de larga frequência dos pulsos emitidos, são utilizadas barramentos de alta frequência, onde uma largura de banda reduzida tende a proporciona uma melhor resolução de profundidade. Dependendo do equipamento utilizado, precisões submilimétricas podem ser registradas com este método. A detecção de cor, pode ser obtida

[‡] O princípio de tempo de voo possibilita o cálculo da distância entre o dispositivo de medição e o seu alvo, pelo tempo que um raio luminoso emitido leva para incidir sobre o objeto e retornar até um sensor de luminosidade

utilizado-se lasers com a cor branca e um sistema óptico de separação de cores (116).

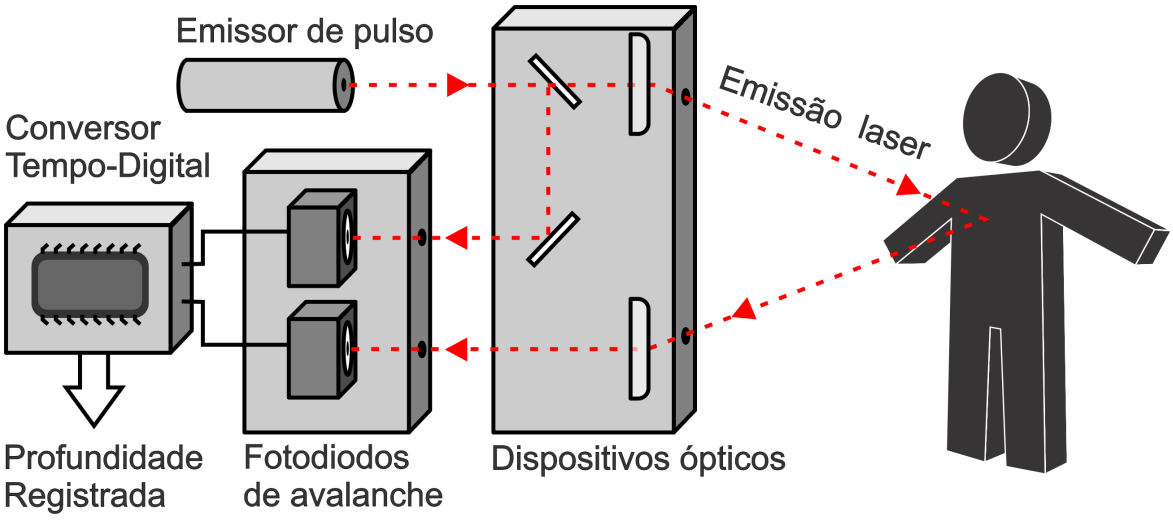


Figura 3.10 – Esquema de uma câmera de tempo de voo.
Fonte: Elaborada pelo autor

A profundidade registrada neste desse tipo de dispositivo pode ser calculada pela equação

$$R = 0.5 \cdot c \cdot T$$

, onde a resolução de profundidade é obtida através de sua derivada, descrita por

$$\partial R = 0.5 \cdot c \cdot \partial T$$

sendo c a velocidade da luz, ou seja $(3 \cdot 10^8 \text{m/s})$, logo para obter-se uma precisão de profundidade de por exemplo 1,0 cm, é exigida uma resolução temporal de 66 ps ou uma largura de banda equivalente a 15 GHz. A precisão na medição está diretamente relacionada com a amplitude do sinal de retorno. A modulação AM é utilizada para medir a fase relativa do sinal de retorno e a modulação FM é utilizada para medir a frequência entre o sinal e sua referência, a frequência do sinal está diretamente relacionada ao alcance máximo do sensor, que pode ser obtido por desses sensores.

As principais vantagens deste tipo de dispositivo sobre as outras câmera de profundidade, são as suas altas taxas de atualização de até 100 quadros por segundo, alcances de detecção de objetos de até 60 metros, apresentarem uma certa resistência a iluminação natural, além da capacidade de gerar matrizes de profundidade sem interpolação e sem necessidade de utilização de algoritmos complexos (135).

O maior empecilho para a popularização desse tipo de sensor para interfaces de usuário costumava ser os valores elevados que esses eram ofertados. No entanto, com o impacto de mercado produzido pela popularização das câmeras baseadas em luz estruturada, muitas

empresas vem oferecendo câmeras de voo de baixo custo (136–139) (140), com bibliotecas de função (141) e *frameworks* para a sua utilização.

3.2.4 Escâneres holográficos

Escâneres baseados em interferometria, também conhecidos como escâneres holográficos, são essencialmente imagens holográficas de polígonos em rotação. Geralmente os escâneres holográficos tem a forma de um disco, que é rotacionado em torno do seu eixo, onde as grades holográficas codificados nos vários setores desse disco representam visões em múltiplos ângulos do holograma registrado, Figura 3.11.

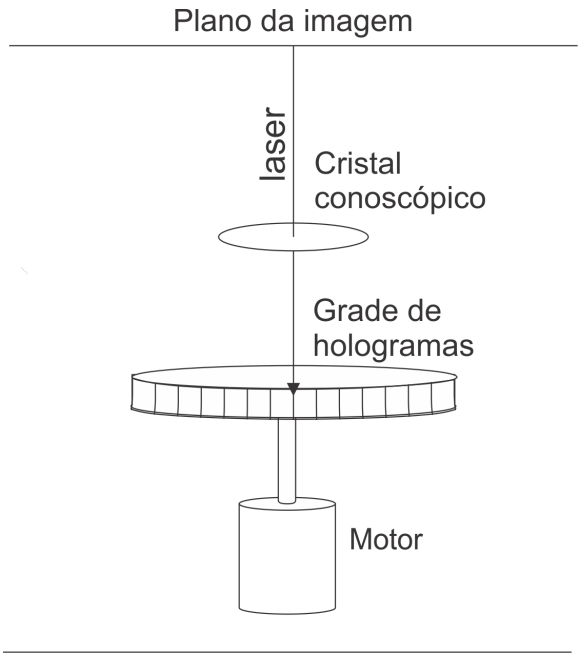


Figura 3.11 – Esquema básico de um escâner holográfico
Fonte: Elaborada pelo autor

Enquanto na holografia clássica, os padrões de interferência de luz são formados através de uma emissão laser refletida a partir de um objeto e uma emissão laser de referência, na holografia conoscópica, uma única emissão laser é divida e projetada sobre uma superfície, em seguida as reflexões imediatas ao longo dessa mesma emissão são filtradas através de um cristal conoscópico, os feixes luminosos divididos se propagam com a mesma velocidade, mas seguem caminhos diferentes e por isso produzem padrões de interferência, devido à diferença de fase de frequência óptica relacionada com o caminho percorrido, o resultado é um padrão de difração, que é analisado para determinar a distância do dispositivo até uma superfície (142).

A principal vantagem da holografia conoscópica é que é necessário apenas um único raio para a medição de uma superfície, sendo possível desse modo medir a profundidade de orifícios e regiões com grandes concavidades. Entre as aplicações que utilizam este tipo de sensor destacam-se a microscopia, reconhecimento de objetos, televisores holográficos, criptografia e sensoriamento remoto (143–146). No entanto apesar de seus benefícios, não foram encontrados na literatura quaisquer trabalhos relacionados ao reconhecimento de gestos com escâneres holográficos, onde os valores e dimensões dos dispositivos pesquisados, no decorrer deste trabalho, se mostraram incompatíveis com o tipo de aplicação proposta.

3.2.5 Câmeras de luz estruturada

O funcionamento das câmeras de luz estruturada, baseiam-se na projeção de padrões luminosos sobre a cena (98, 129), Figura 3.12 (147). Os padrões projetados variam conforme a abordagem utilizada, assim como as características dos dispositivos utilizados para este fim, onde algumas abordagens utilizam múltiplas câmeras para detectar a correlação entre os padrões gerados (123, 148), projeções sequenciais de padrões de codificados (149), interferências de fases de padrões marginais (150) e sistemas ópticos capazes de criar distorções nos padrões de acordo a distância do dispositivo (151). Assumindo lentes perfeitas o critério de Rayleigh é

Figura 3.12 – Ambiente sob a luz estruturada do *Kinect* para *XBOX*
Fonte: LIND, J. (147)

utilizado para calcular o ganho em sistemas baseados em projeção de padrões (98), conforme

$$q = 1.22 \cdot \lambda \cdot fn$$

. O critério de Raleigh indica quão bem uma imagem pode ser computada. Por exemplo, com um comprimento de onda de $\lambda = 680$ nm, e uma lente $fn = 4$, obtém-se uma resolução de profundidade de $\delta p = 1.4\mu m$, o que demonstra que os escâneres baseados em luz coerente são mais precisos a uma mesma configuração óptica.

Os sensores baseados em emissões de padrões luminosos são muito populares por causa da disponibilidade de projetores de baixo custo, além da possibilidade de digitalização de volumes tridimensionais completos em um única imagem. Diferentes modelos de câmeras de luz estruturada, desenvolvidas especialmente para a detecção de gestos em tempo real, foram recentemente difundidas (152, 153) para utilização em jogos e sistemas computacionais. Embora a precisão de profundidade deste tipo de dispositivo seja comparativamente menor do que a de escâneres de fenda, ou escâneres de lasers pontuais, o uso de luz incoerente remove o ruído associado, resultando em dados mais suaves. Este tipo de sensor é bastante utilizado para registrar corpos humanos em movimento, já que para este fim, a precisão das imagens de profundidade não é tão importante quanto o registro em tempo real de superfícies suaves.

As principais desvantagens destes sensores em comparação com outras câmeras de profundidade consistem em: não funcionar sob forte iluminação natural[§], e terem um alcance funcional limitado a poucos metros. No entanto este tipo de câmera foi eleito para este trabalho, por conta de sua recente popularização no cenário mundial, e da vasta gama de informações para a manipulação de interfaces fornecidas por algumas de suas bibliotecas de funções; além da possibilidade de utilização de diversos dispositivos, como os sensores desenvolvidos pela *Microsoft,* Figura 3.13 (154), pela *ASUS*, Figura 3.14 (155) e pela *PrimeSense*, Figura 3.15 (156).

(a) *Kinect* para *XBOX* **(b)** *Kinect* para *Windows*

Figura 3.13 – Câmeras de luz estruturada da *Microsoft*
Fonte: CARTWRIGHT, J. (154)

[§]A luz infravermelha emitida pelo sol se sobrepõe sobre a luz estruturada de baixa potência projetada por este dispositivo, inviabilizando o seu registro.

(a) XTion **(b)** *XTion PRO* **(c)** *Xtion PRO LIVE*

Figura 3.14 – Câmeras de luz estruturada da *ASUS*
Fonte: ASUS MULTIMEDIA (155)

(a) *Carmine* 1.08 **(b)** *Carmine* 1.09 **(c)** *Capri* (Embutido)

Figura 3.15 – Câmeras de luz estruturada da *PrimeSense*
Fonte: PRIMESENSE NATURAL INTERATION (156)

3.3 Modelo escolhido

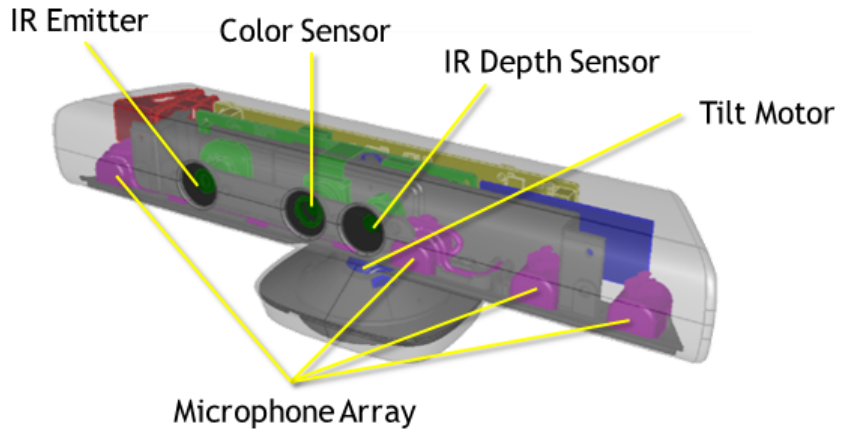
Lançada em novembro de 2010 (153), a câmera de profundidade *Kinect* para *XBOX foi* criada pela companhia israelense *PrimeSenses* (157), sendo a principal contribuição da *Rare* (158), empresa subsidiária da *Microsoft Game Studio*s (159), o desenvolvimento de softwares para a interpretação de gestos a partir deste dispositivo. Este foi então o primeiro modelo de câmera de luz estruturada vendido diretamente ao público em geral. Sendo originalmente desenvolvido para funcionar no console *XBOX 360*, ela teve sua primeira biblioteca de funções para computador, a *libfreenect* produzida por Hector Martin (160), poucos dias depois do seu lançamento. Após um pronunciamento na mídia, onde a *Microsoft* voltou atrás contra a acusação de hackeamento de seu produto (161), foram lançados outros softwares para a utilização do *Kinect* para *XBOX 360* em computadores, primeiramente pela empresa contratada para desenvolver este produto, a *PrimeSense*s, e posteriormente pela própria *Microsoft*, que atualmente já mantém uma segunda versão deste dispositivo especialmente melhorado para funcionar em computadores, denominado então como *Kinect* para *Windows*.

Diferente do novo dispositivo, o *Kinect* para *Xbox 360* foi construído apenas para ser utilizado com o *Xbox 360*, e por isso não é licenciado para uso comercial em qualquer outra plataforma, no entanto essa ainda é a câmera de profundidade baseada em luz estruturada mais facilmente encontrada para aquisição, e mantém o recorde do *Guinness World Records* como: o periférico para jogos mais rapidamente vendável (162), por ter obtido a marca de vendas de mais de oito milhões de unidades em apenas dois meses após o seu lançamento e vinte e quatro milhões em fevereiro de 2013. Essa também é atualmente a câmera de profundidade com maior compatibilidade para as diversas bibliotecas de função criadas para manipulação deste tipo de hardware, onde algumas dessas apresentam compatibilidade para outros tipos de sensores de funcionamento similar, enumerados na Seção 3.2.5, o que o torna o modelo ideal para estudos, bem como para a popularização de interfaces gestuais baseadas nesse tipo de tecnologia.

3.3.1 Especificações

Apesar do *Kinect* para *XBox* ser comumente descrito como um sensor para captura de movimentos de usuários, ele na verdade é uma câmera de profundidade baseada em luz estruturada, conforme visto na seção 3.2.5. Esse tipo de dispositivo é capaz de obter matrizes de profundidade relativas ao ambiente, o que permite a sua utilização para diversos outros tipos de aplicações, sendo a interpretação de gestos realizada por outros aparelhos, como computadores pessoais e o console *XBOX 360*, por meio das informações que ele fornece. Os principais componentes desse sensor consistem em: uma câmera colorida, um emissor laser infravermelho, uma câmera monocromática com filtro para luz infravermelha e um sistema computacional interno.

Além disso também são utilizados um acelerômetro, um conjunto de microfones embutidos e um pequeno motor elétrico (151, 163).A Figura 3.16 (164) exibe a distribuição de alguns desses dispositivos, onde o sistema computacional supracitado pode ser parcialmente visto. Como as especifidades de cada componente do *Kinect* para *XBOX* variam de um aparelho para outro, segue abaixo uma descrição obtida através de engenharia reversa realizada em um *Kinect* para *XBOX* o específico.



picos

Figura 3.16 – Distribuição dos sensores e atuadores
Fonte: KINECT FOR WINDOWS SENSOR COMPONENTS AND SPECIFICATIONS (164)

3.3.1.1 Câmera colorida

A Câmera *RGB* encontrada no sensor em questão, Figura 3.17a (165), utiliza o sensor *CMOS* MT9M112, da *Aptina Imaging* com 1,3 *megapixels* de resolução. A saída de vídeo desta tem por padrão oito bits de profundidade e resolução *VGA* (640 × 480 *pixels*) com um filtro de cor Bayer e taxa de atualização padrão de 30 Hz. Esta câmera é capaz de alternativamente atingir uma resolução de até 1280x1024 *pixels* a uma taxa de atualização de 15 Hz, além de disponibilizar outros formatos de cor como o *UYVY*. O *Kinect* apresenta um ângulo de visão de 57° na horizontal e 43° na vertical para as duas câmeras que o compõe.

(a) Câmera colorida **(b)** Câmera IR **(c)** Emissor infravermelho

Figura 3.17 – Sensores e emissor do *Kinect*
Fonte: MICROSOFT KINECT TEARDOWN (165)

3.3.1.2 Câmera infravermelha e vídeo de profundidade

A câmera infravermelha deste aparelho, utiliza o sensor *CMOS* MT9M001, também da *Aptina Imaging*, com 1,3 *megapixels* de resolução e 5,2 mícron *pixels* de abertura, Figura 3.17b (165). Este modelo, apesar de antigo, provavelmente é empregado por valer-se de *pixels* maiores que os convencionais, já que estes funcionam melhor com a iluminação reduzida ocasionada pela filtragem infravermelha. O *Kinect* para *XBOX* pode transmitir a saída dessa câmera diretamente, funcionando como uma câmera infravermelha comum, onde ele exibe a luz estruturada projetada sobre o ambiente em uma resolução de 640x480 *pixels* a 30 Hz, ou 1280x1024 *pixels* a uma taxa de atualização de 15 Hz. O registro da luz estruturada projetada é utilizado para criar imagens de profundidade, onde o intervalo de percepção deste dispositivo é de 0,6 a 3,5 metros. Fora destes limites as imagens começam a degradar e perder precisão. A saída do vídeo de profundidade também é disponibilizada em resolução *VGA* (640 $\times$ 480 pixels) a 30 Hz com 11 bits de precisão, o que possibilita até 2.048 níveis de sensibilidade.

3.3.1.3 Emissor laser e Padrão de Projeção

Basicamente um emissor laser infravermelho, Figura 3.17c (165), projeta um padrão salpicado, estático e quasi-periódico sobre o ambiente (166), similar ao da Figura 3.18 (167). Esse laser apesar de ser invisível à visão humana, pode ser captado por câmeras com filtros específicos para o seu comprimento de onda. O seu emissor é constituído por um diodo laser, que irradia um comprimento onda de 830nm, com uma potência de saída em torno de 60mW, não se utiliza de modulação e apresenta um nível de saída constante. Provavelmente por medida de segurança é também provido de um sensor térmico e um fotodiodo para calcular a sua potência de saída atual. Sua temperatura é mantida constante através de um pequeno elemento de Peltier, montado entre ele e uma a placa de alumínio, que pode tanto esquentar quanto esfriar o laser e assim estabilizar seu comprimento de onda. A potência desse laser é perigosa aos olhos se não for distribuída por seu gerador de padrão óptico. Conforme visto em uma das patentes referentes a esse dispositivo (168), o padrão emitido é quasi-periódico com múltiplas simetrias, é não contêm qualquer célula repetida ao longo de sua área. Um padrão deste tipo pode ser gerado para uma simetria n, com intensidade local $I(r)$, através da equação

$$I = \left\| \sum_{m=0}^{n-1} e^{ik_m.r} \right\|^2$$

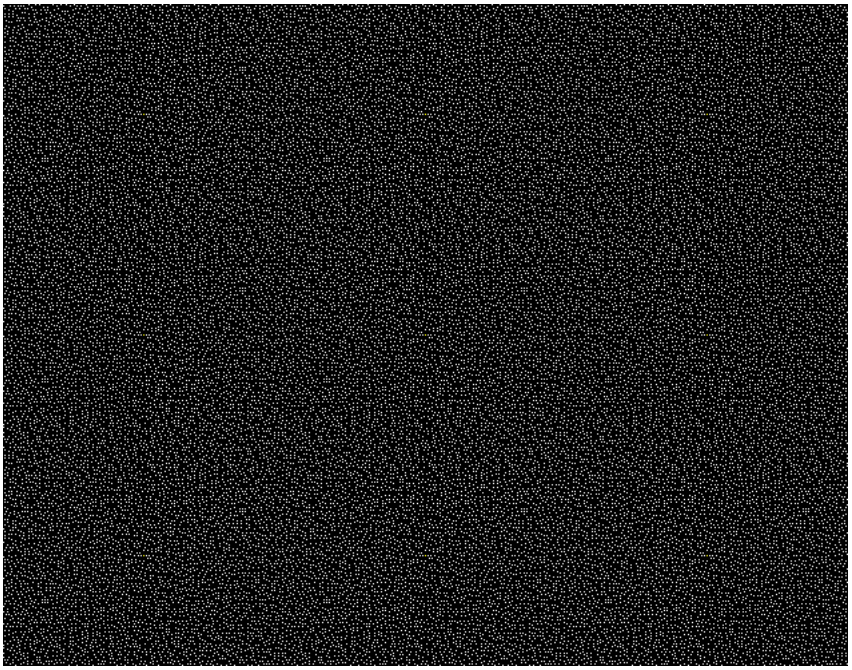


Figura 3.18 – Padrão quasi-periódico emitido pelo *Kinect*
Fonte: REICHINGE, A(167)

, onde k_me obtido utilizando-se a expressão

$$k_m = \left(cos\frac{2\Pi m}{fold}, sin\frac{2\Pi m}{fold} \right)$$

. A utilização de padrões quasi-periódicos provê um espectro de frequência conhecida, com picos facilmente distinguíveis, onde o processador utilizado aproveita esta informação espectral para filtrar as imagens capturadas e assim reduzir os ruídos inseridos pela iluminação ambiente no cálculo de correlação da imagem. Além disso, já que esse padrão é relacionado com as profundidades mapeadas pelo sistema, a probabilidade se produzir resultados errados é reduzida, uma vez que apenas uma correspondência correta entre sua projeção em uma parte da cena e uma área da imagem de referência produz valores significativos. Para produzir o padrão supracitado o laser emitido pelo diodo passa por um difusor, Figura 3.19, que além de espalhar sua luz é utilizado também para filtrar parte das radiações (166). Dessa forma é criado um padrão preliminar, que se completa com a utilização de um elemento óptico de difração, responsável por produzir um fenômeno onde se obtém diferentes focos, para direções angulares distintas. Além disso, elemento óptico de difração é projetado para reduzir o ângulo de divergência e com isto possibilitar que a intensidade da luz varie menos com a distância. Deste modo o padrão quasi-periódico emitido pelo difusor, é então distribuído de forma a espalhar-se sobre diferentes orientações de acordo com a mudança de distância focal, como ilustrado na Figura 3.20.

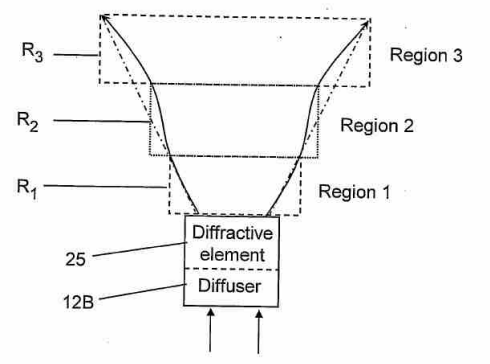


Figura 3.19 – Difusor e elemento óptico de difração
Fonte: FREEDMAN, B et al.(168)

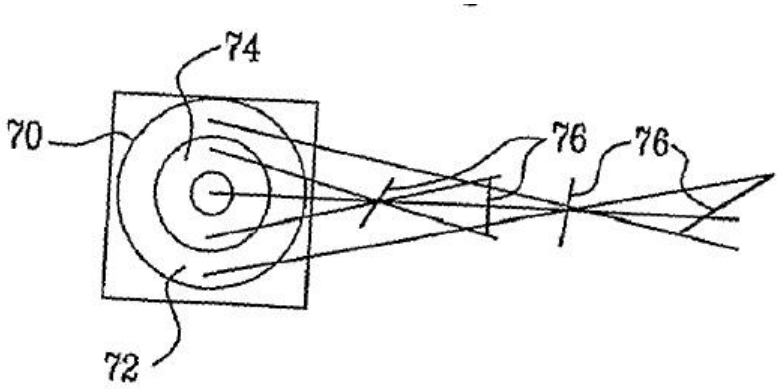


Figura 3.20 – Orientação conforme a distância focal
Fonte: FREEDMAN, B et al.(168)

Já que o padrão produzido varia com a distância, os objetos irradiados por ele também apresentam um padrão de iluminação variado, conforme o seu afastamento para com o sensor.Este padrão de emissão e reconhecimento de formas, através de manchas luminosas infravermelhas corresponde a luz estruturada patenteada pela *PrimeSense* (151, 166), que possibilita o mapeamento tridimensional de cenas do mundo real, analisando as variações de sua forma e espaçamento. Conforme ilustrado na Figura 3.21, essa análise consiste então na correlação do padrão projetado sobre o ambiente, com as possíveis variações de distribuições luminosas registradas, conforme a distância dos objetos em cena para com o dispositivo utilizado.

3.3.1.4 Matriz de microfones

O *Kinect* para *XBOX* pesquisado contém um par de conversores estéreo, WM8737L A/D da *Wolfson Microelectronics*, com pré-amplificadores embutidos, conectados a uma matriz

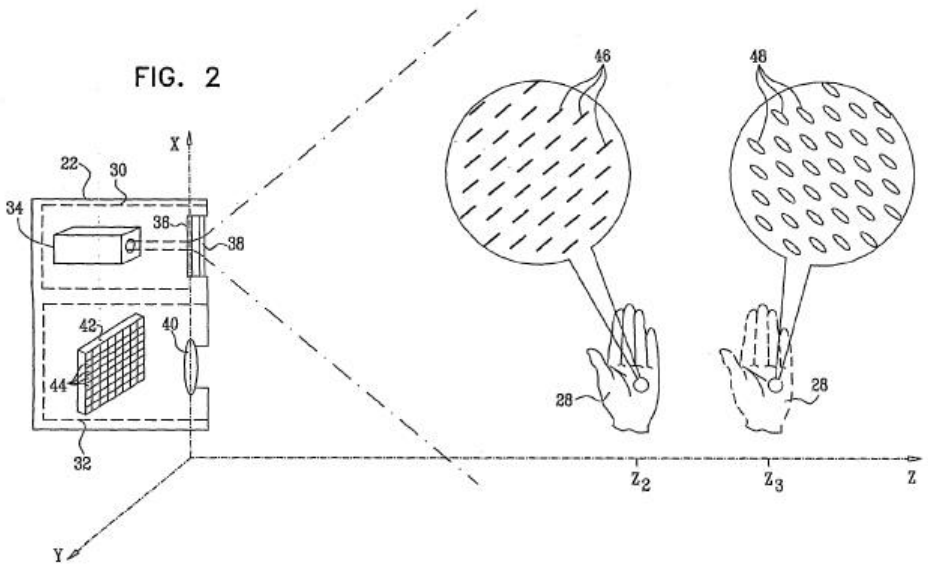


Figura 3.21 – Orientação da luz pela distância
Fonte: FREEDMAN, B et al.(168)

contendo quatro microfones, Figura 3.22 (165). Essa aparelhagem opera sobre canais de áudio de 16-bit a uma taxa de atualização de 16 kHz, e possibilita a captação as vozes mais próximas, filtragem de ruídos externos e diferenciação de pessoas que estejam conversando em um mesmo ambiente.

Figura 3.22 – Microfones do *Kinect* para *XBOX*
Fonte: MICROSOFT KINECT TEARDOWN. (165)

3.3.1.5 Sistema computacional

O chip da *PrimeSense* PS1080, Figura 3.23 (165), é responsável por controlar o projetor infravermelho, interpretar os dados das câmeras, calcular as matrizes de profundidade e coletar os dados de áudio dos microfones. Deste modo esse componente é diretamente responsável pelo gerenciamento de todos os sensores deste dispositivo, além do controle do projetor infra-

vermelho, que é parte indispensável do sistema de captura das imagens de profundidade. Esse

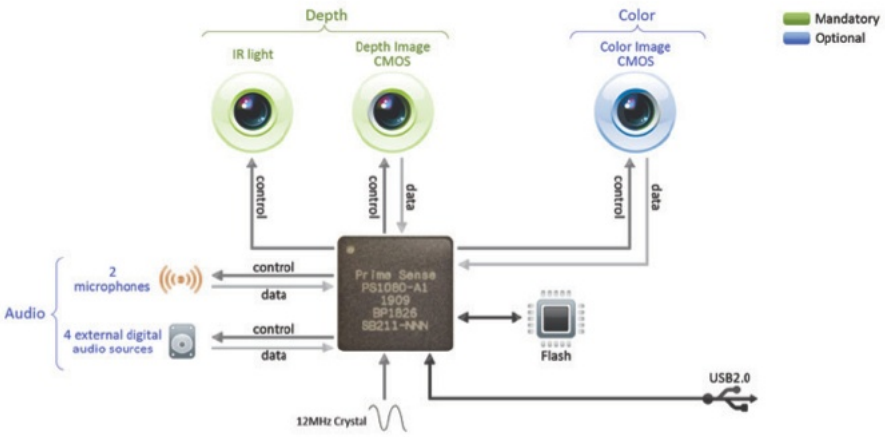


Figura 3.23 – Placa *Prime Sense* PS1089
Fonte: MICROSOFT KINECT TEARDOWN. (165)

chip se encontra em comunicação via *USB* 2.0 com o processador de aplicação PXA168-a da *Marvell Aspen*, que por sua vez é responsável por controlar as funções de áudio do sensor. A comunicação *USB* 2.0 utilizada costuma ser a maior empecilho de velocidade e qualidade para a captura de imagens utilizando o *Kinect*, além de limitar a quantidade de dados transmitidos internamente, restringindo o tamanho padrão das imagens fornecidas. A Figura 3.24 (163) contém as placas de circuitos encontradas dentro do *Kinect* para *XBOX* e a localização dos seus principais componentes.

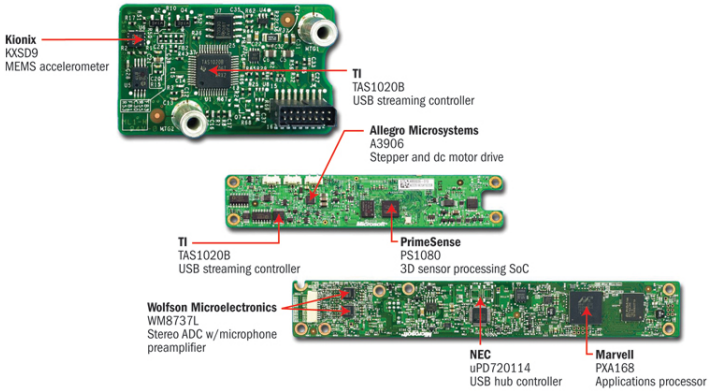


Figura 3.24 – Placas mãe utilizadas pelo *Kinect* para *XBOX*
Fonte: WIDENHOFER, B. (163)

3.3.1.6 Acelerômetro

Este aparelho também possui um acelerômetro de três eixos que utiliza tecnologia micro eletro mecânica (*MEMS*). O acelerômetro encontrado consiste em um KXSD9 da *Kionix*, com capacidade para detecção de movimentos de até duas unidades de aceleração da gravidade. Este dispositivo é responsável por informar a inclinação do sensor em três dimensões, com um limite inferior de precisão de apenas um grau; no entanto, devido à sua sensibilidade ao aquecimento, esse limite pode atingir um desvio de até três graus, ainda dentro do intervalo de temperatura normal para o seu funcionamento.

3.3.1.7 Motor de inclinação e adaptador para portas *USB*

O motor encontrado utiliza um drive DC A3906 da *Allegro Microsystems*, Figura 3.25 (165), capaz de inclinar o dispositivo analisado em até 31°, para cima ou para baixo. Já que esse motor requer mais energia do que as portas *USB* 2.0 do *XBox 360* podem fornecer, a *Microsoft* lançou para a nova versão desse console, o *XBox 360 S*, com um conector especial, capaz de combinar as capacidades de comunicação *USB* com energia adicional requerida para o funcionamento pleno desse sensor. Posteriormente foi lançado um cabo proprietário que serve como adaptador para o *XBox 360 FAT* e como fonte de energia extra para a alimentação do motor embutido no *Kinect* para *XBox*. Este é o mesmo adaptador que possibilita a utilização do *Kinect* para *XBox* em computadores, e que foi empregado para a produção do seu primeiro drive, através de técnicas de engenharia reversa.

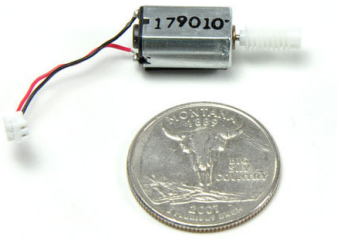


Figura 3.25 – Motor que controla a inclinação do sensor
Fonte: MICROSOFT KINECT TEARDOWN. (165)

[§]Disponível em: `http://www.eetimes.com/document.asp?doc_id=1281322` Acesso em fev. 2014

Capítulo 4

Proposta

Este trabalho visa a criação de uma interface humano computador que se baseie na detecção das poses assumidas pelas mãos de seus usuários, seja eficaz no reconhecimento de um conjunto abrangente de classes, apresente baixo custo computacional, seja portável para múltiplos sistemas operacionais e sensores de profundidade, e extensível para o aprendizado de novas poses. A Interface proposta também deve ser capaz de localizar dedos, mãos e braços dos usuários em coordenadas tridimensionais correspondentes, possibilitando assim a interação com ambientes virtuais; além disso o sistema desenvolvido deve ser capaz de reconhecer um conjunto de poses de mãos distintas, permitindo a utilização de comandos pré-determinados, bem como a emissão de um conjunto mais abrangente de informações, produzidas através da combinação das poses iniciais.

O projeto também tem como meta a utilização de câmeras de profundidade para o registro de um conjunto heterogêneo de candidatos, realizando uma diversidade razoável de poses de mãos, sob ângulos e distâncias variadas do dispositivo de captura. A motivação para este fim consiste na carência de bases de profundidade disponíveis contendo poses de mãos variadas realizadas por múltiplos usuários, e na importância desse instrumento para a avaliação e comparação entre sistemas similares baseados em visão computacional (169).

Pretende-se dessa forma gerar contribuições científicas e tecnológicas, nos modos de um software de baixo custo computacional para o reconhecimento de gestos, e com estudos e desenvolvimento de técnicas, capazes de incorporar informações de profundidade para aprimorar o reconhecimento de gestos. Dessa maneira espera-se promover a popularização deste tipo de interface, possibilitando a sua utilização em dispositivos de pequeno porte, sistemas de realidade aumentada e acessibilidade a surdos-mudos, de forma extensível e configurável; além de prover, através da base de vídeos gerada, um meio de comparação para com métodos de detecção de gestos manuais alternativos, bem como suporte ao desenvolvimento de novos sistemas para esse fim.

4.1 Base de gestos

Para a avaliação e o treinamento da interface descrita nessa monografia, uma base de gestos manuais foi gerada, contendo 26 classes de poses estáticas. Cada pose selecionada foi associada a um símbolo, de modo que o conjunto de todos os símbolos, correspondentes a uma classe de gesto manual a ser reconhecida pelo sistema, pode ser representado pelo vetor:

$$G = \{\rightarrow, \infty, 4, A, B, C, D, E, F, G, I, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y\}$$

. A Figura 4.1 exibe um exemplo de imagem para cada classe registrada, em conjunto com o símbolo utilizado para a sua representação.

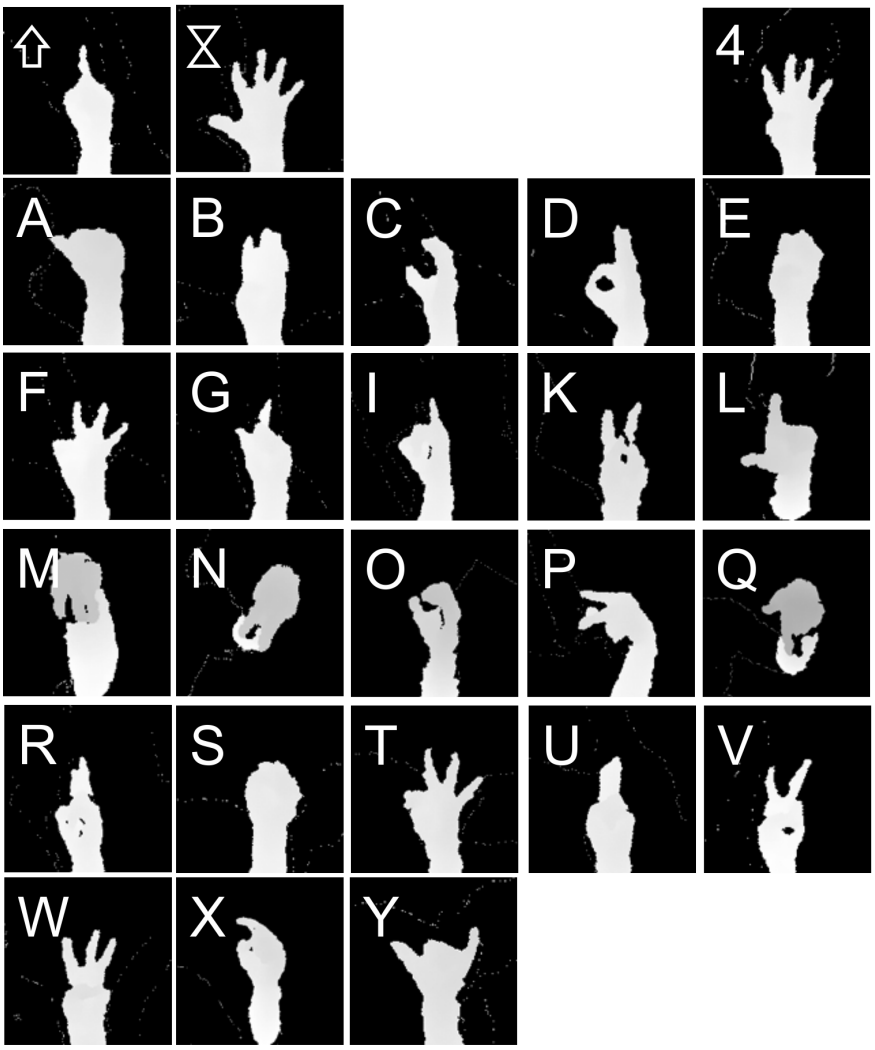


Figura 4.1 – Exemplos das classes selecionadas
Fonte: Elaborada pelo autor

Essas poses foram registradas de modo que a mão dos usuários estivessem a distâncias e orientações variáveis, de aproximadamente 70 cm a 3,2 metros e 360° graus de rotação

em relação ao eixo de profundidade do dispositivo. Esse procedimento foi repetido por cinco voluntários que apresentavam tamanhos e espessuras de mãos e dedos distintos, de forma que no total foram gravados 130 arquivos de vídeo contendo imagens de profundidade, imagens coloridas, rastreamentos e esqueletizações de usuários dos participantes. Essa base ocupa um espaço em disco de 31 Gigabytes, e tem a duração estimada em aproximadamente quatro horas.

As classes supracitadas, assim como os caracteres associados, foram extraídos da Língua de Sinais Brasileira (LIBRAS) (33), com foco nos gestos estáticos de seu alfabeto datilográfico. Visando uma maior variedade de poses com possibilidade de reconhecimento em quadros individuais, foram escolhidas para serem registradas na base proposta todos os gestos que não necessitavam de quaisquer movimentos para a sua caraterização. Foram acrescentadas também outras poses, que apesar de apresentarem movimentos originalmente, poderiam ser diferenciadas através de suas representações estáticas dos sinais supracitados. A Figura 4.2 (170) exibe instruções visuais de como expressar o alfabeto datilográfico de Libras, onde a maioria dos sinais pode ser proferido através de poses simples, sendo os movimentos inclusos descritos através de linhas indicadoras na cor branca.

Figura 4.2 – Sinais do alfabeto datilográfico de libras
Fonte: CURSOS DE LIBRAS ONLINE GRÁTIS COM CERTIFICADO. (170)

Dado a extensão dessa base e a grande variação nas posições das mãos e dedos dos usuários no espaço, optou-se por não fazer anotações referentes às suas posições, devido a falta de tempo hábil disponível para esse fim. Visto que o intuito principal deste trabalho é o reconhecimento de uma vasta quantidade de poses, onde cada vídeo registrado contempla

uma única classe destas, priorizou-se por um maior refinamento dos algoritmos desenvolvidos para o reconhecimento dessas, a custo de uma capacidade de validação numérica mais precisa dos métodos de rastreamento.

4.2 Métodos para a detecção de gestos e rastreamento de mãos e dedos

Uma vez que o sistema proposto tem como meta ser portável para múltiplos dispositivos, ele deve ser capaz de funcionar com um desempenho satisfatório em plataformas com limitações de processamento e armazenamento de dados, onde muitas vezes não haverão GPU's disponíveis para a renderização de modelos geométricos ou paralelização de processos. Deste modo, foram pesquisados e desenvolvidos diversos algorítimos para classificação de imagens baseadas na abordagem de reconhecimento de gestos por extração de características, explanada na Seção 2.4.1.2. A razão para tal é que esta metodologia costumam apresentar menores custos computacionais (171) do os métodos baseados em rastreamento de modelos (54, 74, 122).

Já que muitos métodos de baixa complexidade, são capazes de segmentar as imagens de profundidade e fornecer informações em três dimensões precisas a partir delas, esses dados podem então ser utilizados para obter a localização de suas partes constituintes e para facilitar o reconhecimento dos gestos representados (12).Com base nesse fato, foram pesquisadas diversas formas de estimar a posição dos dedos, mãos e braços dos usuários, afim de possibilitar a interação com cenários virtuais e auxiliar o funcionamento de sistemas de aprendizado de máquina, na classificação dos gestos propostos.

Obedecendo a nomenclatura adotada na Seção 2.4.2, para as fases inerentes ao reconhecimento de gestos manuais, os métodos utilizados estão divididos entre as categorias: identificação de dados, extração de características e classificação de gestos. As duas primeiras categorias deste processo são utilizadas tanto para a extração de imagens da base criada (visando o treinamento e validação da interface, Seção 4.1) quanto para a obtenção em tempo real das imagens de ambas as mãos do usuário fornecidas pelo *Kinect.* O ultimo passo também utiliza esses dois grupos de imagens, onde o primeiro grupo é aplicado ao treinamento do classificador desenvolvido, tornando possível o reconhecimentos das poses registradas no segundo grupo.

Afim de facilitar o entendimento do processo como um todo, uma visão geral de seu fun-

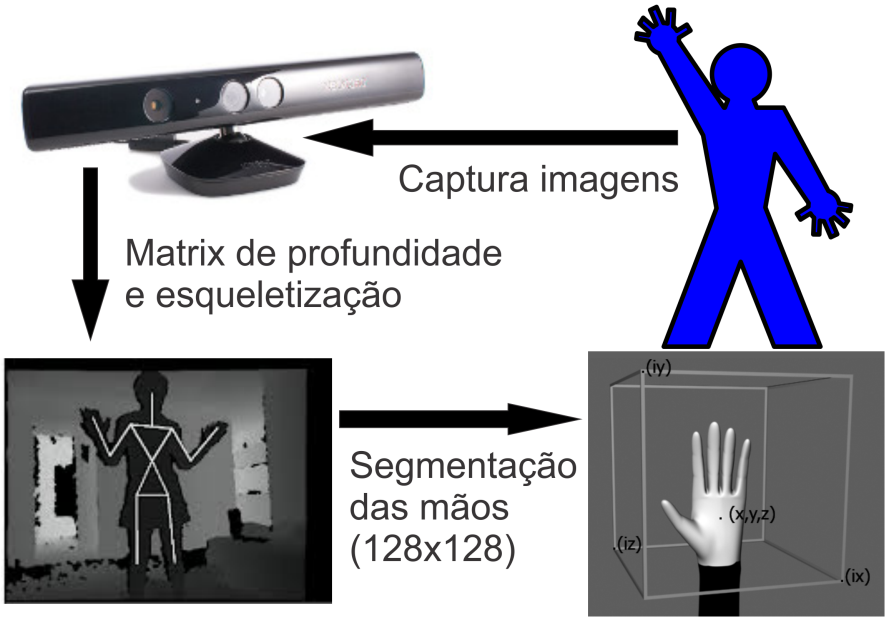


Figura 4.3 – Segmentação das mãos
Fonte: Elaborada pelo autor

cionamento é representada pelas três figuras seguintes, onde são apresentados em sequência as diversas relações entre os métodos descritos nesta mesma seção. Inicialmente são segmentadas as imagens de profundidade capturadas pelo sensor de profundidade, Figura 4.3, afim de isolar as poses das mãos conforme detalhado na Seção 4.2.1.1.

Para obtenção de dados mais propícios a classificação de poses, são processadas as imagens de profundidade segmentadas provindas da base (descrita na Seção 4.1), assim como às do usuário atual, que são adquiridas em tempo real, Figura 4.4. De cada uma dessas imagens de profundidade de pose de mão são extraídas então diversas características, como o contorno (Seção 4.2.1.2) e o sequenciamento de suas informações em um vetor unidimensional (Seção 4.2.3.3).

A partir do contorno da mãos são obtidas então a decomposição de casco convexo (Seção 4.2.2.1) e a esqueletização baseada em triangulação (Seção 4.2.2.2). De maneira similar são utilizadas técnicas de redução de dimensionalidade no vetor contendo os dados da imagem, calcula-se os seus *eigengestures* (172) e *fishergestures* (173), descritos na Seção 4.2.2.3 e na Seção 4.2.2.4, respectivamente.

A partir desse dados é possível realizar então o rastreamento e a classificação das poses de mãos, conforme pode ser visto na Figura 4.20. O rastreamento se baseia na posição das pontas de dedos encontrados pela decomposição de casco convexo, para sistemas com menor capacidade de processamento; ou nos dados fornecidos pela esqueletização baseada em

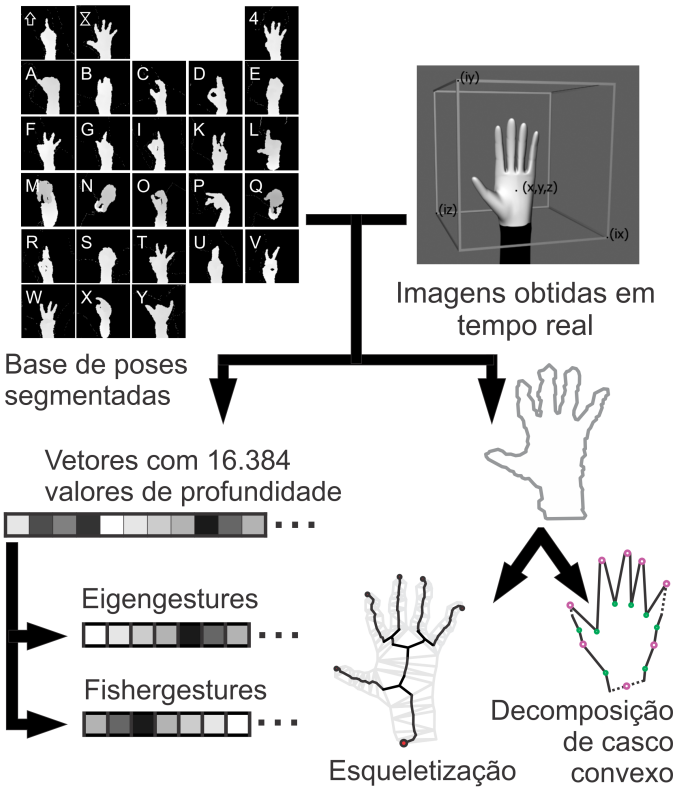


Figura 4.4 – Extração de características
Fonte: Elaborada pelo autor

triangulação, quando se deseja maiores detalhes sobre a disposição de seus membros.

A classificação das poses utiliza prioritariamente os vetores de características gerados pelo sequenciamento dos dados da imagem, e pelas técnicas de redução de dimensionalidade supracitadas; mas também pode se beneficiar da contagem de terminações e dedos encontrada, conforme descrito em detalhes na Seção 4.2.3.4.

4.2.1 Identificação de dados

A identificação de dados é um processo inerente a sistemas de visão computacional, utilizado para obter e isolar informações úteis ao seu funcionamento, a partir de conjuntos de *pixels* contendo informações de tipos variados. Uma vez que as imagens digitais matriciais costumam apresentar informações prejudiciais a sua análise, como: ruídos e objetos indesejados, além de serem compostas por uma quantidade extensa e redundante de informações, como: cores idênticas em células vizinhas; costuma-se utilizar representações reduzidas destas, como as enumeradas na Seção 2.4.2.1.

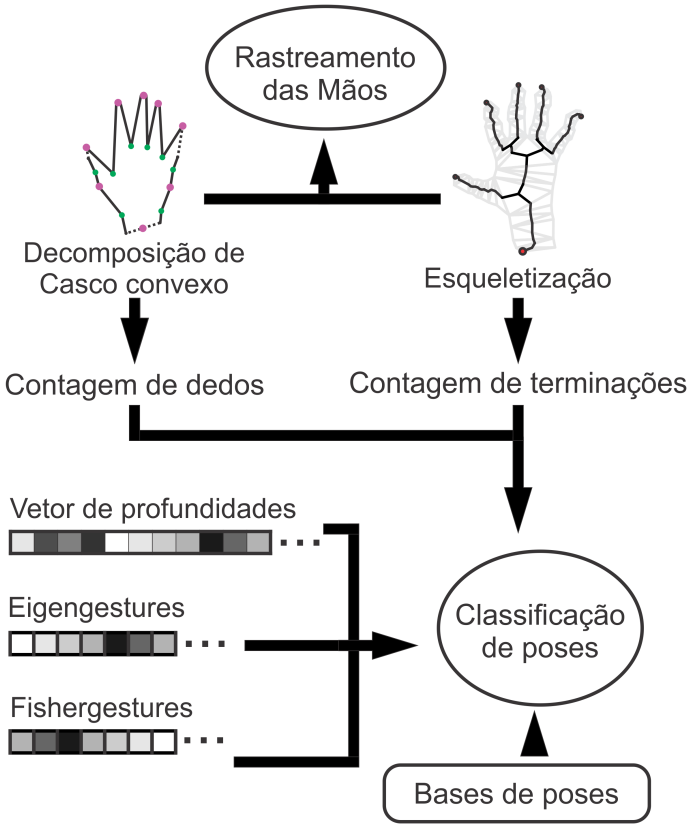


Figura 4.5 – Informações utilizadas para o rastreamento e classificação de poses
Fonte: Elaborada pelo autor

Para o sistema descrito neste trabalho, foram extraídos das imagens de profundidade capturadas dois tipos de dados: linhas poligonais, relativas ao contorno de cada uma das mãos, e imagens de profundidade normalizadas quanto a rotação, escala e posição no espaço tridimensional.

4.2.1.1 Limiarização tridimensional

O primeiro conjunto de dados a ser adquirido consiste nas imagens de profundidade das mãos do usuário, às quais servirão de base para todos os outros métodos desenvolvidos, visando a identificação de suas poses. Com o auxílio da segmentação e da esqueletização geradas pela biblioteca *OpenNI* (13), essas imagens devem ser obtidas de forma que sejam similares quanto a suas posições, orientações e proporções em relação a escala, afim de que sejam processadas independentemente da variação dessas características.

Inicialmente a imagem de profundidade contendo toda a cena é filtrada pela segmentação de usuário supracitada, para que os demais elementos sejam removidos. Com base na es-

queletização correspondente essa imagem é então transladada fazendo que a posição da mão analisada se encontre no seu centro, Figura 4.6.

No entanto o sistema foi configurado para não realizar quaisquer tarefas relacionadas ao rastreamento ou reconhecimento de poses manuais, caso o percentual de confiança das posições da mão e do cotovelo (emitidos pela *OpenNI)* seja inferior a 70%. Essa medida foi tomada uma vez que muitas esqueletizações corporais fornecidas apresentam membros em posições não correspondentes à pose assumida pelos usuários, mesmo quando o seu valor de percentual de confiança é declarado como 100%.

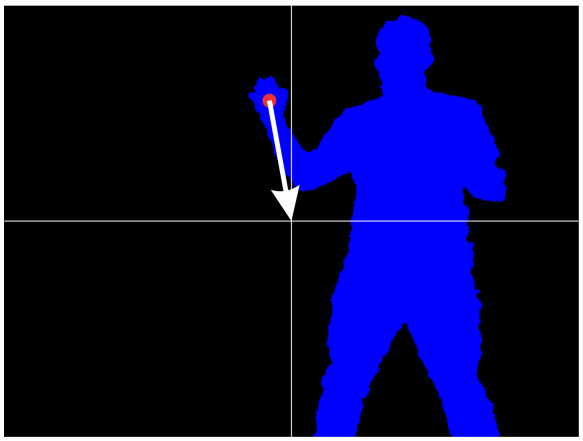


Figura 4.6 – Centralização
Fonte: Elaborada pelo autor

O deslocamento a ser realizado na imagem pode ser calculado através da subtração de um vetor bidimensional c, contendo as coordenadas do centro da imagem, por um vetor m, correspondendo a mão analisada, na forma $\Delta_T = c - m$, onde $c = \{\frac{w}{2}, \frac{h}{2}\}$; sendo w a largura da imagem e h a sua altura.

A rotação da imagem resultante é então aplicada com centro em c, alinhando a mão e o cotovelo correspondente com o eixo vertical, Figura 4.7.

O ângulo dessa rotação em radianos é calculado com base nas coordenadas bidimensionais da mão m e do cotovelo e pela fórmula:

$$\theta = \frac{\Pi}{4} - \arctan(e_y - m_y, m_x - e_x)$$

, sendo o primeiro termo a orientação desejada, e o segundo termo a inclinação atual em que se encontra a pose de mão correspondente.

Uma área quadrada com centro em c é então recortada da imagem rotacionada e copiada para uma nova, de modo a isolar a mão do restante da cena, Figura 4.8.

Para que a pose de mão presente em cada imagem tenha um tamanho proporcional em

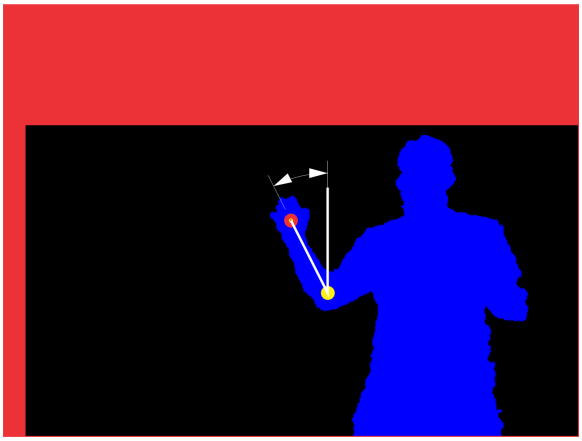


Figura 4.7 – Rotação
Fonte: Elaborada pelo autor

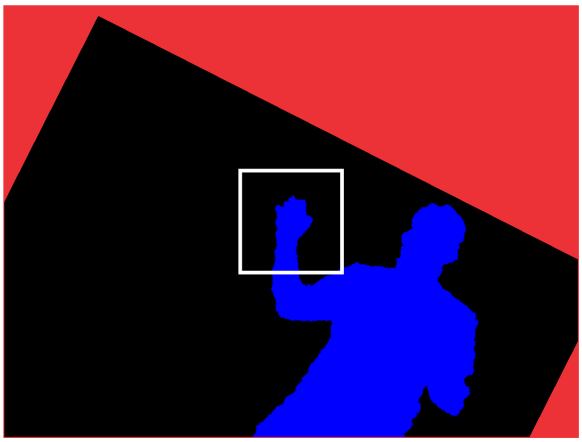


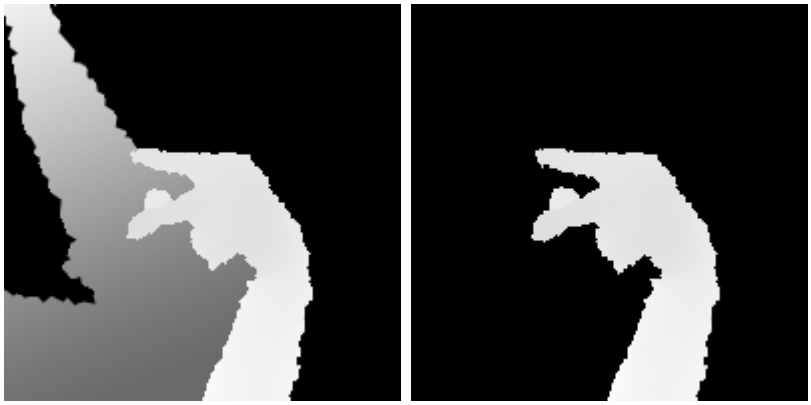
Figura 4.8 – Recorte
Fonte: Elaborada pelo autor

relação a área de recorte, o lado d da área quadrada varia com a profundidade m_z da mão, na forma $d = \frac{k}{m_z}$, onde k é uma constante predefinida para o sistema desenvolvido.

Por fim, é realizada uma análise de cada um dos *pixels* $P(x, y)$, afim de remover todos aqueles que apresentem valores de profundidade fora do intervalo I esperado para a pose de mão em questão, ou $P(x, y) \notin I$, isolando-a tridimensionalmente do restante da cena, Figura 4.9.

O intervalo esperado para a pose é obtido pela profundidade da mão m_z e do cotovelo correspondente e_z, na forma $I = [max, min]$, sendo $min = m_z - q$, (onde q é uma variação de profundidade pré estabelecida), e

$$max = \begin{cases} (m_z + q) < e_z \longrightarrow & m_z + q_z \\ (m_z + q) \geq e_z \longrightarrow & e_z \end{cases}$$



(a) Isolamento bidimensional **(b)** Isolamento tridimensional

Figura 4.9 – Filtro de profundidade
Fonte: Elaborada pelo autor

; de modo a preservar o antebraço na imagem final, mesmo que se encontre a uma diferença de profundidade considerável da mão.

Em outras palavras, este método remove todos os *pixels* da imagem que apresentam valores de profundidade menores que a profundidade da mão subtraída de um valor pré estabelecido, e também remove todos aqueles que apresentam valores maiores que a profundidade do cotovelo. Alternativamente, caso o cotovelo esteja mais próximo do sensor do que a mão do usuário, o valor máximo passa a ser a profundidade da mão somada a um valor preestabelecido.

Deste modo, foram obtidas segmentações de mãos a baixo custo computacional e reduziu-se significativamente a quantidade de *pixels* para futuras análises, uma vez que o tamanho das imagens e a quantidade de pixels presente nelas tende a ser bem menor após esse processo.

4.2.1.2 Extração de contornos

De posse das imagens de profundidade das mãos, obtidas conforme descrito na Seção 4.2.1.1, normaliza-se seus valores de profundidade para transformá-las em imagens de tons de cinza convencionais. Estas são então binarizadas, tendo como resultado imagens representando as suas silhuetas. A extração dos contornos é posteriormente realizada, com o auxílio da biblioteca para visão computacional de código aberto *OpenCV* (174), aplicando-se uma função baseada no algoritmo de Suzuki e Abe (175), que recupera múltiplos contornos a partir de uma imagem binária. Optou-se pelo modo de recuperação mais completo para que sejam gerados todos os contornos possíveis, além de um método de aproximação simplificado para evitar auto intersecções, uma vez que promovem mal funcionamento em outros módulos

do sistema.

O contorno com maior número de pontos é então selecionado dentre todos os outros, removendo ruídos e partes do corpo remanescentes, que não foram filtrados por estarem dentro dos limites tridimensionais estabelecidos para a pose de mão presente na imagem. Para reduzir o número de vértices do contorno encontrado, é utilizado um método de suavização por aproximações poligonais. Cada contorno extraído dessa maneira equivale a uma curva vetorial fechada simples, representando a silhueta da sua respectiva pose.

4.2.2 Extração de características

A extração de características de um conjunto de dados, consiste em um método para combinar esses dados de forma a caracterizá-los com precisão suficiente para resolver um problema proposto. Combinações mais simples podem ser realizadas através de processos de redução de dimensionalidade; no entanto, melhores resultados costumam ser obtidos quando especialistas selecionam informações mais propícias para esta tarefa (176).

Classicamente uma quantidade fixa e pré-determinada de características desejáveis é extraída de um conjunto de dados mais abrangente. Para este fim, um conjunto fixo de dados, denominado como vetor de características, deve ser selecionado de acordo com o problema em questão, de forma que esta representação possa ser utilizada como entrada para os processos aplicados a sua solução no lugar do grupo de dados original.

Visando a obtenção de um conjunto de informações adequados à representação de poses manuais complexas, pesquisou-se por algoritmos para a generalização das imagens de poses manuais obtidas (conforme visto na Seção 4.2.1.1), e métodos para identificar a posição e a orientação dos dedos e das mãos através dos contornos das imagens adquiridos, conforme descrito na Seção 4.2.1.2. A generalização das imagens foi realizada por meio dos métodos *eigengestures* (172) e *fishergestures* (173), consistindo estes em aplicações de algoritmos amplamente utilizados para reconhecimento facial. Por sua vez a identificação da posição e orientação dos membros da mão, foi realizada por meio de uma decomposição de casco convexo (20, 177, 178), e de um método de esqueletização baseado em triangulação de contornos.

Enquanto o popular método de decomposição de casco convexo fornece dados úteis para a dedução da pose realizada, e permite encontrar a posição da ponta e da base de cada dedo que esteja provocando uma protuberância significativa na silhueta das imagens, a custos computacionais mínimos; a esqueletização baseada na triangulação de Delaunay (179) é capaz

de representar com maior precisão as orientações e dimensões das várias regiões de interesse presentes nestas mesmas silhuetas, com um desempenho adequado a interação em tempo real.

Já que a decomposição do casco convexo e o algoritmo de esqueletização proposto podem ser obtidos por meio da triangulação de Delaunay do contorno das mãos (180), (sendo desta forma executados em conjunto sem que haja maiores prejuízos ao desempenho do sistema), além de apresentarem uma variedade de detalhamento e precisão adequadas para assegurar uma maior consistência dos dados, por meio de sua redundância intrínseca. Optou-se pela aplicação de ambos os métodos, visando uma localização mais eficaz dos membros das mãos, e a aquisição de características mais robustas para a identificação de suas poses, do que as fornecidas por meio de qualquer um desses individualmente.

4.2.2.1 Decomposição de casco convexo

O conceito de casco convexo, também conhecido como concha convexa, ou envelope convexo é bastante utilizado para relacionar características encontradas em contornos de imagens, onde muitas pesquisas tem aplicado algoritmos que o implementam para o reconhecimento de gestos manuais simples. Está técnica utiliza contornos extraídos de imagens de mãos para detectar a localização de pontas de dedos e identificar suas partes constituintes, sendo desta forma sensível apenas às informações contidas na silhueta dessas imagens.

O casco convexo aplicado a um conjunto de finito de pontos $S \in \mathbb{R}^2$, ou $E \in \mathbb{R}^3$, retorna um polígono, ou um poliedro formado pela menor quantidade desses mesmos pontos, capaz de delimitar um volume convexo C que contém todos os pontos pertencentes ao conjunto em questão. Esse mesmo conceito aplicado a um conjunto finito de pontos $P \in \mathbb{R}^n$, produz um politopo convexo em $\mathbb{R}^n$.

A Figura 4.10 (181) ilustra o polígono C que corresponde ao casco convexo do conjunto S de pontos contidos em um plano, e a Figura 4.11 (182) ilustra o poliedro C' que corresponde ao casco convexo do pontos E, distribuídos no espaço.

Analogias comumente utilizadas, para explicar o conceito e casco convexo são a de uma borracha esticada em torno de um monte de pregos martelados previamente sobre uma plano; ou uma superfície elástica envolvendo um objeto. A Figura 4.12 (183) ilustra o primeiro caso, onde os pregos de cor verde que tocam a borracha representam o casco convexo do conjunto.

Embora seja um conceito muito simples, encontrar algoritmos que sejam capazes de obter

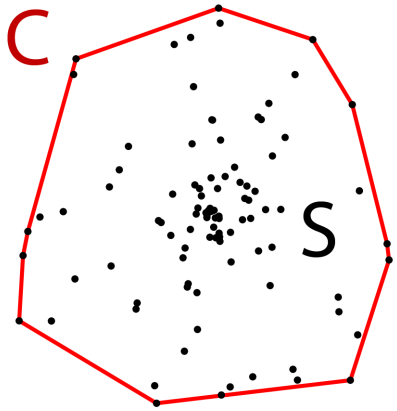


Figura 4.10 – Casco convexo em $\mathbb{R}^2$
Fonte: RUFAT, D. (181)

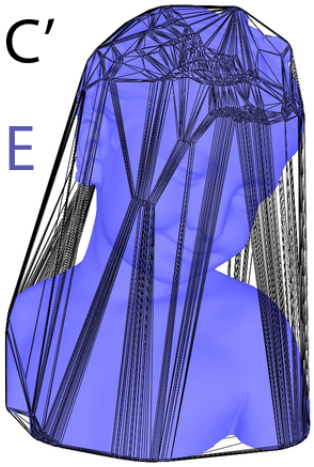


Figura 4.11 – Casco convexo em $\mathbb{R}^3$
FONTE: HERT, S; SCHIRRA, S. (182)

o casco convexo de um conjunto finito qualquer de pontos, em espaços Euclidianos de baixa dimensão, é considerado um dos problemas fundamentais da geometria computacional (184).

Formalmente, o casco convexo de um conjunto de pontos X pode ser definido das seguintes maneiras:

1. Um único conjunto convexo, com a mínima quantidade de pontos capaz de limitar todos os pontos de X.
2. A interseção de todos os conjuntos convexos que contém X.
3. O conjunto de todas as combinações convexas dos pontos de X.
4. A união de todos os simplexos onde todos os seus vértices pertencem a X.

Partindo da terceira definição, uma combinação convexa associa para cada ponto $x_i \in X$ um coeficiente α_i, de modo que todos esses coeficientes sejam não negativos e a sua soma total

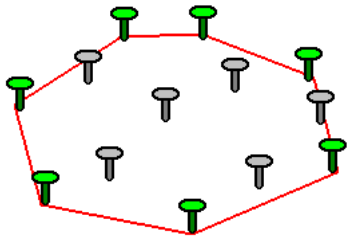


Figura 4.12 – Analogia da borracha esticada
Fonte: FITTING A POLYGON ABOUT A SET OF POINTS. (183)

seja igual a um. Esses valores são então utilizados como o peso de uma média ponderada entre as coordenadas de todos os pontos. Para cada escolha de coeficientes, a combinação convexa resultante é um ponto do casco convexo, e o casco convexo como um todo pode ser obtido pela escolha de coeficientes de todos os modos possíveis. A afirmação anterior pode ser então expressa como a Fórmula 4.2.1:

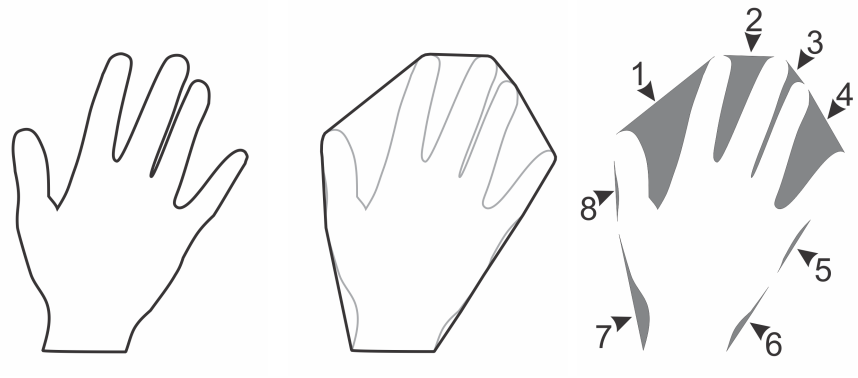
$$\left\{ \sum_{i=1}^{|s|} \alpha_i x_i \,|\, (\forall i : \alpha_i \leq 0) \wedge \sum_{i=1}^{|s|} \alpha_i = 1 \right\} \tag{4.2.1}$$

deste modo, cada ponto $x_i \in X$ que não estiver no casco convexo dos outros pontos, ou ($x_i \notin Conv(X \setminus \{x_i\})$) é parte dos vértices que representam o casco convexo $Conv(X)$. Apesar da utilização dessa definição para a obtenção do casco convexo ser pouco eficiente, existem vários algoritmos para esse fim capazes de obter com precisão e desempenho a representação correta da forma convexa desejada (185, 186).

A complexidade desses algoritmos costuma ser estimada em termos do número de pontos n do conjunto X, e do número de pontos h de seu casco convexo correspondente $Conv(X)$. Para conjuntos de coordenadas bidimensionais e tridimensionais, existem algoritmos sensíveis à saída capazes de computar o casco convexo em $O(n \log h)$. Para o cálculo deste conceito em dimensões maiores, a menor complexidade obtida é de $O(n^{\lfloor d/2 \rfloor})$, onde d é a quantidade de dimensões das coordenadas que descrevem os pontos $x_i \in X$.

Um conceito intimante relacionado ao casco convexo é o defeito convexo. Os defeitos convexos de um contorno podem ser encontrados pela diferença entre esse polígono e casco convexo encontrado para ele, a Figura 4.13 ilustra o processo geral, onde a Figura 4.13a exibe o contorno dado como entrada, a Figura 4.13b representa o casco convexo obtido, e a Figura 4.13c apresenta então os defeitos convexos encontrados.

A seguir é explanado como os conceitos de casco convexo e defeito convexo são utilizados para identificação dos membros da mão. Para a implementação deste método foi utilizada a função *cvConvexityDefect* da biblioteca *OpenCV*, que obtém os n defeitos convexos D_n de um



(a) Contorno de mão **(b)** Casco convexo **(c)** Defeitos convexos

Figura 4.13 – Extração de defeitos convexos
Fonte: Elaborada pelo autor

contorno C qualquer, representando cada um deles pelos dois pontos onde eles tangenciam o casco convexo, e por um terceiro ponto que correspondente a sua coordenada mais profunda. A Figura 4.14 ilustra os pontos obtidos em sentido horário, onde as cores vermelha e azul representam respectivamente as coordenadas iniciais s e finais e de cada defeito convexo e a cor verde sinaliza o seu ponto mais profundo c.

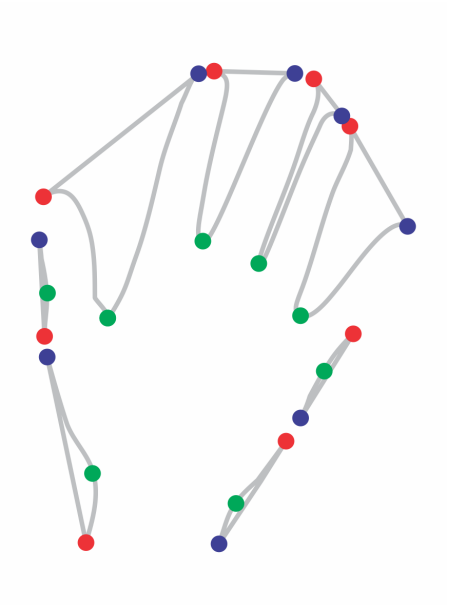


Figura 4.14 – Extremidades
Fonte: Elaborada pelo autor

No intuito de reconhecer terminações, formadas por pares de defeitos convexos consecutivos (D_i, D_{i+1}), com características similares às frequentemente encontradas para às pontas individuais e aglomerados de dedos, várias relações presentes entre os seus dados são analisadas.

O primeiro dado extraído é o inicio de sua base, através do ponto médio entre os centros de seus defeitos convexos $b = \frac{c_i + c_{i+1}}{2}$. A seguir, as coordenadas mais protuberantes da terminação, correspondendo à e_i, s_{i+1}, são armazenadas em um vetor, juntamente com o seu ponto médio $m = \frac{e_i + s_{i+1}}{2}$, na forma $\{e_i, m, s_{i+1}\}$. A partir desses dados, as distâncias d_0, d_1, d_2 do início da base para com os três pontos são então calculadas, como: $d_0 = \|e_i - b\|$, $d_1 = \|m - b\|$, $d_2 = \|s_{i+1} - b\|$, onde a coordenada mais distante $d = \max(d_0, d_1, d_2)$, é escolhida como sendo a localização da ponta

$$p = \begin{cases} d = d_0 & \rightarrow e_1 \\ d = d_1 & \rightarrow m \\ d = d_2 & \rightarrow s_{i+1} \end{cases}$$

, conforme demonstrado na cor roxa pela Figura 4.15.

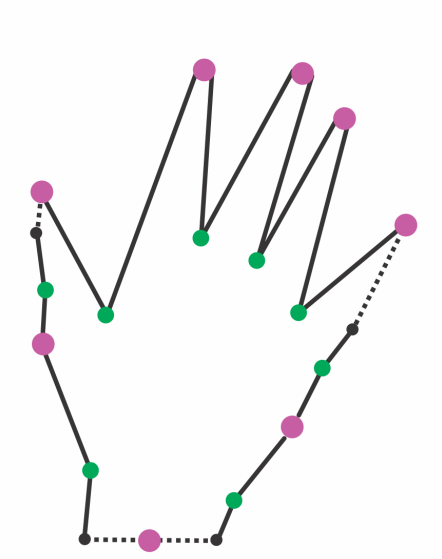


Figura 4.15 – Pontos médios
Fonte: Elaborada pelo autor

Por fim os pontos encontrados são avaliadas quanto a seu comprimento d, a largura de sua ponta l, a largura de sua base w e o ângulo entre as extremidades da base e a ponta a. O comprimento d corresponde ao mesmo valor encontrado durante a análise das protuberâncias, a largura da ponta l é obtida por $\|e_i - s_{i+1}\|$, a largura da base corresponde a $w = \|c_i - c_{i+1}\|$, e o ângulo é calculado com o auxílio dos vetores $v_0 = e_i - m$ e $v_1 = s_{i+1} - m$, na forma

$$a = \arccos\left(\frac{v_0 \cdot v_1}{\|v_0\| \, \|v_1\|}\right) \text{rad}$$

.

Uma terminação válida é confirmada caso está apresente um comprimento maior que um valor mínimo, e um ângulo menor do que um valor máximo pré-determinados. No entanto, quando a largura da ponta e a largura da base são maiores do que o esperado, esse ponto passa a ser identificado como sendo o pulso ou um aglomerado de dedos, onde o pulso corresponde a terminação mais próxima do cotovelo. A Figura 4.16 ilustra os pontos rejeitados na cor azul, os dedos identificados na cor vermelha, e o pulso como um ponto preto, as medidas supracitas (comprimento, ângulo, largura da ponta e largura da base) também são representadas através exemplos. Desse modo, os padrões encontrados a partir das concavidades de um contorno são

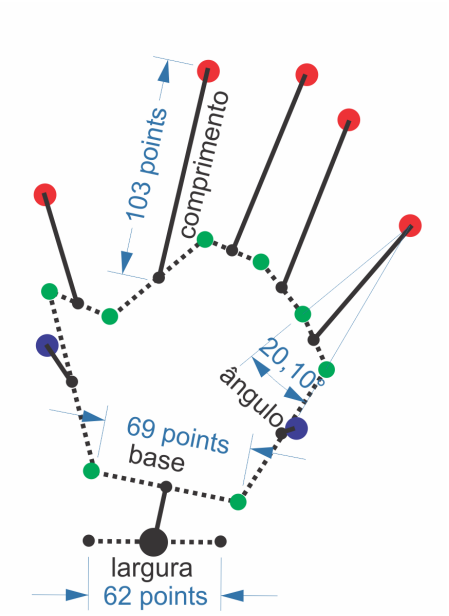


Figura 4.16 – Membros identificados
Fonte: Elaborada pelo autor

modelados como informações relativas a poses simples de mãos, onde é possível reconhecer a posição e a direção dos dedos e do pulso quando estes se fazem proeminentes na silhueta da imagem. A Figura 4.17 ilustra o resultado do método em funcionamento.

4.2.2.2 Esqueletização baseada em triangulação

Entre outras definições (187, 188), uma esqueletização de imagens pode ser descrita como o conjunto de centros, pertencentes a todas as n-esferas* de maior tamanho que podem ser circunscritas em seu interior (189). Uma n-esfera H é dita de maior tamanho em um conjunto

*Uma esfera com n dimensões, podendo corresponder a um ponto, um segmento de reta, um círculo, uma esfera, ou uma hiperesfera, dependendo da quantidade de dimensões assumida.

Figura 4.17 – Padrão modelado via análise de casco convexo
Fonte: Elaborada pelo autor

I se $H \subseteq I$, e se qualquer outra n-esfera $G \supset H \rightarrow G \subseteq I$. Esse conjunto de centros forma estruturas lineares, similares a um esqueleto, bastante utilizadas em reconhecimento de padrões e visão computacional. A Figura 4.18 ilustra a esqueletização de uma imagem retangular bidimensional, formada pelo conjunto infinito dos centros dos círculos de tamanho máximo que podem ser contidos nessa imagem.

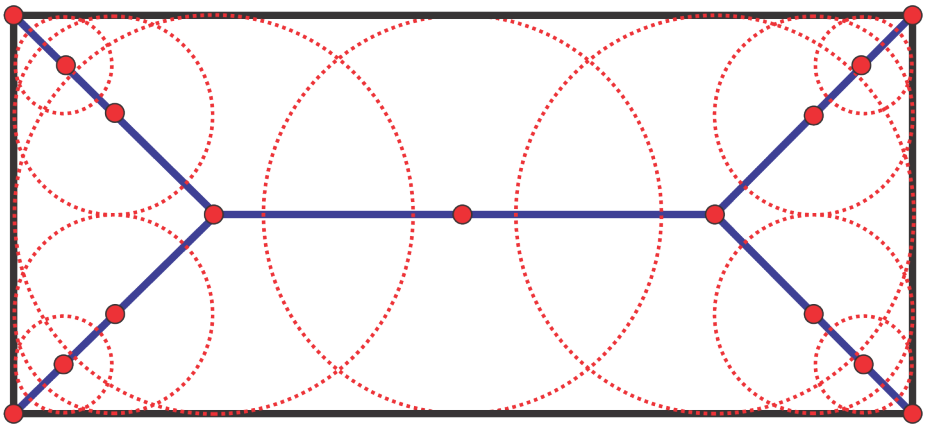


Figura 4.18 – Conjunto dos centros de todas n-esferas máximas
Fonte: Elaborada pelo autor

Além de ser usualmente empregada para obtenção de descritores de formas (19, 190), entre as aplicações que se beneficiam desse conceito podem ser destacadas: a cartografia automática (191), recuperação de conteúdo baseados em imagem (192), reconhecimento de caracteres (193), inspeção de placas de circuito impresso(194), análise de imagens biomédicas (195) e o rastreamento de movimentos humanos (196).

Existem diversos algoritmos de esqueletização fundamentados na definição matemática original (188), de forma que estes são usualmente classificados quanto aos métodos em que se baseiam, como por exemplo: afinamento (197), transformada da distância (198), redução iterativa de contorno (199), morfologia matemática (200), diagrama de Voronoi (201) e triangulação de Delaunay restrita (179). Uma vez que a utilização de cada uma dessas abordagens

tem suas vantagens e desvantagens em comparação às outras (202), após uma avaliação dessas, a esqueletização por triangulação de Delaunay restrita foi escolhida; sendo os seus pontos fortes e fracos em relação a métodos similares destacados a seguir:

Pontos fortes: É capaz de gerar esqueletos vetoriais estruturados, que apresentam baixo custo de armazenamento, processamento e avaliação; de forma rápida, robusta e eficaz.

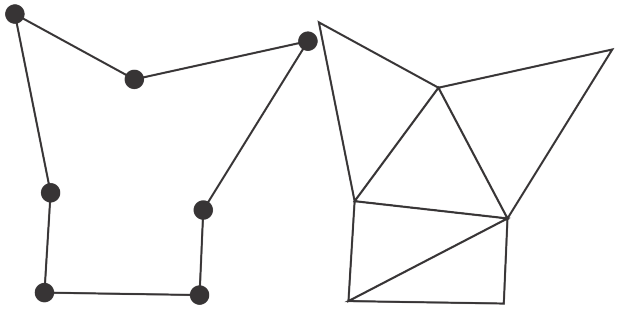
Pontos fracos: Não obedece a definição original de esqueletização, encontrando apenas estruturas aproximadas.

O algoritmo implementado para este trabalho, com base em sua descrição, tem como entrada as coordenadas do cotovelo conectado ao centro da mão (de acordo com a esqueletização de usuário da *OpenNI*), e os contornos do primeiro grupo descrito na Seção 4.2.1.2. Já que cada contorno de mão utilizado é formado por uma única linha poligonal fechada simples, a esqueletização resultante é representada por uma árvore binária.

O processo de esqueletização de contornos é composto por uma sobreposição de funções, sendo estas realizadas inicialmente a partir de relações de vizinhança dos triângulos obtidos, e posteriormente com base na própria árvore gerada, de modo que esta é refinada sucessivamente até a obtenção de sua versão final. A sequência de métodos aplicados para este fim consiste em: triangulação de contorno, classificação de triângulos, criação da árvore inicial e poda com base em comprimento e área.

Triangulação de contornos: Para que a esqueletização poligonal seja realizada é necessário que o contorno da mão seja triangulado, Figura 4.19, e que os triângulos resultantes sejam classificados quanto a sua vizinhança, Figura 4.20. Com este intuito foi utilizada a biblioteca de função *Triangle* (203), já que ela produz diferentes tipos de triangulação com alto desempenho e fornece estruturas auxiliares de fácil acesso e baixo custo computacional, adequadas para a implementação do método aqui descrito.

Inicialmente um contorno poligonal P (Figura 4.19a) é processado por um algoritmo de triangulação de Delaunay restrita (204) $f(:P) \rightarrow T$, resultando em um conjunto de triângulos T, Figura 4.19b. Entre os dados obtidos pela triangulação de contorno da *Triangle*, destacam-se dois vetores, $V_{n\times3\times2}$ e $N_{n\times3}$, onde n corresponde a quantidade de triângulos encontrados para um contorno poligonal P qualquer. Estes dois vetores são indexados pelos números de identificação i, gerados para cada triângulo T_i, onde o primeiro contém as três coordenadas bidimensionais dos vértices dos triângulos e o segundo fornece os índices de cada um seus vizinhos, retornando o valor $\emptyset$ nos casos de referências a vizinhos inválidos.



(a) Contorno poligonal P **(b)** Conjunto de triângulos T

Figura 4.19 – Triangulação de contorno
Fonte: Elaborada pelo autor

Classificação de triângulos: Percorrendo o vetor de vizinhanças N e somando a quantidade de seus vizinhos válidos de cada triângulo, obtém-se o vetor C_n, onde:

$$\begin{aligned} C &= \{C_i \mid \forall i \in \mathbb{N} \wedge 1 \leq i \leq n\}, \\ C_i &= \left\{ \sum_{j=0}^{2} \alpha_{ij} \mid \alpha_{ij} = \begin{cases} 0 & N_{ij} = \emptyset \\ 1 & N_{ij} \neq \emptyset \end{cases} \right\} \end{aligned}$$

, de forma que os valores C_i são utilizados para classificar os triângulos T_i quanto as suas quantidade de vizinhos, durante a fase de criação da árvore inicial. Como consequência do contorno de entrada P consistir em uma única linha poligonal fechada simples, sempre que a triangulação de contorno retornar uma quantidade de triângulos $n > 1$, a classificação de vizinhança resultará em $C_i = \{ x \mid \forall x \in A = \{1, 2, 3\}\}$, não havendo neste caso necessidade de tratamento de exceções, uma vez que o algoritmo utilizado é configurado para funcionar com contornos apresentado uma quantidade mínima de vértices superior a gerada por apenas um triângulo, de modo a evitar ruídos abruptos e aumentar a sua robustez. A Figura 4.20 exibe a classificação realizada, diferenciando os tipos de triângulos pelo número de vizinhos e através de suas cores, sendo o primeiro tipo na cor branca, o segundo cinza claro e o terceiro cinza médio.

Nesta fase também, é gerado um vetor A_n, relativo a área de cada triângulo T_i, que são calculadas através de suas coordenadas V_{ic}, na forma:

$$\begin{aligned} A &= \{A_i \mid \forall i \in \mathbb{N} \wedge 1 \leq i \leq n\}, \\ A_i &= \left\{ \frac{1}{2} \cdot |(V_{i11} - V_{i31})(V_{i22} - V_{i12}) - (V_{i11} - V_{i21})(V_{i32} - V_{i12})| \right\}. \end{aligned}$$

Este vetor é utilizado posteriormente na fase de poda baseada em comprimento e área, em conjunto com o tamanho dos segmentos produzidos na criação da árvore inicial.

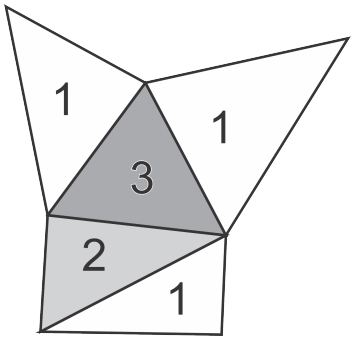


Figura 4.20 – Classificação por tipo de vizinhança
Fonte: Elaborada pelo autor

Criação da árvore inicial: Visando gerar uma árvore com navegação intuitiva e sem a necessidade de futuras ordenações, esse método se inicia pela busca do triângulo contendo o vértice correspondente a raiz da árvore, de modo que a esqueletização inicial é realizada através de uma avaliação recursiva de sua vizinhança.

O primeiro triângulo a ser processado é encontrado por meio de uma busca sequencial pelo ponto v no grupo de vértices

$$H = \{x \mid \forall x \in V_{ke}, x \notin V_{N_k} \wedge C_k = 1\}$$

, onde $k = \{w | \forall w \in \mathbb{N} \wedge 1 \leq w \leq n\}$, $e = \{u | \forall u \in \mathbb{N} \wedge 1 \leq u \leq 3\}$; de modo que $\sqrt{v^2 - o^2} \leq \sqrt{l^2 - o^2}$, dado $l = \{x \mid \forall x \in H\}$ e o ponto o contendo a localização do cotovelo conectado a essa mesma mão,.

Por consequência, todo o vetor V deverá ser percorrido, uma vez que a posição ocupada pelo vértice desejado não é conhecida a priori; no entanto, já que somente o vértice não compartilhado dos triângulos com um único vizinho devem ser analisados, o cálculo de distâncias para com o cotovelo é realizado apenas nesse conjunto reduzido. Cada distância obtida é comparada com uma variável *min*, que armazena sempre o valor mínimo encontrada, possibilitando assim que o índice de seu respectivo triângulo seja registrado na *variável* i_0. A Figura 4.21 exibe esse processo, onde apenas os vértices não compartilhado pelos vizinhos dos triângulos destacados são levados em conta para o cálculo da menor distância até as coordenadas cotovelo.

De posse do índice pertencente ao triângulo inicial, a esqueletização do conjunto de triângulos T é obtida através uma função g, que tem como entrada: o índice do triângulo atual i, o índice do triângulo anterior j, os vetores V, N ,C e um nó pai F gerado previamente. Definida como $g(i, j, V, N, C, F) \rightarrow B$, além de calcular as coordenadas de cada aresta do esqueleto, essa função é capaz de produzir nós de árvore binária concatenados de acordo com a vizinhança

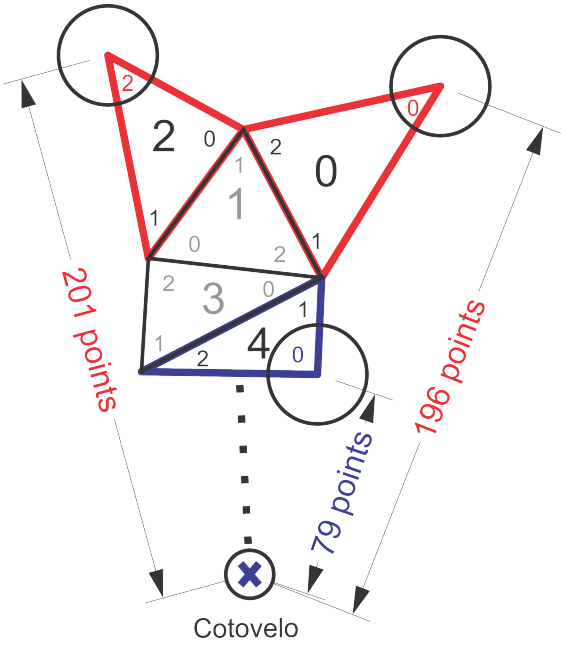


Figura 4.21 – Seleção do ponto inicial
Fonte: Elaborada pelo autor

de triângulos encontrada; onde cada um de seus nós é definido como $b = \{v, S_1, S_2, F\}$, sendo v as suas coordenadas bidimensionais, S_1e S_2 os seus dois nós filhos e F o seu nó pai. Além desses dados, cada nó b gerado pelo sistema tem outros campos que serão utilizando posteriormente na fase de poda, como: índice do triângulo equivalente $b.i$, e o comprimento das arestas $b.s$; gerado por meio da distância de cada vértice até as coordenadas de seu nó pai $b.s = \sqrt{(v - F_p)^2}$, juntamente com o cálculo das posição v de cada nó da esqueletização.

A função g é responsável por processar toda a vizinhança de triângulos através de chamadas recursivas aos índices de seus vizinhos, na forma $g(N_i \neq j, i, V, N, C, B)$. Está função foi implementada de modo a evitar nós redundantes e assegurar uma navegação intuitiva, variando seu funcionamento de acordo com o tipo de triângulo encontrado. Cada tipo de triângulo é reconhecido pela função g através do valor de C_i, onde o triângulo inicial é identificado por meio de seus parâmetros diferenciados $g(i_0, 0, V, N, C, \emptyset)$ fornecidos como nó pai e para o índice do triângulo visitado anteriormente.

A Figura 4.22 exibe a sequência de nós gerados para cada tipo de triângulo identificado, indicando a ordem de inserção através de legendas e cores padronizadas; os triângulos na vizinhança do triângulo principal são representados por linhas pontilhadas, e etiquetados com as letras a e p (com ou sem índices na forma p_n) para referenciar triângulos vizinhos visitados e não visitados respectivamente.

Primeiro triângulo selecionado: Para este tipo de triângulo dois nós são inseridos na árvore

na forma $R = \{v_1, S_1, \emptyset, \emptyset\}$ e $S_1 = \{v_2, S_2, \emptyset, R\}$, onde $v_1 = V_i - V_{N_i}$ equivale as coordenadas do vértice não compartilhado com o triângulo vizinho e $v_2 = \frac{1}{2} \cdot \sum_{l=1}^{2} O_l$ corresponde ao ponto médio da aresta compartilhada $O = \{V_i \bigcap V_{N_i}\}$. O nó S_2 é obtido então através da função g, onde $S_2 = g(N_i, i_0, V, N, C, S_1)$.

Três vizinhos: Já que neste tipo de triângulo são formadas as bifurcações do esqueleto, três nós deverão ser inseridos, sendo estes serem representados como $R = \{v_1, S_1, S_2, F\}$, $S_1 = \{v_2, S_3, \emptyset, R\}$, $S_2 = \{v_3, S_4, \emptyset, R\}$. A primeira coordenada será equivalente ao baricentro do triângulo, $v_1 = \frac{1}{3} \cdot \sum_{l=1}^{3}(V_i)$, e as outras coordenadas equivalem ao ponto médio das arestas não compartilhadas com o vizinho anterior, ou

$$\{A_1, A_2\} = \{V_i - V_j\} \times \left\{V_i \bigcap V_j\right\}$$

, sendo $c_1 = \frac{1}{2} \cdot \sum_{l=1}^{2}(A_1)$ e $c_2 = \frac{1}{2} \cdot \sum_{l=1}^{2}(A_2)$. A sequência de inserção dessas coordenadas como nós filhos do nó central R, obedece então a ordem do menor para o maior valor, obtidos pela equação da reta, Figura 4.22d, resultante de suas coordenadas para com ás de seu nó pai F, como: $r(p(x, y)) \rightarrow ax+by+c = 0$, onde $a = F_y - R_y$, $b = R_x - F_x$, $c = F_x R_y - R_x F_y$. A comparação entre a posição dos pontos

$$o = \begin{cases} r(c_1) < r(c_2) \rightarrow & v_2 = c_1;\ v_3 = c_2 \\ r(c_1) > r(c_2) \rightarrow & v_2 = c_2;\ v_3 = c_1 \end{cases}$$

também afeta os nós restantes

$$\begin{aligned} S_3 &= g(i_1, i, V, N, C, S_1) \\ S_4 &= g(i_2, i, V, N, C, S_2) \end{aligned}$$

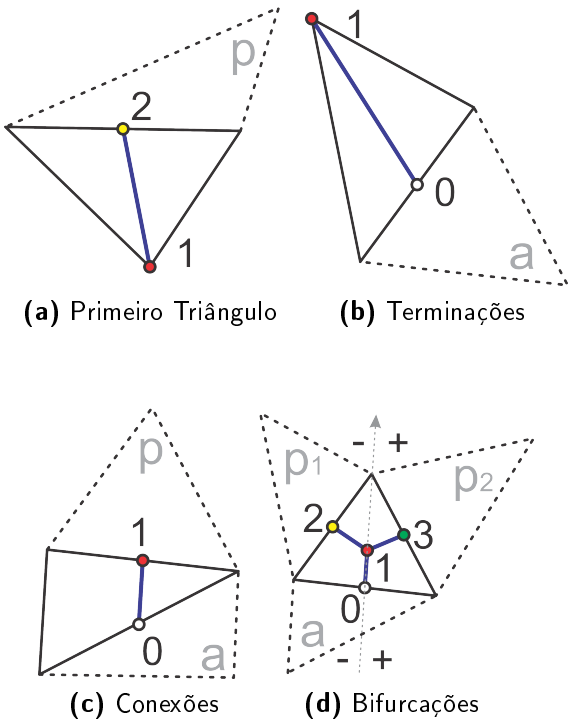
, onde o índice do triângulo vizinho a ser referenciado para cada um deles, correspondem aos que compartilham às arestas obtidas em

$$\begin{aligned} i_1 &= \left\{x \mid V_i \bigcap V_l = A_1,\ l = N_x\right\}, \\ i_2 &= \left\{x \mid V_i \bigcap V_l = A_2,\ l = N_x\right\}. \end{aligned}$$

Dois vizinhos: Forma uma conexão entre o seu nó pai F, para um outro nó, obtida por: $R = p, S, \emptyset, F$, tendo p como ponto médio da aresta compartilhada com o vizinho não visitado, $h = \frac{1}{2} \cdot \sum_{l=1}^{2} O_l$, onde $O = \{V_i \bigcap V_{N_i}\}$, Figura 4.22b. É inserido então o próximo nó a ser visitado em $S = g(N_i \neq j, i, V, N, C, S_1)$.

Um vizinho: Apenas um nó é criado como $T = \{v, \emptyset, \emptyset, F\}$, e as suas coordenadas são equi-

valentes às do único vértice não compartilhado deste triângulo, dado por $p = \{V_i - V_{N_i}\}$, Figura 4.22b. Esse tipo de triângulo não propaga chamadas a novos vizinhos através de g, já que se trata de uma terminação da árvore.



(a) Primeiro Triângulo **(b)** Terminações

(c) Conexões **(d)** Bifurcações

Figura 4.22 – Nós inseridos conforme o tipo de triângulo
Fonte: Elaborada pelo autor

A Figura 4.23 mostra uma possível esqueletização, representativa do contorno fornecido como exemplo. Neste pequeno conjunto de arestas, o ponto 0 é reconhecido como pertencente ao antebraço, e os pontos 1 e 2 são identificados como dois dedos ordenados quando ao seu nó pai.

O método descrito até este ponto funciona bem para contornos com poucos vértices, contendo um espaço razoável entre si. Contornos deste tipo podem ser obtidos por conta da distância das mãos para com o dispositivo, através de algoritmos de simplificação de contornos, ou mesmo utilizando a linha poligonal gerada pela decomposição do casco convexo, vista na Seção 4.2.2.1. Através da Figura 4.24 é possível observar nítidas discrepâncias entre os dois métodos para algumas posições de extremidades.

Poda baseada em comprimento e área: Apesar de apresentar resultados interessantes para linhas poligonais fechadas contendo poucos vértices, o método descrito até então não é capaz de encontrar as posições dos dedos em contornos mais elaborados, sendo para isto

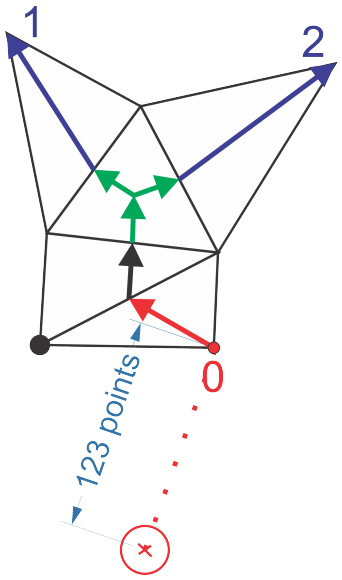


Figura 4.23 – Esqueletização poligonal em contornos com poucos vértices
Fonte: Elaborada pelo autor

necessária uma etapa conhecida como poda, que é comumente utilizada em esqueletizações de imagens de múltiplos tipos. A Figura 4.25, ilustra o resultado do processo, utilizando o mesmo sistema de cores das imagens anteriores, em contornos com detalhamento similar aos obtidos conforme descrito na Seção 4.2.1.2.

O critério de poda proposto, se baseia na análise do comprimento e da largura de cada ramificação das árvores geradas, com a finalidade de remover as cada terminação até a sua bifurcação inicial. Estas sequências de segmentos serão então denominadas neste trabalho como um **galhos terminais**, visando facilitar sua sua referência durante o decorrer dessa seção.

A poda é realizada a partir das folhas da árvore, e consiste em uma comparação dos respetivos valores encontrados em uma esqueletização de entrada, com as constantes estipuladas: para o tamanho mínimo de um dedo s_{min}, área máxima a_{max}, e proporção mínima entre essas medidas $q_{min} = \frac{l_{min}}{a_{max}}$.

Este algoritmo é baseado em processos recursivos, que podem ser divididos em duas fases principais: a busca por galhos terminais para a remoção; e a respectiva poda desses mesmos galhos que apresentem características menos promissoras. A Figura 4.26 ilustra seu funcionamento

Busca por Galhos: Além de encontrar e avaliar todos os galhos terminais da árvore, esse conjunto de métodos também é utilizado para verificar se pelo menos um galho terminal fora dos limites estipulados foi encontrado, que consiste na condição necessária para que os

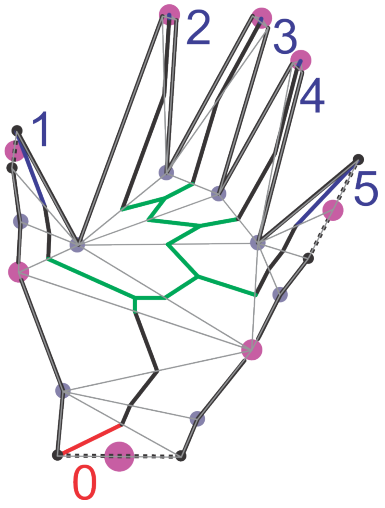


Figura 4.24 – Esqueletização resultante da análise de casco convexo
Fonte: Elaborada pelo autor

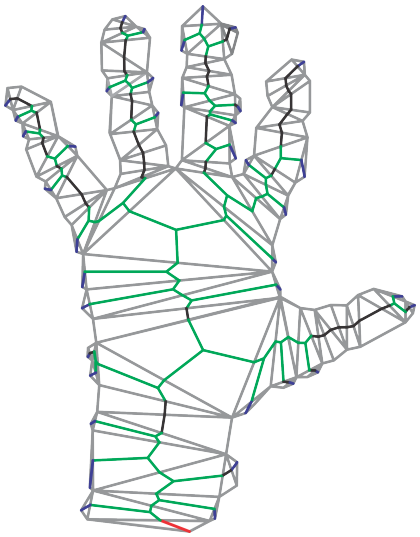


Figura 4.25 – Árvore pré-poda
Fonte: Elaborada pelo autor

algoritmos de poda sejam iniciados. Caso seja encontrado pelo menos galho terminal candidato a poda, o processo de busca por galhos ativa a função de poda de galhos, e então se reinicia, terminando apenas quando nenhum desses é encontrado.

Localização das folhas: O primeiro passo deste método é localizar as folhas da árvore gerada, para este fim uma função recursiva inicia a busca pela raiz da árvore e realiza chamadas binárias a todos os filhos, até chegar a um nó sem filhos, Figura 4.26a. É iniciada então uma rotina de processamento de características com base em seu nó atual.

Processamento de características: A partir de uma folha da árvore gerada, o processamento de característica percorre recursivamente cada nó pai, até chegar ao primeiro nó contendo dois filhos. Durante esse processo, são acumulados os valo-

res dos tamanhos $b_t.s$, de cada segmento percorrido t, na forma $S = \sum_{t=1}^{n} b_t.s$; assim como o valor das áreas $A_{b_t.i}$ de seus triângulos correspondentes, na forma $\alpha = \sum_{t=1}^{n} A_{b_t.i}$. É importante observar que os dados relativos a triângulos com três vizinhos não é adicionado a esses resultados, uma vez que não representam o comprimento ou a área dos galhos terminais adequadamente.

Avaliação de características: Caso um galho terminal apresente comprimento S, área α, ou proporção $Q = \frac{S}{\alpha}$, fora dos limites estipulados, esse galho será marcado para a avaliação de poda e os seus valores serão registrados nos nós filhos de bifurcações correspondentes, para que possam ser utilizados como critério de desempate, caso haja outro galho marcados para a remoção incidindo sobre a mesma bifurcação. A comparação entre os valores obtidos e as constantes supracitadas, são então realizados na forma:

$$isF = (s_{min} < S) \wedge (a_{max} > \alpha) \wedge \left(q_{min} < \frac{S}{\alpha}\right).$$

A função retorna então um valor válido para a presença de galhos terminais candidatos através da árvore, por concatenação de resultados a partir da bifurcação encontrada até a função de *Busca por galhos*, de forma a obter positivo se qualquer nó for diagnosticado como positivo e negativo apenas se todos os nós forem diagnosticados como negativos. A Figura 4.26b, apresenta os galhos terminais selecionados para avaliação de poda na cor vermelha, e os seus respectivos triângulos, em amarelo, além dos nós marcados para análise de poda na cor azul, e por fim os segmentos e os triângulos não atingidos diretamente pela avaliação de poda, nas cores preta e branca respectivamente.

Poda de galhos: Tendo como objetivo promover uma poda gradual da árvore, preservando os segmentos mais longos e que possuam a menor área; esse método é ativado sempre que um processo denominado anteriormente como: *Busca por galhos,* retorna um valor verdadeiro, sendo descartado em muitos contornos com poucos vértices.

Comparação entre galhos: Percorrendo a árvore a partir da raiz, é avaliado o número de requisições de poda $r = \{n \in \mathbb{N} \mid z \in \{0, 1, 2\}\}$ presentes em seus nós de bifurcação. Quando este valor é diferente de zero, a bifurcação correspondente é selecionada para uma avaliação de poda de seus galhos terminais. Se o valor de $r = 1$, o galho terminal correspondente é passado como parâmetro para uma função de remoção. No entanto, caso hajam duas requisições de poda relativas a um mesmo nó de bifurcação, a proporção

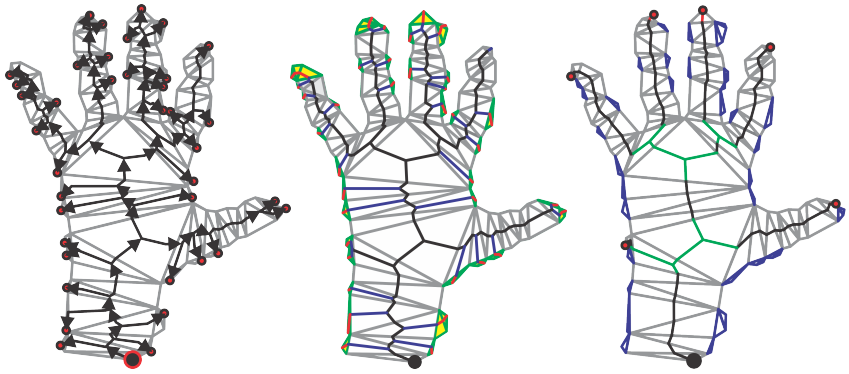
de cada galho terminal $q_g = \frac{S_g}{\alpha_g}$ é utilizada como critério de desempate na forma

$$d = \begin{cases} q_1 < q_2 \rightarrow & Cut(S_1) \\ q_1 \geq q_2 \rightarrow & Cut(S_2) \end{cases}$$

onde q_1, q_2 são as proporções dos galhos terminais relativos ao filhos S_1 (mais a esquerda) e S_2(mais a direita) da respectiva conexão, como definidos previamente; e $Cut(S)$ é a função de poda de galhos terminais descrita a seguir.

Poda de um galho terminal: Em posse de um nó pertencente a um galho terminal da árvore, a função de poda percorre cada um de seus filhos sucessivamente, chegando até a sua folha. A partir dai, o caminho de retorno para o seu nó pai é registrado, e o nó atual é removido da memória, juntamente com o seu respectivo triângulo. Esse processo se repete então para o seu nó pai, terminando ao encontrar o primeiro nó contendo uma bifurcação.

Remoção do nó central: Já que o triângulo onde se localiza a bifurcação passa a ter apenas dois vizinhos, o nó central deve ser retirado de modo que esqueletização passe a ter apenas um seguimento nesta região; para este fim são coletados os endereços de seu nó pai e de seu nó filho restante; o nó central é então excluído da memória e o seu endereço é substituído nesses dois outros, onde o primeiro passa a ser pai do segundo e o segundo filho do primeiro respectivamente. A Figura 4.26c exibe a primeira interação da poda na árvore gerada a partir do contorno de exemplo (que deverá ser reiniciada, já que ainda contém um galho terminal a ser retirado) onde os triângulos com arestas em azul foram removidos da triangulação inicial, assim como os seus respectivos segmentos.



(a) Folhas encontradas **(b)** Avaliação de galhos **(c)** Primeira interação

Figura 4.26 – Fases do processo de poda
Fonte: Elaborada pelo autor

4.2.2.3 Eigengestures

De forma análoga a*o Eingenfaces* (205, 206), esse método tem como finalidade obter generalizações de classes de imagens através de um conjunto de autovetores produzido a partir das amostras de aprendizagem. Tendo sido utilizado esse mesmo algoritmo, mas com nome alterado devido a sua nova aplicação (172), está seção se decida a descrição da história e do funcionamento do método original.

Considerado como a primeira tecnologia eficiente para o reconhecimento de faces, o algoritmo *Eigenfaces* é um princípio largamente utilizado por grandes sistemas computacionais privados para a identificação de seres humanos (207). Mais especificamente, o termo *eigenface* é utilizado para denominar os componentes principais de uma distribuição de probabilidade de imagens faciais, através da utilização do algoritmo *Principal Component Analysis* (*PCA*) (208, 209).

O principio básico do método *PCA*, consiste em transformar um grupo de variáveis possivelmente correlacionadas, em um conjunto menor de variáveis não relacionadas. Para este fim é pressuposto que o conjunto de dados de alta dimensão a ser analisado é descrito por diversas variáveis correlacionadas, de modo que a maior parte de suas informações pode ser representada utilizando uma quantidade de dimensões inferior. Dessa forma o método *PCA* obtém as direções com a maior variância no conjunto de dados em questão, onde estas são denominadas como seus componentes principais, o que consiste nos autovetores correspondentes aos maiores autovalores de sua matriz de covariância.

Essa matriz pode ser gerada então através da distribuição de probabilidade no espaço de alta dimensionalidade das imagens, onde considera-se que cada um de seus *pixels* corresponde a uma dimensão nesse espaço. Os autovetores resultantes formam assim um conjunto gerador linearmente independente, capaz de representar as imagens originais. Desse forma, esse algoritmo pode então ser descrito através dos seguintes passos:

Seja $X = \{x_1, x_2, ..., x_n\}$ um vetor aleatório com $x_i \in \mathbb{R}^d$

1. Calcule a sua média, na forma

$$\mu = \frac{1}{n} \sum_{i=1}^{n} (x_i)$$

2. Compute sua matriz de covariância

$$S = \frac{1}{n} \sum (x_i - \mu) (x_i - \mu)^{T'}$$

3. Compute os autovalores λ_i e os autovetores v_i de S, como

$$Sv_i = \lambda_i v_i,\ i = 1, 2, \ldots, n$$

4. Ordene os autovetores de forma decrescente em relação a cada autovalor correspondente

   (a) Os k componentes principais são os autovetores correspondentes aos k autovalores de valor mais alto

   (b) Os k componentes principais do vetor x observado são dados por: $y = W^T (x - \mu)$, onde $W = (v_1, v_2, \cdots, v_k)$

Dessa maneira a reconstrução da base do PCA é dada por: $x = Wy + \mu$.

No entanto, a solução deste problema pode ser inviável, dado o número selecionado de imagens por classe para o treinamento, e o tamanho padronizado para cada imagem do sistema. Uma vez que para 60 imagens com dimensões de, por exemplo 128×128 *pixels* (conforme descrito na Seção 4.2.3.2), é gerada uma matriz de covariância na forma $S = XX^T$, sendo a quantidade de *pixels* em X igual a $60 \times 128 \times 128$, de forma que a matriz resultante contém então 983.040 elementos. No entanto uma vez que qualquer matriz de $M \times N$ dimensões, com $M > N$ apresenta no máximo N–1 autovalores diferentes de zero, é realizada a decomposição dos autovalores $S = X^T X$ de tamanho $N \times N$, na forma:

$$X^T X v_i = \lambda_i v_i$$

, e assim são obtidos os autovetores originais de $S = XX^T$, por meio de uma multiplicação à esquerda da matriz de dados, como:

$$XX^T (Xv_i) = \lambda_i (Xv_i)$$

. Uma vez que os autovetores resultantes desse processo são ortogonais, os dados a ser comparados devem ser então normalizados em tamanho unitário.

Afim de obter generalizações para a identificação das classes, é criada então uma nova matriz covariância, representando a variação de todos os *eigenfaces* provenientes de suas imagens. Por fim, através da seleção do subconjunto de *eigenfaces* que apresente os maiores autovalores, obtém-se um coleção de características capaz de representar as maiores variâncias presentes nas imagens de sua respectiva categoria.

Desta forma esses autovetores correspondem à solução de um problema de mínimos quadrados (210), que garantem a manutenção da variância dos dados eliminando correlações

desnecessárias entre as características originais. Esse processo promove uma redução da dimensão do dados de aprendizagem, permitindo o armazenamento de uma quantidade reduzida de imagens para representar as informações de aprendizagem, já que os *eigenfaces* resultantes são capaz de descrever a variação global de cada classe desejada.

4.2.2.4 Fishergestures

Uma vez que é utilizado neste trabalho para o reconhecimento de gestos manuais, o nome que descreve esse método foi instituído para se adequar a essa nova aplicação. Mas já que se trata da aplicação do algoritmo *Fisherfaces*, desenvolvido originalmente para o reconhecimento de faces (211), também será descrita nesta seção um pouco do histórico e funcionamento do método original.

Desenvolvida como um aprimoramento do método *Eigenfaces*, o esses algoritmos se utilizam de conceitos bastante similares para a generalização de imagens de faces humanas. A principal diferença entre essas duas abordagens, consiste no princípio empregado por elas para sintetização de dados de aprendizagem. Inicialmente é realizada uma análise de componentes principais (208, 209) em cada imagem de treinamento, obtendo-se assim seus respectivos autovetores. O que corresponde aos mesmos *eigenfaces* utilizados pelo método anterior. No entanto, para encontrar uma representação mais adequada do subespaço de uma classe específica, é utilizado o método *Linear Discriminant Analysis* (212), inserindo como parâmetros de entrada todos os seus autovetores correspondentes.

Para encontrar a combinação de características que melhor representam e diferenciam cada classe analisada, esse método maximiza a proporção de dispersão entre classes para intra classes, em vez de maximizar a dispersão total. Em suma, são selecionadas para representar cada classe características capazes de agrupá-las, e ao mesmo tempo diferenciá-las o máximo possível das outras classes analisadas. Esse algoritmo pode ser então obtido da como:

Seja X um vetor aleatório com amostras retiradas de c classes:

$$\begin{aligned} X &= \{X_1, X_2, \ldots X_c\} \\ X_i &= \{x_1, x_2, \ldots, x_k\} \end{aligned}$$

, e a média total

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

; suas matrizes de dispersão correspondentes a

$$\begin{aligned} S_B &= \sum_{i=1}^{c} N_i \left(\mu_i - \mu\right)\left(\mu_i - \mu\right)^T \\ S_W &= \sum_{i=1}^{c} \sum_{x_j \in x_i} \left(x_j - \mu_i\right)\left(x_j - \mu_i\right)^T \end{aligned}$$

, onde μ_i corresponde a média de cada classe, de forma que $i \in \{1, \ldots, c\}$, calculada por

$$\mu\mu_i = \frac{1}{|X_i|} \sum_{x_j \in x_i} x_j$$

. A partir desses dados, é então obtida a projeção que maximiza o critério de separabilidade da classe

$$W_{opt} = \arg\max_w \frac{\left|W^T S_B W\right|}{\left|W^T S_w W\right|}$$

, sendo uma solução para este problema de otimização (211), a resolver o problema de autovalor geral

$$\begin{aligned} S_B v_i &= \lambda_i S_w v_i \\ S_W^{-1} S_B v_i &= \lambda_i v_i \end{aligned}$$

. No entanto não é possível encontrar essa solução por métodos convencionais, com base nos dados utilizados, já que a matriz de espalhamento S_W se torna singular; esse fato é proveniente da quantidade de amostras c ser menor que a dimensão dos dados de entrada N (quantidade de *pixels* contida em cada imagem), de modo que o *rank* de S_W é no máximo $(N–c)$, dados N amostras e c classes.

Para resolver esse problema é utilizado novamente o método de *Principal Component Analysis* nos dados obtidos, e as amostras utilizadas são projetadas no espaço dimensional $(N - c)$. Deste modo o algoritmo *Linear Discriminant Analysis* pode então processar os dados reduzidos, uma vez que estes não correspondem mais a uma matriz singular. O problema de otimização anterior pode então ser reescrito como

$$\begin{aligned} W_{\text{pca}} &= \arg\max_W \left|W^T S T W\right| \\ W_{\text{fld}} &= \arg\max_W \frac{W^T W_{\text{pca}}^T S_B W_{\text{pca}} W}{W^T W_{\text{pca}}^T S_W W_{\text{pca}} W} \end{aligned}$$

A matriz de transformação W, que projeta uma amostra no espaço dimensional $(c–1)$ é então dada por:

$$W = W_{\mathrm{fld}}^{T} W_{\mathrm{pca}}^{T}$$

.

Os vetores resultantes são denominados então como *fisherfaces*, onde estes correspondem a um conjunto gerador linearmente independente, capaz de definir esse subespaço. Apesar do potencial superior de representação do algoritmo anterior, a introdução deste novo método de generalização é justificado pelo fato de que a solução de mínimos quadrados produz resultados menos efetivos do que o método proposto, quando utilizado especificamente para promover classificações.

Isso ocorre devido a capacidade superior da análise de discriminante linear para a produção de maiores diferenciações entre as generalizações de classes. Essa diferenciação ocorre por meio da concentração dos dados relativos a uma classe especifica em um ponto do espaço n-dimensional[†], concomitante ao aumento da separação dos dados referentes às representações de classes distintas.

4.2.3 Classificação de poses

Visando a classificação das poses manuais realizadas pelos usuários, foram pesquisados diversos algoritmos capazes de relacionar as informações fornecidas pelos métodos desenvolvidos, e agregá-las em grupos conforme as suas similaridades. Uma vez que não encontrou-se um conjunto de regras preestabelecidas para resolver este problema, optou-se por uma abordagem mista, capaz de resolver o problema com um desempenho satisfatório.

Dentre muitas metodologias disponíveis para esta finalidade, conforme pode ser visto na Seção 2.4.2.3, a seleção do tipo de dado empregado para o treinamento, assim como a escolha da categoria de classificador a ser utilizado para a identificação das poses, foram realizadas com base nos requisitos pré-determinados para o sistema como um todo.

As características do sistema proposto, às quais esta decisão exerce influência, podem então ser citadas em ordem de prioridade como: eficácia na discriminação de um conjunto abrangente de poses, eficiência adequada a interação em tempo real, extensibilidade à aprendizagem de novas poses, e portabilidade para sistemas de pequeno porte. A seguir são especificadas as abordagens escolhidas para a classificação, com foco nos atributos supracitados.

[†]Onde n é quantidade de píxels da imagem resultante

4.2.3.1 Abordagem para a classificação:

Uma vez que todas as poses manuais registradas na base desenvolvida para este trabalho são previamente conhecidas, e que a interface projetada não tem como finalidade deduzir relações ocultas entre elas; a abordagem baseada em treinamento supervisionado foi selecionada como a mais adequada para o método de aprendizagem.

Além da escolha do modo treinamento, uma classificação hierárquica (213) foi desenvolvida, por meio da utilização das características extraídas anteriormente. Dessa maneira foram obtidas pré-seleções dos dados de treinamento e testes, possibilitando que os métodos pesquisados sejam utilizados com reduções significativas de processamento.

Devido à necessidade de se obter um equilíbrio entre os requisitos supracitados, pesquisou-se por algoritmos de classificação baseados em generalização de características, assim como métodos de transdução baseados nas generalizações das amostras de dados.

4.2.3.2 Metodologia de avaliação

A partir da base de poses manuais produzida para este trabalho, foram selecionadas aleatoriamente 60 amostras pertencentes a cada uma das 26 classes registradas, correspondente a um total de 1.560 imagens de poses de mãos utilizadas para a avaliação dos métodos de classificação propostos. Além disso todas as imagens foram escalonadas para as dimensões de 128×128, afim permitir a sua utilização pelos sistemas de classificação de imagens utilizados.

Por sua vez, as amostras pertencentes a cada classe específica foram extraídas de seus cinco vídeos correspondentes, gravados pelos respectivos voluntários, equivalendo a doze imagens para cada uma das 130 gravações realizadas. De forma a evitar a introdução de amostras muito semelhantes, as imagens provenientes de uma mesma gravação foram coletadas em intervalos de vinte *frames*, assegurando assim uma alta variedade para os dados de treinamento e teste, bem como uma maior confiabilidade dos resultados obtidos.

A partir das amostras extraídas, todos os testes foram realizados por validação cruzada (214), dividindo a base de poses em cinco pastas distintas, de forma que as 60 amostras provenientes de uma mesma classe foram distribuídas nessas pastas conforme a sua ordem de extração. Por fim, os resultados das avaliações foram então calculados como as médias dos valores obtidos para cada classificador específico, onde foram realizadas cinco avaliações por

classificador proposto, selecionando uma dessas cinco pastas a cada vez como base de teste e todas as pastas restantes como base de treinamento.

4.2.3.3 Tipos de dados utilizados para a classificação:

Constatada a insuficiência de informações para a diferenciação das poses registradas, provenientes da identificação dos membros das mãos (Seção 4.2.2); optou-se pelo uso de imagens de profundidade (Seção 4.2.1.1) e técnicas de generalização, para o treinamento do classificador implementado.

A seguir são listadas as razões para o emprego dessas imagens e generalizações nas fases de treinamento e identificação das poses, as informações necessárias à sua classificação, além do método de estruturação dos dados aplicado para a utilização de imagens nos algoritmos de aprendizado de máquina pesquisados.

Informações para a diferenciação de poses: Analisando as poses manuais selecionadas para a interação com o sistema desenvolvido, Figura 4.1, é notória a similaridade entre várias classes distintas. Muitas vezes a diferença entre elas consiste em uma pequena mudança na posição de um dedo, na presença ou ausência de buracos em sua silhueta, ou projeções de *pixels* que promovem diferenças de profundidade, mas que se encontram internas ao contorno da mão.

Um vez que a localização dos membros das mãos provém de uma única linha poligonal fechada, obtida a partir da binarização dessas imagens, todas as informações presentes em seu interior são desconsideradas; o que inviabiliza a discriminação de poses onde qualquer uma dessas características é necessária à sua diferenciação.

Entre outros exemplos: utilizando apenas esses dados, não é possível verificar a posição de dedos que apresentem proeminências internas ao contorno, como nas poses: K, M, N e Q. De forma similar, classes que muitas vezes só são diferenciáveis pela presença ou ausência de buracos, como: O e S, ou D e I, também não podem ser reconhecidas corretamente. Por fim, as poses onde uma pequena mudança na ordem de profundidade dos dedos consistem em sua única diferença, como: F e T, também não podem ser avaliadas com base nos dados em questão.

No entanto as imagens de profundidade relativas às poses de mãos, assim como as suas generalizações, contém todas as informações registradas em cada pose, e já se encontram

disponíveis para serem utilizadas no treinamento de classificadores, de forma invariante à posição, rotação e escala. Embora a utilização desse tipo de dado seja geralmente menos eficiente, e consuma mais memória do que as técnicas de rastreamento; verificou-se que ele consiste no único tipo de informação obtida pelo sistema que contém todas as características necessárias para a diferenciação de cada pose, além de ser facilmente aplicável às fases de treinamento e teste de muitos métodos de aprendizagem de máquina.

Estruturação dos dados: Afim de utilizar um conjunto de dados para o treinamento dos classificadores pesquisados é necessário inserir as informações pertencentes a cada amostra em vetores com o mesmo tamanho. Assim, cada informação presente em um índice específico de um desses vetores, deve manter uma relação com todos os outros dados contidos nas células correspondentes de todos outros vetores obtidos.

Para as imagens de profundidade de gestos manuais, assim como às suas generalizações, esta tarefa pode ser realizada escalando todas elas para um mesmo tamanho, e posteriormente inserindo sequencialmente cada linha pertencente a estas em um vetor contendo a mesma quantidade de células da imagem resultante, em conjunto com um rótulo utilizado para descrever a sua classe.

Os outros dados extraídos não podem ser organizados desta mesma forma, por conta da alta variabilidade de suas informações. Essa variabilidade se deve ao fato de que nem sempre é possível identificar qual dedo corresponde a uma terminação específica, ou mesmo quantos desses estão presentes em uma única terminação encontrada. No caso específico da esqueletização baseada em triangulação (descrita em detalhes na Seção 4.2.2.2) o número de informações presentes em cada terminação também é um agravante desta condição, já que as linhas que as representam são formadas por quantidades variáveis de pontos.

4.2.3.4 Classificação hierárquica

Esse algoritmo foi desenvolvido visando o aprimoramento da interface proposta, de forma a melhorar a precisão e a velocidade dos métodos de classificação. Uma vez que as características encontradas para alguns grupos de classes possibilitam várias subdivisões da base de treinamento, os métodos de aprendizado de máquina (utilizados para na última fase deste processo) tendem a apresentar maiores taxas de acerto, já que desta forma necessitam discriminar a pose em questão sob quantidades de classes reduzidas. Do mesmo modo a velocidade

para a classificação de poses também tende a aumentar, uma vez que o espaço de busca por características é reduzido.

Tendo sido constatado que o algoritmo de decomposição de casco convexo (descrito em detalhes na Seção 4.2.2.1) apresenta uma taxa de acerto razoável (Seção 5.1.1) para encontrar a quantidade de pontas de dedos, da maioria das poses manuais propostas; uma divisão inicial da base de treinamento foi realizada a partir dos dados obtidos pela aplicação deste processo nas imagens da base registrada, e em testes interativos com o sistema em funcionamento.

Uma vez que são retornados por este método apenas valores naturais entre zero e cinco, equivalendo ao número de dedos que estejam provocando protuberâncias na silhueta de uma imagem de pose manual, as imagens da base de aprendizagem original foram separadas em seis subconjuntos; correspondendo a quantidade de pontas de dedos mais frequentemente obtida para as amostras pertencentes a cada classe.

Já que para algumas dessas classes de poses foram obtidas variações frequentes nos valores encontrados, se tornou necessária a inserção dessas em mais de uma das subdivisões, para que desta forma pudessem ser corretamente classificadas posteriormente. Em contra partida, duas das seis subdivisões obtiveram apenas uma classe de pose manual, o que significa que se tornou desnecessário o armazenamento de amostras de treinamento relativas a estas. A Tabela 4.1 ilustra a subdivisão da base de treinamento, indicando a quantidade de pontas de dedos correspondente a cada uma delas, e as classes das amostras a serem inseridas para o treinamento dos classificadores.

Tabela 4.1 – Primeira subdivisão das poses para a classificação hierárquica

Pontas de dedos encontradas	Classes de poses manuais
0	A B E M N O Q S U X
1	$\rightarrow$ C D G I R U P
2	K L P V Y
3	F T W
4	4
5	∞

Fonte: Elaborada pelo autor

Através de avaliações similares à realizada para a decomposição de casco convexo, verificou-se que o método de esqueletização baseada em triangulação também obtém uma acurácia razoável para encontrar a quantidade de terminações de cada amostra pertencente a uma classe de pose manual, conforme descrito na Seção 5.1.2. Deste modo, as imagens de poses foram subdivididas em novas categorias, conforme a quantidade de terminações mais frequentemente apresentada para cada uma de suas classes.

A Tabela 4.2 exibe esta nova subdivisão, indicando a quantidade de pontas de dedos encontrada previamente, o número de terminações que as poses analisadas apresentam mais frequentemente, e a separação final de suas imagens nas novas bases de treinamento correspondentes.

Tabela 4.2 – Segunda subdivisão das poses para a classificação hierárquica

Pontas de dedos encontradas	Terminações	Classes de poses manuais
0	0	A B E M N O Q S
	1	X U
1	1	→ D G I U
	2	C P

Fonte: Elaborada pelo autor

Dada a subdivisão da base, a classificação das imagens de poses manuais é então realizada de forma análoga; onde inicialmente é feita a contagem de suas pontas de dedos, e dependendo do resultado são analisadas a quantidade de terminações encontradas. A última fase dessa classificação consiste então na utilização de sistemas de reconhecimento com base em imagem, conforme descrito na Seção 4.2.3.5.

A Tabela 4.3 exibe todas as fases desse método, conforme as características encontradas para uma amostra de testes específica. Além disso também é indicado se existe a necessidade da aplicação de classificadores baseados em imagens para a sua identificação, e a quantidade de classes que deverão ser consideradas durante a classificação dependendo dos dados obtidos anteriormente.

Tabela 4.3 – Fases da Classificação

Pontas de dedos	Terminações	Classificação por imagens	Classes
0	**0**	SIM	8
	1	SIM	2
1	**1**	SIM	5
	2	SIM	2
2	-	SIM	5
3	-	SIM	3
4	-	NÃO	1
5	-	NÃO	1

Fonte: Elaborada pelo autor

4.2.3.5 Método de classificação baseados em imagens

Visando obter classificações eficazes e eficientes das poses de mãos, foram analisados dois métodos bastante similares de aprendizado de máquina, afim de encontrar a melhor solução para este problema. Entre outros fatores, os testes realizados com esses algoritmos foram aplicados, afim de comparar o funcionamento do sistema como um todo, sob diferentes abordagens de reconhecimento de imagens.

Além do clássico algoritmo transdutivo *KNN* (215), as generalizações produzidas anteriormente para cada classe de pose foram utilizadas, no intuito de avaliar a eficácia de sistemas indutivos para a resolução deste problema e promover uma melhor performance para a interface implementada.

Essa avaliação partiu do pressuposto que: enquanto os métodos de reconhecimento por generalização de dados costumam necessitar de custos computacionais bem menos elevados (69), beneficiando a eficiência e a portabilidade da interface em tempo real; a utilização de classificadores baseados em transdução tende a ser mais adaptável a adição de novas classes, sem comprometer a aplicabilidade da avaliação do sistema (216), promovendo assim maior eficácia e extensibilidade do método proposto em cenários com uma quantidade elevada de poses manuais a ser reconhecida.

Dessa forma, essas técnicas de generalização de imagens para reconhecimento de faces foram utilizadas, devido a trabalhos anteriores utilizando abordagens similares para a identificação de poses de mãos humanas (173), e da constatação de semelhanças estruturais observadas entre as imagens de ambos os grupos; como por exemplo: alta deformação, necessidade de alinhamento, e uso de classes pré-definidas. Por sua vez, o algoritmo *KNN* foi selecionado como representante dos métodos indutivos, por conta de sua alta flexibilidade e robustez para a classificação de diversos tipos de dados.

KNN: O algoritmo K *Nearest Neighbor* (ou abreviadamente *KNN*) apresenta um funcionamento elementar, dado a sua eficácia e versatilidade para a discriminação de dados em várias aplicações. Além de ser bastante utilizado para a classificação, este método transdutivo é também aplicável à análise de regressão, onde em ambos os casos seus parâmetros de entrada consistem em vetores de características contendo as informações a serem utilizadas para o treinamento.

Sendo um algoritmo transdutivo bastante simples, sua fase de treinamento é considerada

como mínima ou inexistente. Já que não produz generalizações acerca dos dados analisados, esta etapa consome pouco processamento, consistindo basicamente na inserção de informações na memória do sistema. A robustez e a eficácia desse método se devem ao fato de que a avaliação dos dados de testes é realizada através de comparações diretas com as amostras de treinamento. Dessa maneira, são minimizados os erros ocorridos em classificações onde os dados pertencentes a uma mesma categoria são muito heterogêneos, e por isso não podem ser utilizados para produzir generalizações adequadas à sua diferenciação.

Apesar de seus benefícios, este método costuma promover um uso significativo de memória para o seu funcionamento, já que depende da manutenção dos dados supracitados durante a classificação. Além do alto consumo de memória, sua eficiência para a classificação também costuma ser baixa, já que no pior caso todas as amostras analisadas podem interferir no resultado final. Com foco em sua aplicação para a interface proposta, segue uma explicação sucinta das fases de execução do *KNN* para a classificação das imagens de poses de mãos obtidas anteriormente.

Base de treinamento: conforme descrito na Seção 4.2.3.3, cada amostra de treinamento é rotulada afim representar a classe a qual pertence, juntamente com um vetor contendo n células, correspondendo a quantidade de *pixels* fixada para as imagens. Todos esses dados são então inseridos na mé moria do sistema, de forma a acelerar futuras buscas com base em suas similaridades.

Parâmetros para a classificação: após o treinamento, o mesmo processo de estruturação é realizado em tempo real para as imagens a serem discriminadas, com a diferença de que neste caso o objetivo do classificador consiste em encontrar o rótulo mais adequado para elas. O último parâmetro a ser inserido corresponde a um número inteiro K, que representa a quantidade de amostras de treinamento a ser avaliada, para cada atribuição de rótulo à uma imagem pose de mão fornecida pela interface.

Seleção dos vizinhos: afim de que sejam obtidas as K amostras da base de aprendizagem que sejam as mais similares a uma imagem de pose manual qualquer; é realizada uma busca dentre as imagens de treinamento, com base no cálculo da distância euclidiana entre elas. Essa distância é obtida, através da interpretação dos valores inseridos nos vetores de características como sendo coordenadas em um espaço de n dimensões.

Atribuição de rótulos: a última fase desse método, consiste na atribuição de um rótulo à imagem de teste, representando uma classe de pose manual válida do sistema. Esse rótulo é então obtido através de um processo baseado em eleição majoritária, utilizando

como unidade de voto as ocorrências de cada classe de pose nas K amostras selecionadas. Computados esses votos, o sistema infere como resultado a classe que obtiver a maior quantidade de votos, atribuindo por fim o rótulo resultante à imagem analisada.

A Figura 4.27, demonstra o funcionamento do *KNN* para uma pequena quantidade de classes pré-determinadas e de atributos presentes em seus vetores de características. Neste exemplo, dois valores descrevem os dados das amostras de treinamento, onde cada uma destas é rotuladas como A ou B.

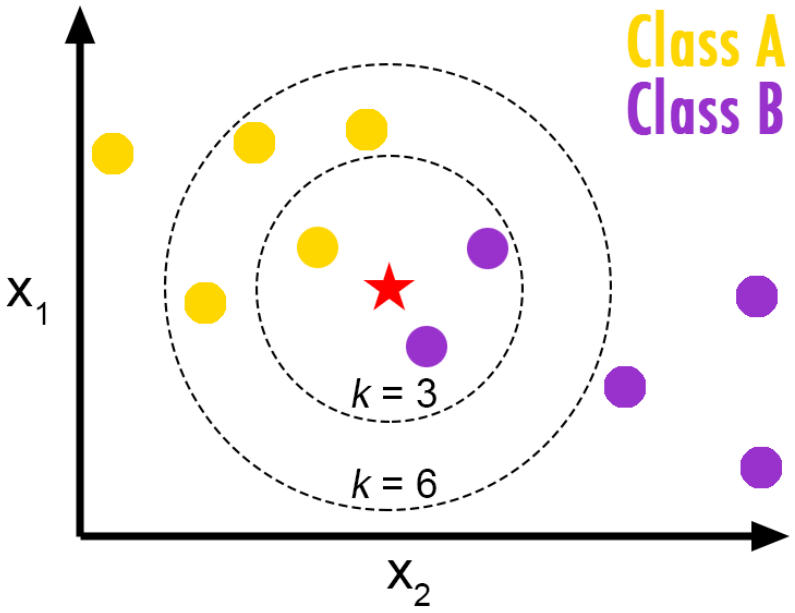


Figura 4.27 – Classificação por votação majoritária
Fonte: DeWilde, B (217)

Considerando os seus atributos como coordenadas nos eixos X_1 e X_2, as características das amostras de treinamento e testes podem ser representadas por sua posição em um plano cartesiano bidimensional. Conforme descrito anteriormente, esse método seleciona os K vizinhos que apresentem a menor distancia do seu objeto de classificação, de forma que a identificação do conjunto de dados representado como uma estrela, será realizada com base na posição dos K círculos mais próximos a ela. Dessa forma, para $K = 3$ e para $K = 6$, são obtidos os resultados B e A respectivamente, correspondendo a quantidade majoritária de seus elementos dentre os K vizinhos avaliados.

Comparação com as generalizações de classes: De acordo com o método desejado, podem ser selecionados para a discriminação de poses manuais os *eigengestures* (gerados conforme descrito na Seção 4.2.2.3), ou os *fishergestures* (produzidos como vistos na Se-

ção 4.2.2.4). Essas generalizações de imagens, correspondentes a cada classe de pose manual registrada, são utilizadas para a classificação das amostras de testes, através de uma comparação holística[‡].

Deste modo, cada imagem a ser avaliada é comparada com cada generalização de classe selecionada, de modo a obter seus respectivos valores de similaridade. Esse valor será próximo de zero nos casos em que a imagem avaliada seja semelhante a uma generalização específica, e valores maiores conforme estas sejam mais distintas.

Com base nesse resultado, o rótulo da generalização que apresente menor valor é então atribuído para a sua respectiva amostra de teste. Esse processo é considerado como um caso especial do algoritmo *KNN*, denominado como *1NN* (*First Nearst Neighbor*), que pode ser descrito sucintamente como um método de classificação pela seleção do vizinho mais próximo.

As comparações entre uma imagem qualquer e as generalizações na forma de *eigengestures* e *fishergestures,* podem então ser realizada pelas seus respectivos métodos de projeção através dos seguintes passos:

1. Projeção das generalizações de classes no subespaço *PCA* encontrado
2. Projeção da imagem a ser classificada nesse mesmo subespaço
3. Atribuição do rótulo pertencente a projeção de generalização de classe mais próxima da projeção da amostra a ser classificada, utilizando uma métrica de distância euclidiana no espaço n-dimensional das imagens.

[‡]Comparação da imagem como um todo, ou contrário à análise de partes específicas destas

Capítulo 5

Resultados

Além dos resultados numéricos, relativos ao desempenho do método desenvolvido, várias de suas características qualitativas e funcionais são avaliadas neste capítulo, afim de relatar adequadamente todas as suas capacidades. Visando uma descrição detalhada do desempenho obtido para as funcionalidade propostas, são especificados os resultados verificados para os métodos de rastreamento de membros das mãos, e algoritmos de reconhecimento de poses manuais implementados.

De modo a evidenciar as inovações científicas e tecnológicas do trabalho realizado, todos os dados obtidos são comparados com as respectivas informações, relatadas para outros métodos voltados ao reconhecimento de gestos manuais, cujo reconhecimento acadêmico foi previamente comprovado.

Por fim, todos os testes referentes ao desempenho dos métodos descritos foram realizados em tempo real, por meio de um *Kinect* para *XBOX* 360, e um computador *desktop* contendo um processador *Intel Core* 2 *Quad* de 2.66 GHz, com memória *RAM* utilizável de 3,24 *GB*, *HD* de 438,7 *GB*, e portas de comunicação *USB* 2.0 dedicadas. No entanto, já que a capacidade computacional não exerce influência sobre a eficácia dos processos utilizados por esse sistema, essa característica pode ser avaliada utilizando um conjunto de imagens previamente extraídas da base registrada, como visto na Seção 4.2.3.2.

5.1 Avaliação dos métodos de rastreamento

Uma vez que decidiu-se por não fazer anotações na base registrada referente às posições dos dedos e das mãos dos usuários, conforme visto na Seção 4.1, verificou-se por observações empíricas realizadas através da visualização do sistema implementado em funcionamento, que este é capaz de obter a posição das mãos e dedos com acurácia e eficiência, apesar de não utilizar técnicas preditivas (218–220) comumente aplicadas em softwares de rastreamento por visão computacional. No entanto, já que o rastreamento das mãos é feito através da esqueletização de usuário da *OpenNI*, o sistema implementado tem influência direta da precisão

deste algoritmo, que tende a variar com a velocidade de movimentação do usuário, além de ser especialmente vulnerável a poses corporais específicas (221, 222).

Uma medida automatizada da acurácia quanto ao reconhecimento de posições deste método foi realizada por meio da taxa de acerto referente a quantidade e tipo de terminações encontrados. A Tabela 5.1 exibe o desempenho desses métodos, incluindo a velocidade de limiarização das imagens, que é utilizada em conjunto com todos os métodos descritos*.

Tabela 5.1 – Desempenho dos métodos de rastreamento de poses manuais

Método	Acurácia	Tempo por pose	Quadros por segundo
Limiarização 3D	Não avaliada	$5,791 \cdot 10^{-3}$ s.	172,674
Casco convexo	70,513%	$4,487 \cdot 10^{-3}$ s.	97,291
Esqueletização	78,647%	$6,730 \cdot 10^{-3}$ s.	79,859

Fonte: Elaborada pelo autor

Já que muitas interfaces de interação por gestos podem ser desenvolvidos com base na posição dos membros das mãos, conforme visto no Capítulo 2, a eficiência do rastreamento desses é avaliada independentemente para cada método utilizado com esta finalidade. Dessa maneira, uma vez que a velocidade de rastreamento sofre altas variações conforme a quantidade de processos ativos durante a execução do sistema, e que a localização dos membros da mão pode ser realizada tanto pela decomposição de casco convexo (descrita na Seção 4.2.2.1), quanto pela esqueletização baseada em triangulação (vista em detalhes na Seção 4.2.2.2), a rapidez de localização dos membros da mão é fornecida separadamente para estes dois casos, ressaltando as vantagens e desvantagens de cada um deles.

5.1.1 Decomposição de casco convexo

Sendo utilizado como a base para diversos métodos de reconhecimento de gestos manuais (20, 177, 178), esse método é aplicado na interface implementada para a localização e diferenciação das terminações de várias poses manuais. Uma vez que esse algoritmo consome recursos computacionais inferiores a esqueletização baseada em triangulação, ele tende a ser utilizado em máquinas mais simples para interações com elementos virtuais, sem a necessidade de utilização de quaisquer outros processos aqui descritos.

Conforme visto anteriormente, esse algoritmo busca por terminações proeminentes no contorno das mãos e às classifica de acordo as características encontradas em cada uma delas.

*No cálculo da quantidade de quadros por segundo de cada método é incluso o tempo dispendido para a limiarização tridimensional, sendo que esse montante não é limitado a taxa de atualização do sensor utilizado

Dessa forma é possível verificar a sua precisão para o rastreamento de pontas e aglomerados de dedos, apenas comparando a quantidade e o tipo de terminações encontradas para cada pose de mão, com as respectivas informações comumente obtidas para as classes a qual pertencem.

Configurando o sistema implementado para ativar somente esse módulo, foi possível verificar que por meio dele é possível processar uma imagem de pose manual padrão (Seção 4.2.3.2) em um tempo médio de $4,487.10^{-3}$ segundos; onde em conjunto com a limiarização tridimensional, pode ser utilizado para rastrear uma pose de mão a uma velocidade de $97,291$ quadros por segundo, ou metade desta quantidade de imagens para a localização em tempo real das duas mãos de um mesmo usuário; sendo esta velocidade limitada pela taxa de atualização do sensor utilizado.

Afim de obter a sua taxa de acerto para as classes registradas como um todo, o processo descrito acima é feito para cada imagem da base, onde a eficácia e_i para uma classe i é calculada como sendo a quantidade de poses com terminações equivalentes a esperadas r_i, divida pelo número total de poses analisadas q_i, ou seja $e_i = \frac{r_i}{q_i}$.

A taxa de acerto geral t, é então calculada como sendo a média das eficácias das n classes do sistema, na forma:

$$t = \frac{e_0 + e_1 + \ldots + e_n}{n}$$

, onde foi verificada uma porcentagem de acerto de 70,513% para este método em relação a base registrada.

5.1.2 Esqueletização baseada em triangulação

Uma vez que é capaz de descrever com precisão cada terminação encontrada, a esqueletização baseada em triangulação de contornos é mais indicada para um rastreamento detalhado das partes que compõem as mãos. Essa característica é relevante em sistemas onde a forma dessas terminações exercem maior influência no resultado final, já que a colisão com os objetos virtuais pode ser feita para com todos os vértices presentes em seus segmentos.

De maneira análoga a realizada para decomposição de casco convexo, a eficácia de rastreamento desse método é calculada com base na comparação do número de terminações encontradas para cada pose, em relação à quantidade de membros mais comumente identificada nas classes correspondentes.

Utilizando a mesma metodologia empregada para o método anterior, obteve-se então uma

taxa de acerto de $78,647\%$, como uma média da eficácia resultante da análise de cada classe individualmente. Processando apenas este algorítimo para localizar os membros das duas mãos, verificou-se ele consome $6,731 \cdot 10^{-3}$ segundos em média para a análise de uma única imagem, obtendo em conjunto com a limiarização tridimensional o desempenho de médio de $79,859$ quadros por segundo, no rastreamento de uma mão em tempo real.

5.2 Desempenho para o reconhecimento de poses

Uma vez que vários sistemas baseados na interface proposta podem necessitar apenas da identificação das poses manuais dos usuários para a sua execução, não necessitando assim das posições dos membros de suas mãos para quaisquer tarefas; cada classificador baseado em imagens foi avaliado individualmente, e em conjunto com a método hierárquico desenvolvido, visto anteriormente na Seção 4.2.3.4. Essa abordagem foi realizada afim de diferenciar adequadamente todas as metodologias de diferenciação de poses manuais pesquisadas, tanto quanto à sua acurácia para a classificação, como também à sua viabilidade para interfaces de usuário em tempo real.

Por meio do processo de validação cruzada, realizou-se os treinamentos e os testes dos algoritmos de classificação descritos, usufruindo de todas as imagens extraídas para este fim. O desempenho de cada método foi verificado com o sistema em funcionamento, onde a interface implementada foi aplicada na identificação de suas poses, de modo que cada gesto referente a uma classe proposta foi avaliado sequencialmente, afim de abranger todas as variações temporais provenientes do processamento de cada uma dessas.

Os valores obtidos desta forma, referentes a acurácia e o desempenho dos classificadores propostos é ilustrado na Tabela 5.2, onde se observa as variações dessas métricas conforme as técnicas utilizadas. Uma vez que para a classificação das imagens de profundidade, vista na Seção 4.2.3.5, encontrou-se $k = 1$ para o do método de classificação *KNN*, considera-se que para cada característica descrita utilizou o mesmo tipo de classificador. Nos métodos de classificação hierárquicos, está incluso o custo da esqueletização por triangulação e decomposição de casco convexo, no tempo dispendido para análise de cada pose manual. /

Tabela 5.2 – Desempenho dos métodos de reconhecimento de poses manuais

Método	Acurácia	Tempo por pose	Quadros por segundo
Imagens de profundidade	$83,013\%$	$8,269 \cdot 10^{-2}$ s.	12,093
Imagens hierarquizadas	$59,359\%$	$9,615 \cdot 10^{-3}$ s.	37,559
Eigengestures	$81,539\%$	$2,506 \cdot 10^{-1}$ s.	3,900
Eigengestures hierárquico	$69,744\%$	$7,821 \cdot 10^{-2}$ s.	10,502
Fishergestures	$35,641\%$	$7,051 \cdot 10)^{-3}$ s.	77,866
Fishergestures hierárquico	$66,603\%$	$9,615 \cdot 10^{-3}$ s.	37,559

Fonte: Elaborada pelo autor

5.3 Comparações entre sistemas para reconhecimento de gestos manuais

Afim de enfatizar a contribuição da interface implementada para as áreas de interação usuário computador e visão computacional, foram analisados vários sistemas para identificação de gestos manuais, quanto às suas funcionalidades, eficácias, desempenhos, recursos utilizados e variedades de poses identificadas.

No entanto, entre diversos softwares, kits de desenvolvimento, e bibliotecas de função com a finalidade de rastreamento e reconhecimento de gestos manuais (223–230) disponíveis na Internet, foram encontrados apenas três métodos que contam com divulgações científicas reconhecidas (54, 122, 231).

O conjunto de tarefas realizado por cada uma desses sistemas, juntamente com as funcionalidades do software implementado, podem ser visto na Tabela 5.3, correspondendo ao rastreamento dos membros das mãos, a classificação de suas poses, e a análise das variações espaçotemporais para identificação de gestos.

Tabela 5.3 – Tarefas realizadas por cada sistema avaliado

Aviso: as condições de comparações com outros métodos não foram as mesmas			
Método	**Identificação de membros**	**Classificação**	**Temporal**
Método proposto	SIM	SIM	NÃO
SigmaNIL	SIM	SIM	SIM
Kinect 3D Hand tracking	SIM	NÃO	NÃO
HandGKET	NÃO	NÃO	SIM

Fonte: Elaborada pelo autor

Embora esses sistemas realizem tarefas relativamente distintas, a funcionalidade básica

descrita para cada um deles consiste no rastreamento das mãos, e na possibilidade de reconhecimento dos gestos manuais de seus usuários, a partir dos dados obtidos.

Um conjunto de indicadores referentes ao desempenho desses softwares pode ser visto na Tabela 5.4, onde esses valores estão relacionados a realização das tarefas a que se propõem, sendo estes: a sua acurácia para o reconhecimento de gestos, o número de poses analisadas, a velocidade para a realização dessas tarefas[†], e se é aplicado o uso de *GPU´s* para o seu funcionamento básico.

Tabela 5.4 – Indicadores pesquisados para cada sistema avaliado

Aviso: as condições de comparações com outros métodos não foram as mesmas				
Método	**Acurácia**	**Poses**	**Eficiência**	***GPU***
Método proposto	66,6%	26	$9,615 \times 10^{-3}$s/imagem	Não
SigmaNIL	82.1%	40	$4,5 \times 10^{-3}$s/imagem	Não
Kinect 3D Hand tracking	74%	Não se aplica	15 Hz	SIM
HandGKET	94.2%	17	Não especificado	Não

Fonte: Elaborada pelo autor

Apesar dos números apresentados serem úteis para a compreensão do potencial de cada interface analisada, eles não servem como método de comparação entre elas, uma vez que são expostos nesse trabalho conforme foram fornecidos pelos autores de cada ferramenta. Além da diferença de velocidade entre as máquinas utilizadas, também é de grande influência para o resultado desses métodos as bases para treinamento e testes utilizadas para a validação desses resultados. Na Seção 6.1 este problema é descrito com mais detalhes, juntamente com uma sugestão de trabalho para a sua resolução.

[†]A velocidade do método é especificada conforme descrito nos artigos relacionados

Capítulo 6

Conclusão e trabalhos futuros

Conforme a proposta desse trabalho, foi desenvolvida uma interface genérica e abrangente para detecção de poses manuais, com uma precisão adequada para o reconhecimento de uma quantidade significativa de gestos, e robustez para o rastreamento dos membros das mãos de seus usuários. Desse modo, ficou comprovada a efetividade das metodologias baseadas em extração de características (176) para a identificação de poses manuais, onde são destacados os diversos benefícios da utilização de sensores de profundidade (98) para a identificação de gestos (27). Em especial, a utilização do sensor de profundidade selecionado foi de grande auxílio o obtenção de um método eficaz de reconhecimento e rastreamento de poses. Uma vez que o rastreamento das mãos e cotovelos é fornecido de previamente, é que as informações de profundidade simplificam a segmentação das imagens com invariância a iluminação.

Verificou-se que dentre as características e métodos de classificação utilizados, as imagens de profundidade sem hierarquização tendem a apresentar melhores resultados quando deseja-se maior eficácia de classificação e dispõe-se de recursos computacionais mais poderosos; enquanto o *fishergestures* com hierarquização apresenta um desempenho mais satisfatório, embora apresente uma acurácia inferior. Verifica-se também que o sistema apresenta acurácia e o desempenho satisfatórios para o rastreamento de poses, sendo estes resultados similares quanto a utilização da decomposição de casco convexo, e a esqueletização baseada em triangulação.

Além disso, esse sistema apresenta alta portabilidade, podendo ser utilizado através de diversos tipos de sensores de profundidade e plataformas computacionais; já que têm como base para o seu funcionamento um conjunto de bibliotecas de função (13, 174) voltadas para este fim. Ademais, uma vez que muitos de seus métodos são independentes, eles podem ser aplicados em diferentes propósitos e condições de funcionamento, possibilitando que sejam configurados de acordo com as necessidades específicas de cada software, e que os seus custos computacionais sejam assim adequados para interação em tempo real em vários dispositivos eletrônicos.

Não tendo sido encontrada nenhuma base de dados com informações compatíveis, outra contribuição importante deste trabalho consiste nos registros disponibilizados contendo

informações adequadas a avaliações, treinamentos e comparações de sistemas para o reconhecimento de gestos manuais. Acredita-se que essas imagens poderão ser utilizadas futuramente para comparações entre sistemas com finalidades similares ao desenvolvido para este trabalho, uma vez que além de informações de profundidade relativas a cada pose manual registrada, também são disponibilizadas nessa base suas respectivas imagens coloridas, concomitante a segmentações e esqueletizações de seus usuários.

6.1 Sugestões para o prosseguimento da pesquisa

Por meio dos subprodutos desse trabalho, três projetos acadêmicos distintos podem ser realizados; afim de avaliar as técnicas utilizadas por esta área de pesquisa, registrar informações mais acuradas sobre gestos manuais diversos, ou aprimorar os métodos descritos nessa monografia. A seguir são listadas as respectivas justificativas de suas importâncias científicas e tecnológicas, e um conjunto de sugestões para as suas realizações.

A primeira sugestão consiste na comparação de diversos algoritmos para o reconhecimento de gestos manuais por visão computacional a partir da base registrada, utilizando assim um mesmo conjunto de imagens em cada teste. Isto é importante, uma vez que a área de reconhecimento de gestos manuais (6, 15, 26) necessita de uma avaliação acurada de seus métodos, onde deve ser utilizada uma base comum para a comparação entre os resultados obtidos. Essa técnica também conhecida como *benchmark* (232) é fundamental para que seja realizada uma avaliação justa entre sistemas computacionais com finalidades similares.

Essa proposta é relevante, devido ao fato de que o reconhecimento de gestos manuais conta atualmente com vários sistemas computacionais implementados, assim como um grande número de métodos acadêmicos publicados; onde geralmente esses algoritmos são complexos e extensos, e não sãos disponibilizados os seus códigos fonte para alterações. Embora a acurácia de muitos métodos esteja devidamente documentada, uma comparação direta entre esses valores não é adequada por vários motivos, entre esses é possível citar:

- Utilizam de um conjunto de dados próprio, para o seu treinamento e validação;
- Se propõem ao reconhecimento de gestos em quantidade, formas e variações temporais distintas;
- Cada software avaliado dispõe de formatos particulares de arquivo, geralmente incompatíveis entre si.

A segunda sugestão consiste na gravação de uma nova base, utilizando os métodos desenvolvidos para este trabalho, e visando assim obter resultados superiores aos aqui relatados. Isso se deve ao fato de que os dados registrados são aplicados para o treinamento e avaliação dos classificadores implementados, de modo que a sua corretude exerce influência direta sobre eles. Dessa maneira, a interface implementada pode ser utilizada, de para oferecer aos voluntários uma visualização em tempo real da classificação obtida para os gestos realizados. Com isso, evita-se que ocorram erros comuns, que tornam as suas poses inadequadas para uma classe específica, ou que impossibilitam a sua identificação pelo sistema. Entre diversos exemplos desses erros podem ser citados:

- Distribuição de dedos sob variações inadequadas para a pose em questão;
- Rotações de mãos e braços que impossibilitam a visualização dos membros das mãos;
- Posicionamento das mãos fora dos limites espacias de captura do sensor utilizado;
- Mudanças da pose realizada durante a sua gravação, muitas vezes imperceptíveis pelos envolvidos, ocorridas por falta de atenção, cansaço físico, ou baixa flexibilidade dos membros do corpo envolvidos nesses movimentos.

Além do software utilizado, a seleção de usuários com experiência na execução de gestos manuais complexos, ou o seu treinamento prévio na a execução dos gestos desejados, tende a auxiliar a obtenção de um registro de poses mais adequado, e mitigar a vulnerabilidade da esqueletização de usuário a posturas corporais específicas. Por fim, para obter registros mais adequados ao treinamento e validação desse sistema, deve-se prover um controle rígido da iluminação infravermelha ambiente durante as gravações. Desta forma, tanto a qualidade das imagens, quanto a capacidade de rastreamento da *OpenNI* tendem a ser incrementadas, já que estas são extremamente influenciadas por esta frequência eletromagnética específica.

A terceira é ultima sugestão corresponde ao aprimoramento dos algoritmos implementados, afim melhorar a sua performance como um todo. Entre outros aprimoramentos identificados são aqui sugeridos:

- Utilização de técnicas de esqueletizações baseadas em contornos contendo buracos internos. Visando assim a discriminação de um maior número de poses por este método.
- Aplicação de métodos de rastreamento preditivo (218–220), visando maior acurácia e eficiência dos algoritmos de localização de membros das mãos.

- Criação de uma metodologia mista, aproveitando os dados de rastreamento extraídos, como pontos de interesse para técnicas baseadas em modelos geométricos (54, 122).

REFERÊNCIAS

1 DA ROCHA, H. V.; BARANAUSKAS, M. C. C. *Design e avaliação de interfaces humano-computador*. Campinas: UNICAMP, 2003.

2 CARROLL, J. M. *Human computer interaction*: brief intro. Aarhus, Denmark: The Interaction Design Foundation, 2013.

3 ANGKANANON, K.; WALD, M.; GILBERT, L. Designing mobile web solutions for interaction scenarios involving disabled people. 2013. Disponível em: <http://eprints.soton.ac.uk/352503/1/801-017.pdf>. Acesso em: 25 fev. 2014.

4 CHANG, Y.-J.; CHEN, S.-F.; HUANG, J.-D. A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in Developmental Disabilities*, v. 32, n. 6, p. 2566–2570, 2011.

5 GRUDIN, J. The computer reaches out: the historical continuity of interface design. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (SIGCHI), 1990, Seattle, Washington, USA. *Proceedings...* New York, NY, USA: ACM, c1990.

6 RAUTARAY, S.; AGRAWAL, A. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial IntelligenceReview*, 2012. Diponível em: <http://download.springer.com/static/pdf/935/art%253A10.1007%252Fs1046 2-012-9356-9.pdf? auth66=1393194280_e74008ef56216da00618024d9b7c5472& ext=.pdf>. Acesso em: 21 fev. 2014.

7 JØRGENSEN, A. H.; MYERS, B. A. User interface history. 2008. Disponível em: <http://dl.acm.org/citation.cfm?id=1358696>. Acesso em: 25 fev. 2014.

8 MUSTAQUIM, M. Automatic speech recognition- an approach for designing inclusive games. *Multimedia Tools and Applications*, v. 66, n. 1, p. 131–146, 2013. Disponível em: <http://download.springer.com/static/pdf/768/art%253A10.1007%252Fs11042-011-0918-7.pdf? auth66=1393195511_f0b2600a1e49ef71a3b0ca54b4192733&ext=.pdf>. Acesso em: 25 fev. 2014.

9 LEE, K.-Y.; JANG, D. Ethical and social issues behind brain-computer interface. In: INTERNATIONAL WINTER WORKSHOP ON BRAIN-COMPUTER INTERFACE (BCI), 2013, Seoul, Korea. *Proceedings...* Seoul, Korea: IEEE, c2013. p. 72–75.

10 LAAR, B.; GÜRKÖK, H.; BOS, D.-O.; NIJBOER, F.; NIJHOLT, A. Brain-computer interfaces and user experience evaluation. In: ALLISON, B. Z.; DUNNE, S.; LEEB, R.;

MILLÁN, J.; NIJHOLT, A. (Ed.) *Towards practical brain-computer interfaces, biological and medical physics, biomedical engineering*. Berlin Heidelberg: Springer, 2013. p. 223–237.

11 SONG, L.; HU, R.; XIAO, Y.; GONG, L. *Real-time 3d hand gesture recognition from depth image*. 2013. Disponível em: <SE0347.pdf (884 K)>. Acesso em: 25 fev. 2014.

12 CHEN, L.; WEI, H.; FERRYMAN, J. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, v. 34, n. 15, p. 1995–2006, 2013.

13 OPENNI the standart framework for 3D sensing. Disponível em: <http://www.openni.org/>. Acesso em: 25 fev. 2014.

14 KINECT for windows. Disponível em: <http://www.microsoft.com/en-us/kinectforwindowsdev/start.aspx>. Acesso em: 25 fev. 2014.

15 CHAUDHARY, A.; RAHEJA, J. L.; DAS, K.; RAHEJA, S. Intelligent approaches to interact with machines using hand gesture recognition in natural way: a survey. *International Journal of Computer Science & Engineering Survey*, v. 2, n. 1, 2011. Disponível em: <http://arxiv.org/ftp/arxiv/papers/1303/1303.2292.pdf>. Acesso em: 25 fev. 2014.

16 LEE, S.-H.; SOHN, M.-K.; KIM, D.-J.; KIM, B.; KIM, H. Smart tv interaction system using face and hand gesture recognition. In: INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS (ICCE), 2013, Las Vegas, NV, USA. *Proceedings...* Las Vegas, NV, USA: IEEE, c2013. p. 173–174.

17 HASAN, M.; MISRA, P. Gesture recognition using modified hsv segmentation. 2011. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5966463>. Acesso em: 25 fev. 2014.

18 KROEKER, K. L. Alternate interface technologies emerge. *Communications of the ACM*, v. 53, n. 2, p. 13–15, 2010.

19 DI BAJA, G. S.; SERINO, L.; ARCELLI, C. 3D curve skeleton computation and use for discrete shape analysis. In: BREUŸ, M.; BRUCKSTEIN, A.; MARAGOS, P. (Ed.) *Innovations for shape analysis, mathematics and visualization*. Berlin Heidelberg: Springer, 2013. p. 117–136. doi: 10.1007/978-3-642-34141-0_6.

20 LI, X.; HONG, K.-S. Korean chess game implementation by hand gesture recognition using stereo camera. In: INTERNATIONAL CONFERENCE ON COMPUTING TECHNOLOGY AND INFORMATION MANAGEMENT(ICCM), 8., 2012, Seoul, Korea (South). *Proceedings...* Seoul, Korea (South): IEEE, c2012. p. 741–744.

21 YANG, M.; KPALMA, K.; RONSIN, J.; OTHERS. A survey of shape feature extraction techniques. *Pattern recognition*, p. 43–90, 2008.

22 MURTHY, G.; R.S.JADON. A review of vision based hand gestures recognition. *International Journal of Information Technology and Knowledge Management*, v. 2, n. 2, p. 405–410, 2009.

23 LEE, B.; CHUN, J. Manipulation of virtual objects in marker-less ar system by fingertip tracking and hand gesture recognition. In: INTERNATIONAL CONFERENCE ON INTERACTION SCIENCES, 2., 2009, New York, USA. *Proceedings...* New York, USA: ACM, c2009. p. 1110–1115.

24 MAHMOUDI, F.; PARVIZ, M. Visual hand tracking algorithms. 2006. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01648771>. Acesso em: 25 fev. 2014.

25 NGUYEN, D. D.; PHAM, T. C.; JEON, J. W. Fingertip detection with morphology and geometric calculation. In: INTELLIGENT ROBOTS AND SYSTEMS, 2009, St. Louis, USA. *Proceedings ...* St. Louis, USA: IEEE/RSJ, c2009. p. 1460–1465.

26 KHAN, R. Z.; IBRAHEEM, N. A. Survey on gesture recognition for hand image postures. *Computer and Information Science*, v. 5, n. 3, p. 110–121, 2012.

27 MITRA, S.; ACHARYA, T. Gesture recognition: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*: applications and reviews, v. 37, n. 3, p. 311–324, 2007.

28 BRAHEM, M. B.; MENELAS, B.-A. J.; OTIS, M. J.-D. Use of a 3dof accelerometer for foot tracking and gesture recognition in mobile hci. *Procedia Computer Science*, v. 19, p. 453 – 460, 2013. doi: 10.1016/j.procs.2013.06.061.

29 AUEPHANWIRIYAKUL, S.; PHITAKWINAI, S.; SUTTAPAK, W.; CHANDA, P.; THEERA-UMPON, N. Thai sign language translation using scale invariant feature transform and hidden markov models. *Pattern Recognition Letters*, v. 34, n. 11, p. 1291 – 1298, 2013.

30 YIN, P.; STARNER, T.; HAMILTON, H.; ESSA, I.; REHG, J. Learning the basic units in american sign language using discriminative segmental feature selection. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2009, Taipei, Taiwan. *Proceedings...* Taipei, Taiwan: IEEE, c2009. p. 4757–4760.

31 CASSANO, A. Illustrated guide to italian hand gestures. Disponível em: <http://www.amusingplanet.com/2010/11/illustrated-guide-to-italian-hand.html>. Acesso em: 25 fev. 2014.

32 CUSTÓDIO, L. N.; REZENDE, L.; DE FARIA, E. R. Por dentro dos jogos de salão. Disponível em: <http://portaldoprofessor.mec.gov.br/fichaTecnicaAula.html?aula=52069>. Acesso em: 25 fev. 2014.

33 CAPOVILLA, F. C.; RAPHAEL, W. D. *Dicionário enciclopédico ilustrado trilíngüe da língua de sinais brasileira*: sinais de M a Z. 3. ed. São Paulo: EDUSP, 2001. 1620 p. ISBN: 9788531406690.

34 LINGUA gestual portuguesa.

35 ASL - american sign language. Disponível em: <http://lifeprint.com/>. Acesso em: 25 fev. 2014.

36 ITEC - let's enjoy learning japanese sign language. Disponível em: <http://www.kyoto-be.ne.jp/ed-center/gakko/jsl/>. Acesso em: 25 fev. 2014.

37 GESTUNO: instituto emanuel. Disponível em: <http://institutoemanuel.webnode.com.br/lingua-de-sinais/gestuno/>. Acesso em: 25 fev. 2014.

38 RAID, J. Onze sinais de LIBRAS e um sinal de ASL. Disponível em: <http://angel-raid.deviantart.com/art/Libras-Signs-213488523>. Acesso em: 25 fev. 2014.

39 LIBRAS versão 2.1 - web - 2008, dicionário da língua brasileira de sinais. Disponível em: <http://www.acessobrasil.org.br/libras/>. Acesso em: 25 fev. 2014.

40 HALL, E. T. The silent language. *Behavioral Science*, v. 7, n. 4, p. 477–478, 1962.

41 JOWERS, I.; PRATS, M.; MCKAY, A.; GARNER, S. Evaluating an eye tracking interface for a two-dimensional sketch editor. *Computer-Aided Design*, v. 45, n. 5, p. 923 – 936, 2013.

42 HIKITA, S.; SETO, Y. Point-and-Click Interface Based on Parameter-Free Eye Tracking Technique Using a Single Camera. In: STEPHANIDIS, C. (Ed.) *HCI International 2013 - poster's extended abstracts*. Heidelberg: Springer Berlin, 2013. p. 608-612 (Communications in computer and information science, v. 373).

43 MCGUINNESS, R. WHAT the future holds for eye tracking technology. Disponível em: <http://metro.co.uk/2011/12/05/what-the-future-holds-for-eye-tracking-technology-245368>. Acesso em: 25 fev. 2014.

44 HUSSAIN, M. S.; CALVO, R. A.; CHEN, F. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers*, 2013. Disponível em: <http://iwc.oxfordjournals.org/content/early/2013/06/06/iwc.iwt032.short>. Acesso em: 25 fev. 2014.

45 WEISE, T.; BOUAZIZ, S.; LI, H.; PAULY, M. Realtime performance-based facial animation. *ACM Transaction on Graphics*, v. 30, n. 4, 2011. doi: 10.1145/2010324.1964972.

46 HOLT, B.; ONG, E.-J.; COOPER, H.; BOWDEN, R. Putting the pieces together: connected poselets for human pose estimation. *IEEE International Conference on Computer Vision Workshops*, p. 1196–1201, 2011. doi: 10.1109/ICCVW.2011.6130386.

47 SADAGIC, A.; KOLSCH, M.; WELCH, G.; BASU, C.; DARKEN, C.; WACHS, J. P.; FUCHS, H.; TOWLES, H.; ROWE, N.; FRAHM, J.-M.; ET AL. Smart instrumented training ranges: bringing automated system solutions to support critical domain needs. *The Journal of Defense Modeling and Simulation*: applications, methodology, technology, v. 10, n. 3, p. 327–342, 2013.

48 PARK, J.-W.; OH, C.-M.; LEE, C.-W. Virtual Flying Experience Contents Using Upper-Body Gesture Recognition. In: STEPHANIDIS, C. (Ed.) *HCI International 2013*: posters' extended abstracts. Heidelberg: Springer Berlin, 2013 (Communications in Computer and Information Science, v. 373), p. 367–371.

49 WII official site at nintendo. Disponível em: <http://www.nintendo.com/wii/>. Acesso em: 25 fev. 2014.

50 XBOX.COM brasil - xbox.com. Disponível em: <http://www.xbox.com/pt-BR/Home-2/>. Acesso em: 25 fev. 2014.

51 PLAYSTATION: PS4, PS3, PS Vita, PSP, PSPgo, PS2, playstation jogos - playstation-network. Disponível em: <http://br.playstation.com/>. Acesso em: 25 fev. 2014.

52 LI, Y.-T.; WACHS, J. P. Hegm: a hierarchical elastic graph matching for hand gesture recognition. *Pattern Recognition*, v. 47, n. 1, p. 80–88, 2014.

53 ZHANG, Q. Hand gesture tracking based on particle filtering aiming at real-time performance. *Journal of Information and Computational Science*, v. 10, n. 4, p. 1149–1157, 2013.

54 KESKIN, C.; KIRAC, F.; KARA, Y.; AKARUN, L. Real Time Hand Pose Estimation Using Depth Sensors. In: FOSSATI, A.; GALL, J.; GRABNER, H.; REN, X.; KONOLIGE, K. (Ed.) *Consumer depth cameras for computer visionn: research topics and applications*. London: Springer-Verlag London, 2013. p. 119–137. doi: 10.1007/978-1-4471-4640-7_7. (Advances in computer vision and pattern recognition).

55 BILAL, S.; AKMELIAWATI, R.; EL SALAMI, M.; SHAFIE, A. Vision-based hand posture detection and recognition for sign language: a study. In: INTERNATIONAL CONFERENCE ON MECHATRONICS, 4., 2011, Kuala Lumpur, Malaysia. *Proceedings ...* Kuala Lumpur, Malaysia: IEEE, c2011. p. 1–6.

56 GESTURE recognition - wikipedia, the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Gesture_recognition>. Acesso em: 25 fev. 2014.

57 HAND gesture recognition via model fitting in energy minimization w/opencv. Disponível em: <http://www.morethantechnical.com/2010/12/28/hand-gesture-recognition-via-model-fitting-in-energy-minimization-wopencv/>. Acesso em: 25 fev. 2014.

58 HAI TRAN, T. T. How can human communicate with robot by hand gesture? In: INTERNATIONAL CONFERENCE ON COMPUTING, MANAGEMENT AND TELECOMMUNICATIONS (COMMANTEL), 2013, Ho Chi Minh City, Vietnam. *Proceedings ...* Ho Chi Minh City, Vietnam: IEEE, c2013. p. 235–240.

59 EROL, A.; BEBIS, G.; NICOLESCU, M.; BOYLE, R. D.; TWOMBLY, X. Vision-based hand pose estimation: a review. *Computer Vision and Image Understanding*, v. 108, n. 1-2, p. 52 – 73, 2007.

60 CHAUDHARY, A.; RAHEJA, J. L. Abhivyakti: a vision based intelligent system for elder and sick persons. 2011. Disponível em: <http://arxiv.org/ftp/arxiv/papers/1109/1109.6442.pdf>. Acesso em: 25 fev. 2014.

61 SUN, L.; LIU, G. Hand tracking based on the combination of 2d and 3d model in gaze-directed video. In: INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, 2011, Barcelona, Spain. *Proceedings ...* Barcelona, Spain: IEEE, c2011. p. 1–6.

62 PAVLOVIC, V.; SHARMA, R.; HUANG, T. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 7, p. 677–695, 1997.

63 BROX, T.; ROSENHAHN, B.; CREMERS, D.; SEIDEL, H.-P. High accuracy optical flow serves 3-D pose tracking: exploiting contour and flow based constraints. In: LEONARDIS, A.; BISCHOF, H.; PINZ, A. (Ed.) *Computer Vision ECCV*. Heidelberg: Springer Berlin, 2006 (Lecture notes in computer science, v. 3952), p. 98–111.

64 QIN, S.; ZHU, X.; YANG, Y.; JIANG, Y. Real-time hand gesture recognition from depth images using convex shape decomposition method. *Journal of Signal Processing Systems*, v. 74, n. 1, p. 47–58, 2014.

65 KOLSCH, M.; TURK, M. Fast 2d hand tracking with flocks of features and multi-cue integration. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOP (CVPRW), 2004, Washington, USA. *Proceedings ...* Washington, USA: IEEE, c2004. p. 158–158.

66 DARDAS, N.; GEORGANAS, N. D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and Measurement*, v. 60, n. 11, p. 3592–3607, 2011.

67 SAMUEL, A. L. Programming computers to play games. New York: Academic Press, 1960. p. 165-192 (Advances in computers, v. 1).

68 MITCHELL, T. M. *Machine learning*. New York: McGraw-Hill Higher Education, 1997. (McGraw-Hill series in computer science, v. 45).

69 BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. New York: Springer, 2006. (Information Science and Statistics). Disponível em: <http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738>. Acesso em: 25 fev. 2014.

70 KIM, S.; SON, J.; LEE, G.; KIM, H.; LEE, W. Tapboard: making a touch screen keyboard more touchable. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2009, Paris. *Proceedings ...* Paris: ACM, c2013. doi:10.1145/2470654.2470733.

71 ZHONG, H.; WACHS, J. P.; NOF, S. Y. A collaborative telerobotics network framework with hand gesture interface and conflict prevention. *International Journal of Production Research*, v. 51, n. 15, p. 4443–4463, 2013. doi:10.1080/00207543.2012.756591.

72 NAM, Y.; RHO, S.; LEE, C. Physical activity recognition using multiple sensors embedded in a wearable device. *Journal ACM Transactions on Embedded Computing Systems*, v. 12, n. 2, p. 26:1–26:14, 2013. doi:10.1145/2423636.2423644.

73 FARRINGDON, J.; MOORE, A.; TILBURY, N.; CHURCH, J.; BIEMOND, P. Wearable sensor badge and sensor jacket for context awareness. In: INTERNATIONAL SYMPOSIUM ON WEARABLE COMPUTERS, 3., 1999, San Francisco. *Proceedings...* San Francisco: IEEE, c1999. p. 107–113.

74 KANG, J.; ZHONG, K.; QIN, S.; WANG, H.; WRIGHT, D. Instant 3d design concept generation and visualization by real-time hand gesture recognition. *Computers in Industry*, v. 64, n. 7, p. 785–797, 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016636151300095X>. Acesso em: 25 fev. 2014.

75 GUO, Y.; WANG, Q.; HUANG, S.; ABRAHAM, A. Hand gesture recognition system using single-mixture source separation and flexible neural trees. *Journal of Vibration and Control*, 2013. doi: 10.1177/1077546313481001.

76 JACOB, M. G.; WACHS, J. P.; PACKER, R. A. Hand-gesture-based sterile interface for the operating room using contextual cues for the navigation of radiological images. *Journal of the American Medical Informatics Association*, v. 20, n. e1, p. e183–e186, 2013. doi: 10.1136/amiajnl-2012-001212.

77 DARDAS, N.; PETRIU, E. Hand gesture detection and recognition using principal component analysis. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE FOR MEASUREMENT SYSTEMS AND APPLICATIONS (CIMSA), 2011, Ottawa, Canada. *Proceedings...* Ottawa, Canada: IEEE, c2011. p. 1–6.

78 GUAYARA, E. Realidad virtual. Disponível em: <http://larealidad-virtual.blogspot.com.br/>. Acesso em: 25 fev. 2014.

79 WOLF, M.; ASSAD, C.; STOICA, A.; YOU, K.; JETHANI, H.; VERNACCHIA, M.; FROMM, J.; IWASHITA, Y. Decoding static and dynamic arm and hand gestures from the jpl biosleeve. In: THE INTERNATIONAL CONFERENCE FOR AEROSPACE EXPERTS, ACADEMICS, MILITARY PERSONNEL, AND INDUSTRY LEADERS, 2013, Big Sky Resort, MT, USA. *Proceedings...* Big Sky Resort, MT, USA: IEEE, c2013. p. 1–9.

80 MULATTO, S.; FORMAGLIO, A.; MALVEZZI, M.; PRATTICHIZZO, D. Using postural synergies to animate a low-dimensional hand avatar in haptic simulation. *IEEE Transactions on Haptics*, v. 6, n. 1, p. 106–116, 2013. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6171185&isnumber =6479194>. Acesso em: 25 fev. 2014.

81 CHUN, W. H.; HÖLLERER, T. Real-time hand interaction for augmented reality on mobile phones. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES, 2013, Santa Monica, California, USA. *Proceedings...* New York, NY, USA: ACM, c2013. p. 307–314.

82 MINORITY report de steven spielberg. Disponível em: <http://aripictures.com/minority-report>. Acesso em: 25 fev. 2014.

83 WU, Y.; ZHAO, L.; DING, H. Robust hand gesture recognition with feature selection and hierarchical temporal self-similarities. *International Journal of Information and Electronics Engineering*, v. 3, n. 5, p. 510–515, 2013. doi: 10.7763/IJIEE.2013.V3.368.

84 RAHEJA, J. L.; CHAUDHARY, A.; SINGAL, K. Tracking of fingertips and centers of palm using kinect. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE, MODELLING AND SIMULATION (CIMSIM), 3., 2011, Langkawi, Malaysia. *Proceedings...* Langkawi, Malaysia: IEEE, c2011. p. 248–252.

85 XU, L.; FANG, Y.; WANG, K.; LI, J. Plug&touch: a mobile interaction solution for large display via vision-based hand gesture detection. In: INTERNATIONAL CONFERENCE ON MULTIMEDIA, 20., 2012, Nara, Japan. *Proceedings...* New York: ACM, c2012. p. 1177–1180.

86 HANNUKSELA, J.; BARNARD, M.; SANGI, P.; HEIKKILA, J. Adaptive Motion-Based Gesture Recognition Interface for Mobile Phones. Berlin Heidelberg: Springer, 2008 (Lecture notes in computer science, v. 5008), p. 271–280. doi: 10.1007/978-3-540-79547-6_26.

87 LINGRAND, D.; RENEVIER, P.; PINNA-DERY, A.-M.; CREMASCHI, X.; LION, S.; ROUEL, J.-G.; JEANNE, D.; CUISINAUD, P.; SOULA, J. Gestaction3D: a platform for studying displacements and deformations of 3D objects using hands. In: CALVARY, G.; PRIBEANU, C.; SANTUCCI, G.; VANDERDONCKT, J. (Ed.). INTERNATIONAL CONFERENCE

ON COMPUTER-AIDED DESIGN OF USER INTERFACES(CADUI), 6., 2006, Bucharest, Romania. *Proceedings...* Bucharest, Romania: Springer, c2006. p. 101–110.

88 JOHNSON, R.; O?HARA, K.; SELLEN, A.; COUSINS, C.; CRIMINISI, A. Exploring the potential for touchless interaction in image-guided interventional radiology. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2011, Vancouver. *Proceedings...* New York: ACM, c2011. p. 3323–3332.

89 GALIANA, I.; BARRIO, J.; BRENOSA, J.; FERRE, M. Modular multi-finger haptic device: mechanical design, controller and applications. In: GALIANA, I.; FERRE, M. (Ed.) *Multi-finger haptic interaction touch and haptic systems*. Lodon: Springer, 2013. p. 55–83. doi: 10.1007/978-1-4471-5204-0_4.

90 DONOVAN, J.; SADE, G.; SEEVINCK, J. Gestural, emergent and expressive: three research themes for haptic interaction. In: MARCUS, A. (Ed.) *Design, user experience, and usability*: user experience in novel technological environments. Berlin Heidelberg: Springer, 2013 (Lecture notes in computer science, v. 8014). doi: 10.1007/978-3-642-39238-2_39.

91 WANG, R. Y.; POPOVIć, J. Real-time hand-tracking with a color glove. *Transactions on Graphics (TOG)*, New York, v. 28, n. 3, p. 63:1–63:8, 2009. doi: 10.1145/1576246.1531369.

92 GARDNER, A.; DUNCAN, C. A.; SELMIC, R.; KANNO, J. Real-time classification of dynamic hand gestures from marker-based position data. In: INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES COMPANION, 2013, Santa Monica, California, USA. *Proceedings...* New York: ACM, c2013. p. 13–16.

93 KIM, H.-J.; KIM, H.; CHAE, S.; SEO, J.; HAN, T.-D. Ar pen and hand gestures: a new tool for pen drawings. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2013, Paris. *Proceedings...* New York: ACM, c2013. p. 943–948.

94 PIUMSOMBOON, T.; CLARK, A.; BILLINGHURST, M.; COCKBURN, A. User-defined gestures for augmented reality. In: *Human-computer interaction*. Berlin Heidelberg: Springer, 2013 (Lecture notes in computer science, v. 8118), p. 282–299. doi: 10.1007/978-3-642-40480-1_18.

95 MAYA: software de animação 3D. Disponível em: <http://www.autodesk.com.br/products/autodesk-maya/overview>. Acesso em: 25 fev. 2014.

96 3DS MAX | software de modelagem e renderização 3D. Disponível em: <http://www.autodesk.com.br/products/autodesk-3ds-max/overview>. Acesso em: 25 fev. 2014.

97 BLENDER.org - home of the blender project - Free and Open 3D creation software. Disponível em: <http://www.blender.org/>. Acesso em: 25 fev. 2014.

98 BLAIS, F. Review of 20 years of range sensor development. In: SOCIETY OF PHOTO-OPTICAL INSTRUMENTATION ENGINEERS (SPIE), 2003, Santa Clara, CA. *Proceedings...* Santa Clara, CA: SPIE, c2003. p. 62–76, doi: 10.1117/12.473116.

99 BAKER, M. L.; DALRYMPLE, G. V. Biological effects of diagnostic ultrasound: a review. *Radiology*, v. 126, n. 2, p. 479–83, 1978. Disponível em: <http://europepmc.org/abstract/MED/622502/reload=0;jsessionid=e3wXR4qN7J XnEQXQemLh.12>. Acesso em: 25 fev. 2014.

100 ROSALES, R.; ATHITSOS, V.; SIGAL, L.; SCLAROFF, S. 3D hand pose reconstruction using specialized mappings. In: INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2001, Vancouver, Canada. *Proceedings...* Vancouver, Canada: IEEE, c2001. p. 378–385.

101 BARKANA, Y.; BELKIN, M. Laser eye injuries. *Survey of Ophthalmology*, v. 44, n. 6, p. 459–478, 2000.

102 HACIHALILOGLU, I.; GUY, P.; HODGSON, A. J.; ABUGHARBIEH, R. Volume-specific parameter optimization of 3d local phase features for improved extraction of bone surfaces in ultrasound. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2014. doi: 10.1002/rcs.1552.

103 BARNETT, S. B.; TER HAAR, G. R.; ZISKIN, M. C.; ROTT, H.-D.; DUCK, F. A.; MAEDA, K. International recommendations and guidelines for the safe use of diagnostic ultrasound in medicine. *Ultrasound in Medicine & Biology*, v. 26, n. 3, p. 355–366, 2000.

104 WERGHI, N. Segmentation and modeling of full human body shape from 3-D scan data: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*: applications and reviews, v. 37, n. 6, p. 1122–1136, 2007.

105 DEICHMANN, R.; HAHN, D.; HAASE, A. Fast t1 mapping on a whole-body scanner. *Magnetic Resonance in Medicine*, v. 42, n. 1, p. 206–209, 1999.

106 GU, J.; CHANG, T.; MAK, I.; GOPALSAMY, S.; SHEN, H.; YUEN, M. A 3D reconstruction system for human body modeling. In: MAGNENAT-THALMANN, N.; THALMANN, D. (Ed.) *Modelling and motion capture techniques for virtual environments*. Berlin Heidelberg: Springer, 1998 (Lecture notes in computer science, v. 1537), p. 229–241. doi: 10.1007/3-540-49384-0_18.

107 FOSSUM, E. R. Digital camera system on a chip. *Journal of Microelectromechanical Systems*, Los Alamitos, USA, v. 18, n. 3, p. 8–15, 1998. doi: 10.1109/40.683047.

108 EXPEED - wikipedia, the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Expeed>. Acesso em: 25 fev. 2014.

109 HAN, Y.; WANG, B.; IDESAWA, M.; SHIMAI, H. Recognition of multiple configurations of objects with limited data. *Pattern Recognition*, v. 43, n. 4, p. 1467 – 1475, 2010.

110 BACH, M.; POLOSCHEK, C. Optical illusions. *Advances in Clinical Neuroscience and Rehabilitation*, v. 6, n. 2, p. 20–21, 2006.

111 LOWE, D. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, v. 60, n. 2, p. 91–110, 2004.

112 PURGHE, F. Privileged directions for subjective contours: horizontal and vertical versus tilted. *Perception*, v. 18, n. 2, p. 201–213, 1989.

113 AMES room - Wikipedia, the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Ames_room>. Acesso em: 25 fev. 2014.

114 ZIMMERMAN, C. How to remove a finger - Curtis Zimmerman - youtube. Disponível em: <www.youtube.com/watch?v=CIv0-Uaim8>. Acesso em: 25 fev. 2014.

115 SPINNING dancer - wikipedia, the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Spinning_Dancer>. Acesso em: 25 fev. 2014.

116 BARIBEAU, R.; RIOUX, M.; GODIN, G. Color reflectance modeling using a polychromatic laser range sensor. *Transactions on Pattern Analysis and Machine Intelligence*, v. 14, n. 2, p. 263–269, 1992.

117 MCDONALD, K. Kinect point cloud with depth of field, because we always need more dof. Disponível em: <http://www.flickr.com/photos/28622838@N00/5174106004>. Acesso em: 25 fev. 2014.

118 MOLYNEAUX, D. Kinectfusion rapid 3d reconstruction and interaction with microsoft kinect. In: INTERNATIONAL CONFERENCE ON THE FOUNDATIONS OF DIGITAL GAMES, 2012, Raleigh, North Carolina. *Proceedings...* New York: ACM, c2012. p. 3–3.

119 BESL, P. Active, optical range imaging sensors. *Machine Vision and Applications*, v. 1, n. 2, p. 127–152, 1988.

120 STOWERS, J.; HAYES, M.; BAINBRIDGE-SMITH, A. Altitude control of a quadrotor helicopter using depth map from microsoft kinect sensor. In: INTERNATIONAL CONFERENCE ON MECHATRONICS (ICM), 2011, Istanbul, Turkey. *Proceedings...* Istanbul, Turkey: IEEE, c2011. p. 358–362.

121 LIU, X.; FUJIMURA, K. Hand gesture recognition using depth data. In: INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION, 6., 2004, Seoul, Korea. *Proceedings...* Seoul, Korea: IEEE, c2004. p. 529–534.

122 OIKONOMIDIS, I.; KYRIAZIS, N.; ARGYROS, A. Efficient model-based 3d tracking of hand articulations using kinect. In: BRITISH MACHINE VISION CONFERENCE, 22., 2011, Dundee, Scotland. *Proceedings ...* Dundee, Scotland: BMVA Press, c2011.

123 CYGANEK, B.; SIEBERT, J. *An introduction to 3D computer vision techniques and algorithms*. New York: Wiley, 2011. Disponível em: <http://books.google.com.br/books?id=4leU0ZBkHngC>. Acesso em: 25 fev. 2014.

124 DEVERNAY, F.; FAUGERAS, O. Computing differential properties of 3-d shapes from stereoscopic images without 3-d models. In: COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 1994, Seattle. *Proceedings...* Seattle: IEEE, c1994. p. 208–213.

125 EPIPOLAR GEOMETRY - wikipedia, the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Epipolar_geometry>. Acesso em: 25 fev. 2014.

126 UM, G.-M.; KIM, K. Y.; AHN, C.; LEE, K. H. Three-dimensional scene reconstruction using multiview images and depth camera. In: WOODS, A.J., et al (Ed.). *Stereoscopic displays and virtual reality systems XII*. 2005. p. 271-280. (Proceedings.of the SPIE, v. 5664). doi: 10.1117/12.586764.

127 NICKEL, K.; STIEFELHAGEN, R. Pointing gesture recognition based on 3d-tracking of face, hands and head orientation. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERFACES, 5., 2003, Vancouver, British Columbia, Canada. *Proceedings...* New York: ACM, c2003. p. 140–146.

128 MAYER, R. *Scientific canadian*: invention and innovation from canada's national research council. University of Michigan: Raincoast Books, 1999. 192p.

129 BARIBEAU, R.; RIOUX, M. Influence of speckle on laser range finders. *Applied Optics*, v. 30, n. 20, p. 2873–2878, 1991.

130 SMITH, W. *Modern optical engineering*. McGraw-Hill Professional, 2008. McGraw-Hill series on optical and electro-optical engineering. Disponível em: <http://www.amazon.com/Modern-Optical-Engineering-4th-Ed/dp/0071476873>. Acesso em: 25 Fev. 2014.

131 LANGE, R.; SEITZ, P. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics*, v. 37, n. 3, p. 390–397, 2001.

132 DROESCHEL, D.; STUCKLER, J.; BEHNKE, S. Learning to interpret pointing gestures with a time-of-flight camera. In: INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION (HRI), 6., 2011, Lausanne, Switzerland. *Proceedings ...* Lausanne, Switzerland: ACM/IEEE, c2011. p. 481–488.

133 KOLB, A.; BARTH, E.; KOCH, R.; LARSEN, R. Time-of-flight cameras in computer graphics. *Computer Graphics Forum*, v. 29, n. 1, p. 141–159, 2010.

134 WANG, J. Y. Imaging laser radar - an overview. In: LASES '86; INTERNATIONAL CONFERENCE ON LASERS AND APPLICATIONS, 9., 1987, Orlando, FL, USA. *Proceedings...* McLean, VA, USA: STS Press, c1987. p. 19–29.

135 FOIX, S.; ALENYA, G.; TORRAS, C. Lock-in time-of-flight (tof) cameras: a survey. *IEEE Sensors Journal*, v. 11, n. 9, p. 1917–1926, 2011. doi: 10.1109/JSEN.2010.2101060.

136 3D image sensor - panasonic. Disponível em: <http://pewa.panasonic.com/components/built-in-sensors/3d-image-sensors/>. Acesso em: 25 fev. 2014.

137 FOTONIC-TIME of flight-TOF-range camera-3D camera. Disponível em: <http://www.fotonic.com/content/Default.aspx>. Acesso em: 25 fev. 2014.

138 LEADING 3D chip technology provider | pmdtec.com. Disponível em: <http://www.pmdtec.com/>. Acesso em: 25 fev. 2014.

139 MESA imaging AG - SwissRanger SR4000 - miniature 3D time-of-flight range camera, 3D camera. Disponível em: <http://www.mesa-imaging.ch/>. Acesso em: 25 fev. 2014.

140 RUSSELL, J.; COHN, R. *Softkinetic*. [S. l]: Book on Demand, 2012. 84 p.

141 INTEL perceptual computing SDK 2013. Disponível em: <http://software.intel.com/en-us/vcsource/tools/perceptual-computing-sdk>. Acesso em: 25 fev. 2014.

142 LUM, Z. M. A.; LIANG, X.; PAN, Y.; ZHENG, R.; XU, X. Increasing pixel count of holograms for three-dimensional holographic display by optical scan-tiling. *Optical Engineering*, v. 52, n. 1, p. 015802–015802, 2013. doi: 10.1117/1.OE.52.1.015802.

143 BOGUE, R. Three-dimensional measurements: a review of technologies and applications. *Sensor Review*, v. 30, n. 2, p. 102–106, 2010. doi: 10.1108/02602281011022670.

144 SPAGNOLO, G. S. Potentiality of 3d laser profilometry to determine the sequence of homogenous crossing lines on questioned documents. *Forensic Science International*, v. 164, n. 2-3, p. 102 – 109, 2006. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2005.12.004>. Acesso em: 25 fev. 2014.

145 POON, T.-C. Recent progress in optical scanning holography. *Journal of Holography and Speckle*, v. 1, n. 1, p. 6–25, 2004. doi:10.1166/jhs.2004.003.

146 SIRAT, G.; PSALTIS, D. *Conoscopic holography*. 1985. p. 324-330 (Proceedings of SPIE 0523). doi: 10.1117/12.946301.

147 LIND, J. Xbox kinect dots. Disponível em: <http://www.youtube.com/watch?v=p2nqxyywbQI#aid=P-MMwGwAC3o>. Acesso em: 25 fev. 2014.

148 SJÖDAHL, M.; SYNNERGREN, P. Measurement of shape by using projected random patterns and temporal digital speckle photography. *Applied Optics*, v. 38, n. 10, p. 1990–1997, 1999. doi: 10.1364/AO.38.001990.

149 VUYLSTEKE, P.; OOSTERLINCK, A. 3-d perception with a single binary coded illumination pattern. 1987. doi: doi:10.1117/12.937839, Disponível em: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1246372>. Acesso em 25 fev. 2014.

150 SANSONI, G.; CORINI, S.; LAZZARI, S.; RODELLA, R.; DOCCHIO, F. Three-dimensional imaging based on gray-code light projection: characterization of the measuring algorithm and development of a measuring system for industrial applications. *Applied Optics*, v. 36, n. 19, p. 4463–4472, 1997. doi: 10.1364/AO.36.004463.

151 PRIME SENSE Ltd. Zeev Zalevsky; Alexander Shpunt; Aviad Malzels; Javier Garcia. *Method and System for Object Reconstruction*. US Patent 8,400,494, 14 Mar. 2006, 19 Mar. 2013. Disponível em: <http://www.google.com/patents/US8400494>. Acesso em: 25 fev. 2014.

152 LITOMISKY, K. *Consumer RGB-D cameras and their applications*. Technical report, University of California, Riverside, 2012. Disponível em: <http://alumni.cs.ucr.edu/~klitomis/files/RGBD-intro.pdf>. Acesso em: 25 fev. 2014.

153 GILES, J. Inside the race to hack the kinect. *New Scientist*, v. 208, n. 2789, p. 22 – 23, 2010. doi: 10.1016/S0262-4079(10)62989-2.

154 CARTWRIGHT, J. Kinectc for xbox 360, meet kinect for windows. Disponível em: <http://techau.com.au/kinect-for-xbox-360-meet-kinect-for-windows/>. Acesso em: 25 fev. 2014.

155 ASUS multimedia. Disponível em: <http://www.asus.com/Multimedia/>. Acesso em: 25 fev. 2014.

156 PRIMESENSE natural interation. Disponível em: <http://www.primesense.com/wp-content/uploads/2012/12/PrimeSenses_3DsensorsWeb.pdf>. Acesso em: 25 fev. 2014.

157 3D sensing technology solutions - primesense. Disponível em: <www.primesense.com>. Acesso em: 25 fev. 2014.

158 RARE - creators of kinect sports rivals, banjo-kazooie, viva piñata, perfect dark and more. Disponível em: <www.rare.net>. Acesso em: 25 fev. 2014.

159 MICROSOF games. Disponível em: <http://www.microsoft.com/games/>. Acesso em: 25 fev. 2014.

160 ROGERS, R. Kinect with linux. *Linux Jornal*, v. 2011, n. 207, 2011.

161 DUTTON, F. Kinect "hackers"won't be prosecuted. Disponível em: <http://www.eurogamer.net/articles/2010-11-22-kinect-hackers-wont-be-prosecuted>. Acesso em: 25 fev. 2014.

162 FASTEST-SELLING gaming peripheral. Disponível em: <http://www.guinnessworldrecords.com/records-9000/fastest-selling-gaming-peripheral/>. Acesso em: 25 fev. 2014.

163 WIDENHOFER, B. Inside xbox 360's kinect controller, 2010. Disponível em: <http://www.eetimes.com/document.asp?doc_id=1281322>. Acesso em: 25 fev. 2014.

164 KINECT for windows sensor components and specifications. Disponível em: <http://msdn.microsoft.com/en-us/library/jj131033.aspx>. Acesso em: 25 fev. 2014.

165 MICROSOFT kinect teardown. Disponível em: <http://www.ifixit.com/Teardown/Microsoft+Kinect+Teardown/4066>. Acesso em: 25 fev. 2014.

166 PRIMESENSE Ltd. Alexander Shpunt, Zeev Zalevsky. *Three-dimensional sensing using speckle patterns*. US Patent 8,390,821, 8 mar. 2007, 5 mar. 2013. Disponível em: <http://www.google.com/patents/US8390821>. Acesso em: 25 fev. 2014.

167 REICHINGER, A. Kinect pattern uncovered. Disponível em: <http://azttm.wordpress.com/2011/04/03/kinect-pattern-uncovered/>. Acesso em: 25 fev. 2014.

168 PRIMESENSE Ltd. Barak Freedman; Alexander Shpunt; Meir Machline; Yoel Arieli. *Depth mapping using projected patterns*. US Patent 8,150,142, 6 Set. 2007, 3 Abr. 2012. Disponível em: <http://www.google.com/patents/US8150142>. Acesso em: 25 fev. 2014.

169 GALL, J.; GRABNER, H. *Consumer depth cameras for computer vision*: research topics and applications. Heidberg: Springer, 2013. 210 p. (Advances in computer vision and pattern recognition).

170 CURSOS de libras online grátis com certificado. Disponível em: <http://gfcursosgratis.com/cursos-de-libras-online-gratis/>. Acesso em: 25 fev. 2014.

171 JAIMES, A.; SEBE, N. Multimodal human computer interaction: a survey. In: COMPUTER VISION IN HUMAN-COMPUTER INTERACTION, 2005, Beijing, China. *Proceedings...* Berlin Heidelberg: Springer, c2005. p. 15–21, (Lecture notes in computer science, v. 3766).

172 BIRK, H.; MOESLUND, T. B.; MADSEN, C. B. Real-time recognition of hand alphabet gestures using principal component analysis. In: . c1997 (. p. 261–268.

173 HASANUZZAMAN, M.; UENO, H. Face and gesture recognition for human-robot interaction. In: DELAC, K.; GRGIC, M. (Ed.) *Face Recognition*. Vienna: I-Tech Education and Publishing, 2007. p. 149–182. doi: 10.5772/4836.

174 OPENCV (open source computer vision library). Disponível em: <http://www.opencv.org/>. Acesso em: 25 fev. 2014.

175 SUZUKI, S.; BE, K. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, v. 30, n. 1, p. 32 – 46, 1985. doi: 10.1016/0734-189X(85)90016-7.

176 GUYON, I.; GUNN, S.; NIKRAVESH, M.; ZADEH, L. *Feature extraction*: foundations and applications. Palo Alto, California: Springer, 2006. 778 p. (Studies in fuzziness and soft computing, v. 207).

177 MAIDI, M.; PREDA, M. Interactive media control using natural interaction-based kinect. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2013, Vancouver, Canada. *Proceedings...* Vancouver, Canada: IEEE, c2013. p. 1812–1815, doi: 10.1109/ICASSP.2013.6637965.

178 FRATI, V.; PRATTICHIZZO, D. Using kinect for hand tracking and rendering in wearable haptics. In: WORLD HAPTICS CONFERENCE (WHC), 2011, Istanbul, Turkish. *Proceedings...* Istanbul, Turkish: IEEE, c2011. p. 317–321, doi: 10.1109/WHC.2011.5945505.

179 WANG, T. Extraction of optimal skeleton of polygon based on hierarchical analysis. In: INTERNATIONAL CONFERENCE ON GEO-SPATIAL SOLUTIONS FOR EMERGENCY MANAGEMENT AND THE 50TH ANNIVERSARY OF THE CHINESE ACADEMY OF SURVEYING AND MAPPING, 2009, Beijing, China. *Proceedings...* Beijing, China: ISPRS Archives, c2009. p. 272–276.

180 MAUS, A. Delaunay triangulation and the convex hull ofn points in expected linear time. *BIT Numerical Mathematics*, v. 24, n. 2, p. 151–163, 1984. doi: 10.1007/BF01937482.

181 RUFAT, D. Convex hulls and mesh boundaries. Disponível em: <http://dzhelil.info/triangle/convex.html>. Acesso em: 25 fev. 2014.

182 HERT, S.; SCHIRRA, S. The convex hull of a model made of 192135 points. Disponível em: <http://doc.cgal.org/latest/Convex_hull_3/index.html>. Acesso em: 25 fev. 2014.

183 Fitting a polygon about a set of points. Disponível em: <http://www.idlcoyote.com/math_tips/convexhull.html>. Acesso em: 25 fev. 2014.

184 DE BERG, M.; VAN KREVELD, M.; OVERMARS, M.; SCHWARZKOPF, O. C. *Computational geometry*: algorithms and applications. 3rd. ed. Santa Clara, CA: Springer-Verlag, 2008. 386p.

185 KIM, Y.-J.; KIM, M.-S.; ELBER, G. Precise convex hull computation for freeform models using a hierarchical gauss map and a coons bounding volume hierarchy. *Computer-Aided Design*, v. 46, p. 252 – 257, 2014. doi: 10.1016/j.cad.2013.08.041.

186 KARAVELAS, M. I.; SEIDEL, R.; TZANAKI, E. Convex hulls of spheres and convex hulls of disjoint convex polytopes. *Computational Geometry*, v. 46, n. 6, p. 615 – 630, 2013. doi: 10.1016/j.comgeo.2013.02.001.

187 GONZÁLEZ, R. C.; WOODS, R. E. *Digital image processing. 2nd ed.* Ann Arbor, Michigan: Prentice-Hall, 2002. 793 p. ISBN: 0201180758, 9780201180756.

188 BLUM, H. Biological shape and visual science (part I). *Journal of Theoretical Biology*, v. 38, n. 2, p. 205 – 287, 1973. doi: 10.1016/0022-5193(73)90175-6.

189 JAIN, A. K. *Fundamentals of digital image processing*. Upper Saddle River, NJ: Prentice-Hall Englewood Cliffs, 1989. v. 3. ISBN: 0-13-336165-9.

190 GATRAM, R. M. B.; BABU, B. R.; SRIKRISHNA, A.; RAO, N. V. Shape matching and recognition using hybrid features from skeleton and boundary. *International Journal of Computers & Technology*, v. 7, n. 2, p. 558–564, 2013. Disponível em: <http://cirworld.com/index.php/ijct/article/view/1488>. Acesso em: 25 fev. 2014.

191 KUMAR, K. M.; KANTHAVEL, R. Analysis of various road pattern recognition methods for satellite images. *International Journal of Advanced Research in Electronics and Communication Engineering*, v. 2, n. 3, p. 277–283, 2013. Disponível em: <http://ijarece.org/wp-content/uploads/2013/08/IJARECE-VOL-2-ISSUE-3-277-283.pdf>. Acesso em: 25 fev. 2014.

192 CHI, Z. Data management for live plant identification. In: FENG, D.; SIU, W.-C. Z. H.-J. (Ed.) *Multimedia information retrieval and management, signals and communication technology*. Berlin Heidelberg: Springer, 2003. p. 432–457. doi: 10.1007/978-3-662-05300-3_20.

193 LAM, L.; SUEN, C. Y. An evaluation of parallel thinning algorithms for character recognition. *Transactions on Pattern Analysis and Machine Intelligence*, v. 17, n. 9, p. 914–919, 1995. doi: 10.1109/34.406659.

194 OGUZ, S.; ONURAL, L. An automated system for design-rule-based visual inspection of printed circuit boards. In: IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION, 1991, Sacramento, CA, USA. *Proceedings...* Sacramento, CA, USA: IEEE, c1991. p. 2696–2701, doi: 10.1109/ROBOT.1991.132038.

195 PUDNEY, C. Distance-ordered homotopic thinning: a skeletonization algorithm for 3d digital images. *Computer Vision and Image Understanding*, v. 72, n. 3, p. 404–413, 1998. doi: 10.1006/cviu.1998.0680.

196 FUJIYOSHI, H.; LIPTON, A. J. Real-time human motion analysis by image skeletonization. In: WORKSHOP ON APPLICATIONS OF COMPUTER VISION, 3., 1998, Princeton, NJ, USA. *Proceedings...* Princeton, NJ, USA: IEEE, c1998. p. 15–21.

197 LAM, L.; LEE, S.-W.; SUEN, C. Thinning methodologies-a comprehensive survey. *Transactions on Pattern Analysis and Machine Intelligence*, v. 14, n. 9, p. 869–885, 1992. doi: 10.1109/34.161346.

198 NIBLACK, C.; GIBBONS, P. B.; CAPSON, D. W. Generating skeletons and centerlines from the distance transform. *Graphical Models and Image Processing*, v. 54, n. 5, p. 420 – 437, 1992. doi: 10.1016/1049-9652(92)90026-T.

199 KRINIDIS, S.; CHATZIS, V. A skeleton family generator via physics-based deformable models. *Transactions on Image Processing*, v. 18, n. 1, p. 1–11, 2009. doi: 10.1109/TIP.2008.2007351.

200 DIMITROV, P.; DAMON, J. N.; SIDDIQI, K. Flux invariants for shape. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 1., 2003, Madison, Wisconsin, EUA. *Proceedings...* Madison, Wisconsin, EUA: IEEE, c2003. p. I–835–I–841.

201 OGNIEWICZ, R.; ILG, M. Voronoi skeletons: theory and applications. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 1992, Champaign, IL, EUA. *Proceedings...* Champaign, IL, EUA: IEEE, c1992. p. 63–69.

202 LAKSHMI, J. K.; PUNITHAVALLI, M. A survey on skeletons in digital image processing. In: INTERNATIONAL CONFERENCE ON DIGITAL IMAGE PROCESSING, 2009, Bangkok, Thailand. *Proceedings...* Bangkok, Thailand: IEEE, c2009. p. 260–269.

203 SHEWCHUK, J. R. Triangle: a two-dimensional quality mesh generator and Delaunay triangulator. Disponível em: <http://www.cs.cmu.edu/ quake/triangle.html> Acesso em: 25 fev. 2014.

204 YANG, Y.-J.; ZHANG, H.; YONG, J.-H.; ZENG, W.; PAUL, J.-C.; SUN, J. Constrained delaunay triangulation using delaunay visibility. In: INTERNATIONAL CONFERENCE ON ADVANCES IN VISUAL COMPUTING, 2., 2006, Lake Tahoe, Nevada, EUA. *Proceedings...* Berlin, Heidelberg: Springer, c2006. p. 682–691.

205 SIROVICH, L.; KIRBY, M. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, v. 4, n. 3, p. 519–524, 1987. doi: 10.1364/JOSAA.4.000519.

206 KIRBY, M.; SIROVICH, L. Application of the karhunen-loeve procedure for the characterization of human faces. *Transactions on Pattern Analysis and Machine Intelligence*, v. 12, n. 1, p. 103–108, 1990. doi: 10.1109/34.41390.

207 MASSACHUSETTS INSTITUTE OF TECHNOLOGY. Alex P. Pentland; Mathew Turk. *Face recognition system*. US Patent 5,164,992, 1 Nov. 1990, 17 Nov. 1992. Disponível em: <http://www.google.com/patents/US5164992>. Acesso em: 25 fev. 2014.

208 PEARSON, K. LIII. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, v. 2, n. 6, p. 559–572, 1901. doi: 10.1080/14786440109462720.

209 HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, v. 24, n. 6, p. 417, 1933. doi: 10.1037/h0071325.

210 KÜNDIG, W. A least square fit program. *Nuclear Instruments and Methods*, v. 75, n. 2, p. 336–340, 1969. doi: 10.1016/0029-554X(69)90624-7.

211 BELHUMEUR, P. N.; HESPANHA, J. A. P.; KRIEGMAN, D. J. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *Transactions on Pattern Analysis and Machine Intelligence*, Washington, DC, USA, v. 19, n. 7, p. 711–720, 1997. doi: 10.1109/34.598228.

212 FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

213 SILLA JR, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, v. 22, n. 1-2, p. 31–72, 2011. doi: 10.1007/s10618-010-0175-9.

214 WAHBA, G. *A survey of some smoothing problems and the method of the generalized cross-validation for solving them*. Madison, Wisconsin, USA: University of Wisconsin, Department of Statistics, 1976. 19 p. Disponível em: <https://getinfo.de/app/A-survey-of-some-smoothing-problems-and-the-method/id/TIBKAT%3A646616218>. Acesso em: 25 fev. 2014.

215 AL-NABI, D. L. A.; AHMED, S. S. Survey on classification algorithms for data mining: (comparison and evaluation). *Computer Engineering and Intelligent Systems*, v. 4, n. 8, p. 18–24, 2013. Disponível em: <http://www.iiste.org/Journals/index.php/CEIS/article/view/6575>. Acesso em: 25 fev. 2014.

216 GEORGE MASON INTELLECTUAL PROPERTIES, inc. Fayin Li; Harry Wechsler. *Open set recognition using transduction*. US Patent 7,492,943, 10 Mar. 2005, 17 Fev. 2009. Disponível em: <http://www.google.com/patents/US7492943>. Acesso em: 25 fev. 2014.

217 DEWILDE, B. Data science rules. Disponível em <http://datasciencerules.blogspot.com.br/2012/10/classification-of-hand-written-digits-3.html>. Acesso em fev. 2014.

218 PATWARDHAN, K. S.; ROY, S. D. Hand gesture modelling and recognition involving changing shapes and trajectories, using a predictive eigentracker. *Pattern Recognition Letters*, v. 28, n. 3, p. 329 – 334, 2007. doi: 10.1016/j.patrec.2006.04.002.

219 AHMAD, T.; TAYLOR, C.; LANITIS, A.; COOTES, T. Tracking and recognising hand gestures, using statistical shape models. *Image and Vision Computing*, v. 15, n. 5, p. 345 – 352, 1997. doi: 10.1016/S0262-8856(96)01136-5.

220 DARRELL, T.; PENTLAND, A. Space-time gestures. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 1993, New York. *Proceedings...* New York: IEEE, c1993. p. 335–340.

221 SUMA, E.; LANGE, B.; RIZZO, A.; KRUM, D.; BOLAS, M. Faast: the flexible action and articulated skeleton toolkit. In: VIRTUAL REALITY CONFERENCE (VR), 2011, Singapore. *Proceedings...* Singapore: IEEE, c2011. p. 247–248.

222 ALEXIADIS, D. S.; KELLY, P.; DARAS, P.; O'CONNOR, N. E.; BOUBEKEUR, T.; MOUSSA, M. B. Evaluating a dancer's performance using kinect-based skeleton tracking. In: INTERNATIONAL CONFERENCE ON MULTIMEDIA, 19., Scottsdale, Arizona, USA. *Proceedings...* New York, NY, USA: ACM, c2011. p. 659–662.

223 KESKIN, C.; KIRAC, F.; KARA, Y.; AKARUN, L. Sigmanil - the most powerful vision framework for natural user interfaces. Disponível em: <http://www.sigmanil.com/>. Acesso em: 25 fev. 2014.

224 OIKONOMIDIS, I.; KYRIAZIS, N.; ARGYROS, A. Kinect 3d hand tracking. Disponível em: <http://cvrlcode.ics.forth.gr/handtracking/>. Acesso em: 25 fev. 2014.

225 SOHN, M.-K.; LEE, S.-H.; KIM, D.-J.; KIM, B.; KIM, H. Handgket - hand gesture key emulation toolkit. Disponível em: <https://sites.google.com/site/kinectapps/handgket>. Acesso em: 25 fev. 2014.

226 Flutter - control music and movies with gestures. Disponível em: <https://flutterapp.com/>. Acesso em: 25 fev. 2014.

227 HAND gesture recognition - enables human-computer interaction with hand motions and gestures using standard USB web camera. Disponível em: <http://www.eyedea.cz/hand-gesture-recognition/>. Acesso em: 25 fev. 2014.

228 GRAB detector - openNI. Disponível em: <http://www.openni.org/files/grab-detector/>. Acesso em: 25 fev. 2014.

229 IGESTURE3D - openNI. Disponível em: <http://www.openni.org/files/igesture3d/>. Acesso em: 25 fev. 2014.

230 TIPTEP - touclees HCI | virtual reality - touchlees games. Disponível em: <http://tiptep.com/>. Acesso em: 25 fev. 2014.

231 SOHN, M.-K.; LEE, S.-H.; KIM, D.-J.; KIM, B.; KIM, H. 3D hand gesture recognition from one example. In: INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS (ICCE), 2013, Las Vegas, USA. *Proceedings...* Las Vegas, USA: IEEE, c2013. p. 171–172.

232 GRAY, J. *Benchmark handbook*: for database and transaction processing systems. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. ISBN: 1558601597.