Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Alinhamento de Imagens de Profundidade com Aplicação no Reconhecimento da Língua de Sinais

Juarez Paulino da Silva Júnior

Brasília
2014

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Alinhamento de Imagens de Profundidade com Aplicação no Reconhecimento da Língua de Sinais

Juarez Paulino da Silva Júnior

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Orientador
Prof. Dr. Jacir Luiz Bordim

Coorientador
Prof. Dr. Marcus Vinicius Lamar

Brasília
2014

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Coordenador: Prof. Dr. Ricardo Pezzuol Jacobi

Banca examinadora composta por:

Prof. Dr. Jacir Luiz Bordim (Orientador) — CiC/UnB
Prof. Dr. Hani Camille Yehia — DELT/UFMG
Prof. Dr. Bruno Luiggi Macchiavello Espinoza — CiC/UnB

CIP — Catalogação Internacional na Publicação

da Silva Júnior, Juarez Paulino.

Alinhamento de Imagens de Profundidade com Aplicação no Reconhecimento da Língua de Sinais / Juarez Paulino da Silva Júnior. Brasília : UnB, 2014.
96 p. : il. ; 29,5 cm.

Tese (Mestrado) — Universidade de Brasília, Brasília, 2014.

1. Alinhamento ICP, 2. Casamento de Modelos, 3. Correspondência de padrões, 4. Interface Natural com o Usuário, 5. Reconhecimento Automático da Língua de Sinais, 6. Sensores de Profundidade

CDU 004

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Alinhamento de Imagens de Profundidade com Aplicação no Reconhecimento da Língua de Sinais

Juarez Paulino da Silva Júnior

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Prof. Dr. Jacir Luiz Bordim (Orientador)
CiC/UnB

Prof. Dr. Hani Camille Yehia
DELT/UFMG

Prof. Dr. Bruno Luiggi Macchiavello Espinoza
CiC/UnB

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Programa de Pós-Graduação em Informática

Brasília, 24 de fevereiro de 2014

Agradecimentos

Após cumprir esta jornada de conhecimentos, gostaria de demonstrar meu respeito e gratidão a todos aqueles que estiveram ao meu lado e compartilharam experiências ao longo do processo: família, amigos, colegas e professores.

Agradeço em especial a minha mãe, Aramildes de Sousa Silva, a meu pai, Juarez Paulino da Silva, e aos meus irmãos, Luciano e Mateus, que estiveram sempre comigo, acreditando e confiando em minha superação, principalmente nos momentos mais difíceis.

Agradeço ao meu orientador, prof. dr. Jacir Bordim, e ao meu coorientador, prof. dr. Marcus Lamar, pelo tempo e dedicação que se dispuseram a me auxiliar, e por seus esclarecimentos e direcionamentos apontados. Agradeço, ainda, aos demais membros da banca, prof. dr. Hani Yehia e prof dr Bruno Espinoza, por compartilharem suas experiências e contribuírem de forma positiva com a versão final do trabalho.

Por fim, porém de igual importância, agradeço a meus companheiros do laboratório COMNET, aos amigos do "curso" e "fora" do curso, e aos queridos professores que tive o privilégio de conhecer na Universidade de Brasília. Nunca esquecerei dos bons momentos convividos e as experiências profissionais e pessoais trocadas.

Resumo

Gestos são utilizados desde tempos remotos como um mecanismo natural de comunicação. Como elemento de exteriorização da cultura surda, as línguas de sinais (línguas gestuais) possuem um importante papel na formação de uma unidade social. Neste contexto, sistemas de reconhecimento automático das línguas de sinais podem ser valiosos instrumentos de integração, ao passo que atenuam as barreiras impostas e estreitam os laços culturais entre surdos e ouvintes. Recentemente, surgiram novas pesquisas nesta linha que utilizam os chamados sensores RGB-D. Estes sensores caracterizam-se por serem de baixo custo e fácil uso, além de permitirem a aquisição de imagens de profundidade em tempo-real. Por sua vez, estas imagens carregam informações da localização espacial dos objetos da cena, simplificam tarefas de pré-processamento e contribuem para a proposição de novas metodologias de reconhecimento. Este trabalho propõe um sistema de reconhecimento automático das 26 posturas estáticas representantes das letras dos alfabetos manuais: da Língua de Sinais Americana (*ASL*), e da Língua Brasileira de Sinais (*Libras*). Para alcançar este objetivo, a metodologia do sistema emprega um sensor RGB-D na fase de aquisição de dados; e, de posse das imagens de profundidade, aplica a combinação da estratégia de Casamento de Modelos com o algoritmo de alinhamento *Iterative Closest Point* (*ICP*) na fase de reconhecimento. Como contribuição deste trabalho, a técnica *ICP* é aprimorada de forma a verificar possíveis parâmetros de entrada e saída no alinhamento de instâncias de teste com a base de modelos. Em seguida, utiliza estes parâmetros como determinantes da acurácia e eficiência do reconhecimento. Além disto, a estratégia de Casamento de Modelos é aperfeiçoada de forma a considerar partições de imagens aleatoriamente escolhidas das classes de modelos, visando reduzir o tempo de reconhecimento e aproximando a metodologia aos contextos de tempo-real. Os resultados apresentados mostram que o algoritmo *ICP* pode ser utilizado para produzir casamentos corretos entre as classes do alfabeto, mesmo quando um conjunto próximo (ambíguo) de posturas gestuais é aplicado. Quanto à acurácia da metodologia implementada, estes resultados indicam um desempenho máximo obtido de $99{,}04\%$ de taxa de acerto no reconhecimento da *ASL* e de $99{,}62\%$ para a *Libras*. Verificou-se ainda que o sistema atingiu seu melhor desempenho em eficiência com frequência média de processamento de $7{,}41$ *FPS*, utilizando uma máquina de processador único de 2,4 GHz.

Palavras-chave: Alinhamento ICP, Casamento de Modelos, Correspondência de padrões, Interface Natural com o Usuário, Reconhecimento Automático da Língua de Sinais, Sensores de Profundidade

Abstract

Gestures are used since ancient times as a natural mechanism of communication. As an element of externalisation of the deaf culture, sign languages (gestural languages) have an important role in forming a social unit. In this context, automatic sign language recognition systems can be valuable tools of integration, while mitigating the barriers and strengthening cultural ties between deaf and hearing people. Recently, new research has emerged in this area which uses the so called RGB-D sensors. These devices are characterized by the low cost and ease of use, also allowing the depth image acquisition in real-time. In turn, these images carry information of the spatial location of the objects in the scene, simplify preprocessing tasks and contribute to propose new recognition methodologies. This work proposes a system for automatic recognition of the 26 static postures representatives of the letters in the manual alphabets of: the American Sign Language (*ASL*), and the Brazilian Sign Language (*Libras*). To achieve this objective, the system methodology employs an RGB-D sensor in the phase of data acquisition; and, once in possession of the depth images, applies the combination of the *Template Matching* strategy with the *Iterative Closest Point* (ICP) alignment algorithm in the recognition phase. As contributions of this work, the *ICP* technique is improved in order to verify possible input and output parameters in the alignment of test instances with the model database. Then, it uses these parameters as accuracy and efficiency determinants of the recognition. Moreover, the Template Matching strategy is enhanced to consider image partitions randomly chosen from the model classes, aiming time reduction of recognition and approaching the methodology to real-time contexts. The presented results show that the *ICP* algorithm can be used to produce correct matches between the alphabet classes, even when a close (ambiguous) set of sign postures is applied. Regarding the accuracy of the implemented methodology, these results indicate a maximum performance of 99.04% of success rate in the *ASL* recognition and of 99.62% for *Libras*. It was also verified that the system reach its best efficiency performance with an average processing frame frequency of 7.41 *FPS*, using a 2.4 GHz single processor based machine.

Keywords: Automatic Sign Language Recognition, Depth Sensors, ICP Alignment, Natural User Interface, Pattern Matching, Template Matching.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Siglas e Acrônimos

2.5D	Duas Dimensões e meia ou Pseudo-3D.
2D	Bidimensional.
3D	Tridimensional.
ANMM	Maximização da Margem de Vizinhança Média (do Inglês, *Average Neighborhood Margin Maximization*).
ANN	Redes Neurais Artificiais (do Inglês, *Artificial Neural Networks*).
API	Interface de Programação para Aplicações (do Inglês, *Application Programming Interface*).
ASL	Língua de Sinais Americana (do Inglês, *American Sign Language*).
BSL	Língua de Sinais Britânica (do Inglês, *British Sign Language*).
CMOS	Semicondutor Metal-Óxido Complementar (do Inglês, *Complementary Metal-Oxide-Semiconductor*).
DOF	Graus de Liberdade (do Inglês, *Degrees of Freedom*).
FPS	Quadros por Segundo (do Inglês, *Frames per Second*).
GPGPU	Unidade de Processamento Gráfico de Propósito Geral (do Inglês, General-Purpose Computing on Graphics Processing Units).
HMM	Modelos Ocultos de Markov (do Inglês, *Hidden Markov Models*).
ICP	Busca Iterativa do Ponto mais Próximo (do Inglês, *Iterative Closest Point*).
INES	Instituto Nacional da Educação de Surdos.
IR	Infra-vermelho (do Inglês, *Infrared*).

JSL	Língua de Sinais Japonesa (do Inglês, *Japanese Sign Language*).
K-D Tree	Árvore K-Dimensional (do Inglês, *K-Dimensional Tree*).
K-NN	K-Vizinhos mais Próximos (do Inglês, *K-Nearest Neighbors*).
LADAR	Detecção e Obtenção de Distância por Laser (do Inglês, *Laser Detection and Ranging*).
LCP	Maior Subconjunto Comum entre Pontos (do Inglês, *Largest Common Pointset*).
Libras	Língua Brasileira de Sinais.
LSF	Língua de Sinais Francesa (do Francês, *Langue des Signes Française*).
NiTE	*Middleware* de Interação Natural (do Inglês, *Natural Interaction Middleware*).
NUI	Interface Natural de Usuário (do Inglês, *Natural User Interface*).
OpenNI	Interação Natural Livre (do Inglês, *Open Natural Interaction*).
PCA	Análise em Componentes Principais (do Inglês, *Principal Component Analysis*).
PC	Computador Pessoal (do Inglês, *Personal Computer*).
pixel	Elemento de Figura (do Inglês, *Picture Element*).
PSL	Língua de Sinais Portuguesa (do Inglês, *Portuguese Sign Language*).
RDF	Floresta de Decisão Aleatória (do Inglês, *Random Decision Forest*).
RGB	Referente à formação da imagem de intensidade luminosa (do Inglês, *red*, *green*, *blue*).
RGB-D	Referente à formação da imagem de intensidade (canais *RGB*) e da imagem de profundidade (canal D, do Inglês *depth*).

RMS	Valor Quadrático Médio ou Raíz da Média Quadrática (do Inglês, *Root Mean Square*).
SDK	Conjunto de Desenvolvimento de Software (do Inglês, *Software Development Kit*).
SVM	Máquina de Vetores Suporte (do Inglês, *Support Vector Machine*).
TDD	Telefone de Texto (do Inglês, *Telecommunications Device for the Deaf*).
USB	Barramento Serial Universal (do Inglês, *Universal Serial Bus*).
VGA	Arranjo Gráfico de Vídeo (do Inglês, *Video Graphics Array*).
voxel	Elemento de volume (do Inglês, *Volumetric Element*).

Capítulo 1

Introdução

Gestos são recursos legítimos utilizados para expressar ideias e transmitir informações. O uso de gestos é considerado uma das formas mais primitivas de comunicação. Antes mesmo de aprender a falar, o homem já é capaz de realizar e interpretar algumas das expressões gestuais que observa ao seu redor. Parte desta pronta adaptação se deve a um complexo sistema de visão, do qual o homem se vale para extração de características e reconhecimento de padrões das representações do mundo [1]. Infelizmente, os sinais emitidos pelo homem, como gestos ou expressões corporais, ainda não são igualmente assimiláveis pelos atuais sistemas de computação [2]. Tal limitação motiva o avanço e crescimento de pesquisas de *hardware* e *software* voltadas para áreas como a interação humano-computador ou de aplicação de interfaces naturais com o usuário [3, 4].

Neste contexto, do conjunto de avanços mais recentes quanto à aplicação de sistemas baseados em visão, destaca-se a introdução dos sensores *RGB-D*, tais como o *Microsoft® Kinect* [5], liberado para pesquisas no ano de 2010. Estes sensores são assim chamados por permitirem a obtenção em conjunto: (i) de imagens contendo informação de intensidade luminosa (canais *RGB*, do Inglês *red*, *green*, *blue*); e (ii) de "imagens de profundidade" (canal D, do Inglês *depth*), que associam para cada *pixel*, um valor correspondente da distância dos objetos em cena ao sensor. As principais vantagens do uso destes dispositivos advém de características como [6]: (a) baixo custo; (b) facilidade de uso; e (c) aquisição de quadros em tempo-real. Por sua vez, sob uma perspectiva social, estas são também propriedades desejáveis, quando não essenciais, para a construção de sistemas computacionais voltados ao uso e integração na sociedade.

1.1 Visão Geral

O reconhecimento automático de gestos é um campo de pesquisa relacionado às mais diversas aplicações tais como [3]: assistência ao ensino a distância, manipulação de ambientes virtuais, monitoramento do grau de alerta de motoristas, monitoramento médico de pacientes e técnicas de identificação forense. Deste conjunto de aplicações, o "reconhecimento da língua de sinais" por meio de sistemas computacionais apresenta um importante apelo social ao viabilizar:

(i) uma maior integração da comunidade surda e de deficientes auditivos;

(ii) o auxílio em sistemas de ensino específicos;

(iii) a disseminação da língua de sinais; e

(iv) a instrumentalização de uma sociedade capacitada a compreender sem grandes esforços este importante recurso de comunicação.

Catalogam-se hoje pelo menos 130 diferentes línguas de sinais espalhadas pelo mundo, com a expectativa de que número real seja ainda maior [7]. Em cada uma destas línguas, uma quantidade enorme de gestos estáticos e dinâmicos são aplicáveis, permitindo equipará-las, quanto ao poder de expressão, ao das línguas orais praticadas. Além disso, a consolidação de uma língua de sinais costuma derivar de uma longa história de formação, carregando, em si, parte da cultura de toda uma comunidade.

No Brasil, o censo demográfico de 2000 contou aproximadamente 166 mil pessoas com surdez completa. Destas pessoas, contou-se que apenas 344 cursavam universidades brasileiras [8]. A interpretação estatística destes dois números revela características impactantes quanto ao ensino dos surdos no Brasil, isto é, até o ano de 2000 verificava-se que apenas $0{,}22\%$ da comunidade surda possuía vínculo ativo como estudantes das universidades do país. Considerando que o mesmo censo contou um percentual de $3{,}18\%$ de brasileiros, em geral, cursando as universidades brasileiras [9], evidencia-se uma possível afronta ao direito de igualdade no rol de princípios fundamentais impetrado na Carta Magna brasileira.

Dessa forma, entende-se, pelos motivos já indicados, que os sistemas de reconhecimento automático aplicados às línguas de sinais sejam uma opção atrativa na tentativa de reduzir ou amenizar parte das desigualdades sociais, e de aumentar as possibilidades de interação entre as culturas surda e ouvinte.

À parte das justificativas sociais apontadas, a investigação de técnicas para o reconhecimento automático da língua de sinais apresenta relevante motivação técnica, da qual se apontam os seguintes exemplos:

- Hoje é possível encontrar uma variedade de soluções [10, 11, 12] e produtos [13] que sejam eficazes especialmente no reconhecimento de gestos das línguas de sinais. Estas ferramentas normalmente utilizam um conjunto adicional de acessórios, tal como "luvas de dados", que permite aprimorar a inferência de características extraídas dos gestos. Entretanto, verifica-se um grupo de problemas que limitam o uso destas ferramentas, tais como [3]: (a) maior custo; (b) complicada configuração ou calibração inicial; e (c) redução da naturalidade de interação do usuário com o sistema.

- Como alternativa às limitações identificadas no item anterior, vários sistemas que utilizam imagens de intensidade e são baseados unicamente em visão foram propostos [14, 15, 16]. Nestas pesquisas as informações características dos gestos são extraídas com respeito a elementos observados na imagem como: forma, cor, textura, movimento, ou contorno [3]. Porém um segundo conjunto de restrições pode ser identificado nestas pesquisas, incluindo-se [17]: (a) segmentação ineficiente ou ineficaz; (b) inaptidão no tratamento de oclusões; e (c) não-adaptabilidade a variações do ambiente ou restrições ao grau de intensidade luminosa.

- Recentemente, sensores *RGB-D* de fácil acesso e baixo custo surgiram, permitindo que a maior parte das limitações acima fossem superadas ou, pelo menos, abordáveis. A possibilidade de tratamento destes problemas com o uso de imagens de profundidade motivou a implementação de diversos trabalhos originais em temas da visão computacional como: reconstrução em três dimensões (*3D*) de superfícies densas [18], rastreamento *3D* de objetos, mãos e corpo humano [19, 20], e interação com ambientes de realidade

aumentada [19, 21]. No entanto, ainda se observa certa carência por sistemas que utilizem as imagens de profundidade no contexto de reconhecimento da língua de sinais.

De forma geral, este trabalho investiga e propõe o desenvolvimento de sistemas baseados em visão que utilizem imagens de profundidade no reconhecimento da língua de sinais. Como mecanismo de integração na sociedade, exige-se que tais sistemas sejam de fácil acesso e configuração, ao mesmo tempo que implementem estratégias inteligentes no atendimento de requisitos específicos como acurácia, eficiência e interface natural com o usuário.

1.2 Objetivos e Metodologia

O processo de interação pela língua de sinais é uma poderosa ferramenta de comunicação, que permite aos seus interlocutores compreenderem-se mutuamente e expor seus argumentos e convicções. Dessa forma, sistemas baseados em visão para o reconhecimento automático de gestos da língua de sinais devem ser capazes de lidar com um grande conjunto de posturas estáticas ou dinâmicas, e rastrear não só gestos com as mãos, mas também com a combinação de diferentes partes do corpo.

Este trabalho limita o problema de reconhecimento da língua de sinais, e apresenta como objetivo geral construir um sistema de identificação das posturas estáticas para as letras dos alfabetos manuais. Como objetos de estudo são utilizados os alfabetos manuais da "Língua de Sinais Americana" e da "Língua Brasileira de Sinais". Adicionalmente, condiciona-se o sistema à aquisição e processamento realizados unicamente com o uso de imagens de profundidade, adquiridas a partir do sensor *Microsoft® Kinect* [5].

Para atingir este objetivo, a metodologia proposta adota o uso combinado de duas técnicas: o (i) algoritmo de alinhamento pareado "Busca Iterativa do Ponto mais Próximo" (do Inglês, *Iterative Closest Point – ICP*), que ao ser aprimorado permite inferir similaridades na comparação de duas imagens de profundidade; e a (ii) estratégia de "Casamento de Modelos" (do Inglês, *Template Matching*), utilizada para realizar a classificação de uma dada imagem de teste contra uma base de modelos.

Embora as duas técnicas abordadas na metodologia sejam largamente referenciadas na literatura [6, 22, 23, 24, 25], a proposta deste trabalho é inovadora ao passo que introduz aprimoramentos no uso de ambas para o reconhecimento da língua de sinais:

- Da parte do Casamento de Modelos, é proposta a aplicação de um método de força bruta integrado ao alinhamento *ICP* para o reconhecimento do alfabeto manual. Em seguida, propõe-se uma técnica aproximada de partição e comparação de amostras das classes de reconhecimento. O objetivo desta última é reduzir o custo computacional do método, e aproximá-lo de implementações práticas ao contexto de tempo-real.
- O algoritmo *ICP* serve precipuamente à obtenção do alinhamento entre dois conjuntos de pontos. O seu uso no reconhecimento de padrões é bastante restrito na literatura, sendo apenas brevemente explorado em aplicações como de biometria, onde dificilmente se aplicam as exigências de tempo-real. De fato, até onde se investigou, nenhuma pesquisa relacionada utilizou com sucesso o algoritmo *ICP* no reconhecimento da língua de sinais. Na verdade, do levantamento bibliográfico, apenas o trabalho de Trindade *et al.* [23] propôs o seu uso neste escopo. No entanto os argumentos apresentados foram desfavoráveis à

aplicação do alinhamento *ICP*, concluindo que a técnica não seria capaz de realizar o casamento de padrões para algumas classes de letras da Língua de Sinais Americana. Assim, constitui uma hipótese da pesquisa averiguar a aplicabilidade do algoritmo e apresentar propostas de melhorias com o objetivo de torná-lo eficiente e eficaz ao computar similaridades entre dois padrões de imagens de profundidade.

Como pode ser observado, a metodologia utilizada neste trabalho possui forte inclinação para avaliar a "acurácia", estimada pela taxa de acerto, e "eficiência", custo computacional, das técnicas propostas no escopo de reconhecimento da língua de sinais. Dessa forma, os seguintes objetivos específicos foram estabelecidos:

- Avaliar o estado da arte em técnicas de reconhecimento da língua de sinais, em especial, aquelas voltadas para a identificação do alfabeto manual e que utilizem imagens de profundidade.
- Estudar em detalhes o algoritmo de alinhamento *ICP* de forma a identificar sob quais aspectos este pode ser utilizado para o casamento de padrões. Após a análise, apresentar proposta de melhorias à técnica.
- Estudar em detalhes a estratégia de Casamento de Modelos de forma a identificar sob quais aspectos esta pode ser utilizada para o casamento de padrões. Após a análise, apresentar proposta de melhorias à técnica.
- Construir testes e verificar resultados para a metodologia implementada, permitindo certificar a relevância das contribuições introduzidas.

1.3 Contribuições

Deste trabalho resultaram as seguintes contribuições:

- um levantamento bibliográfico organizado dos métodos de aquisição de imagens de profundidade, incluindo a apresentação de uma taxonomia de classificações;
- a caracterização do cenário atual de reconhecimento da língua de sinais por sistemas baseados em visão e o levantamento das principais técnicas que abordam o problema com o uso de sensores de profundidade;
- a definição de uma arquitetura original de casamento de padrões, utilizada na comparação pareada de imagens de profundidade a partir de uma base de dados de posturas estáticas, sem esforço adicional de aprendizado;
- a proposição de aprimoramentos ao algoritmo de registro *ICP*, viabilizando sua aplicação para o reconhecimento automático de alfabetos manuais;
- a proposição de uma técnica de particionamento da base de dados empregada na estratégia conjunta de Casamento de Modelos e registro *ICP*, limitando o custo computacional da metodologia proposta e permitindo aproximá-la de aplicações para o contexto de tempo-real.
- variadas técnicas de construção de experimentos e análise de resultados que permitem avaliar soluções que apliquem a estratégia de Casamento de Modelos e inferir considerações sobre o reconhecimento dos alfabetos manuais.

1.4 Publicações Resultantes deste Trabalho

Os artigos descritos a seguir foram elaborados ao longo do Mestrado e forneceram os subsídios para a construção deste texto:

1. J. P. da Silva, M. V. Lamar, e J. L. Bordim. Accuracy and efficiency performance of the ICP procedure applied to sign language recognition. Em *CLEI Electronic Journal*, *aceito em 2014*.

2. J. P. da Silva, M. V. Lamar, e J. L. Bordim. A study of the ICP algorithm for recognition of the hand alphabet. Em *Computing Conference (CLEI), 2013 XXXIX Latin American*, páginas 1–9, 2013.

1.5 Estrutura do Documento

Esta dissertação está dividida em 6 capítulos, considerando esta introdução. O restante do documento está organizado da seguinte forma:

- O **Capítulo 2** contém uma síntese dos conceitos tido como fundamentais ao entendimento do escopo deste trabalho: apresenta-se um estudo dos sistemas baseados em visão, e em especial, quanto à aquisição de imagens de profundidade com o sensor *Kinect*; introduz-se, ainda, uma visão geral da aplicação destes sistemas para o reconhecimento de posturas da língua de sinais.

- O **Capítulo 3** apresenta a revisão teórica correlata à área de reconhecimento de gestos com imagens de profundidade. Esta revisão inclui um resumo do estado da arte das principais técnicas que melhor se adequam às propostas desta pesquisa; também detalha aspectos técnicos da estratégia de Casamento de Modelos e do algoritmo de registro *ICP*, principais componentes do sistema proposto.

- O **Capítulo 4** introduz a definição formal do problema de reconhecimento dos alfabetos manuais; descreve as contribuições originais deste trabalho quanto ao aprimoramento das técnicas aplicadas ao casamento de padrões; e explora a arquitetura geral da metodologia proposta, incluindo detalhes de sua implementação para a obtenção dos resultados.

- O **Capítulo 5** demonstra a efetividade do sistema proposto sob aspectos de acurácia e eficiência computacional no reconhecimento dos alfabetos manuais da Língua de Sinais Americana e da Língua Brasileira de Sinais. A metodologia implementada é avaliada sob diversos cenários, obtendo resultados que permitam clarificar e certificar a aplicação das contribuições descritas.

- Por fim, o **Capítulo 6** conclui esta dissertação, reunindo os principais aspectos abordados ao longo do texto; e indica as possíveis extensões e evoluções identificadas no escopo desta pesquisa.

Capítulo 2

Reconhecimento de Gestos por Sistemas Baseados em Visão

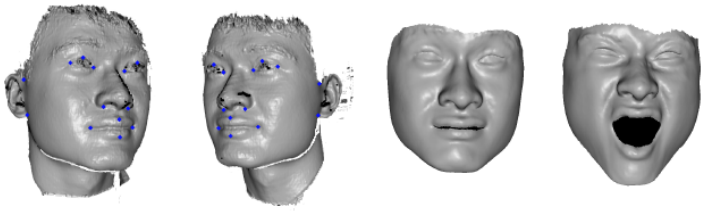
O homem é capaz de perceber com facilidade a estrutura tridimensional do mundo em sua volta. No entanto, a visão é tida como o sentido de maior complexidade e ainda hoje é um mistério o seu completo funcionamento [26]. A percepção visual é um mecanismo sofisticado, pois procura-se resolver um problema inverso pelo qual estimam-se elementos constitutivos – cores, localização e profundidade – dado um conjunto de informações desagregadas em uma cena observada. Em seguida, o sistema visual recorre a um conhecimento prévio baseado em noções físicas do mundo adquiridas ao longo da vida e o correlaciona a modelos probabilísticos para eleger a melhor escolha de interpretação. Além de interpretações estáticas para cada cena, este complexo sistema é capaz de associá-las em diferentes instantes de tempo, auxiliando assim na dinâmica de iterações do homem com o mundo [1].

Conjuntamente ao estudo de se mapear a visão humana, os sistemas computacionais baseados em visão surgiram com o objetivo de desenvolver técnicas para recuperar e interpretar informações do mundo real a partir de projeções em imagens obtidas por câmeras e sensores [2]. Pesquisas nesta área incluem temas como *image stitching* (Figura 2.1a), reconstrução em três dimensões (*3D*) (Figura 2.1b), e correspondência ou reconhecimento de imagens (Figura 2.1c). Com maior ênfase no tópico desta dissertação, há o interesse em pesquisas que permitam a aquisição espacial (*3D*) de segmentos de uma imagem e o seu reconhecimento em determinadas classes (gestos) segundo um conjunto de posturas previamente definidas (Figura 2.1d).

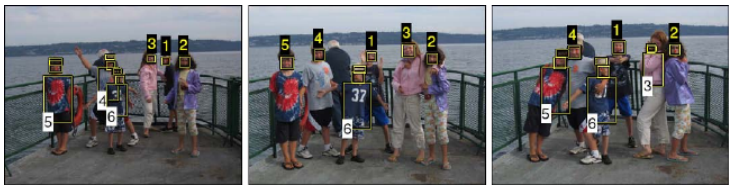
Enquanto significativos progressos da máquina em relação ao homem são observados em outros campos (poder de processamento, armazenamento de dados, entre outros), os sistemas baseados em visão encontram uma grande margem de diferença quando comparados à visão humana. Este contraste é ainda mais perceptível quando se aborda o tema de reconhecimento de gestos, onde diversos subproblemas são encontrados, tais como [17]: localização e segmentação de partes do corpo, tratamento de oclusões, mudanças da luz ambiente, ou rastreamento e reconhecimento de múltiplos gestos.

Este capítulo apresenta um levantamento geral dos principais conceitos para o entendimento do escopo desta dissertação. A partir da revisão bibliográfica, são introduzidos conceitos associados às imagens digitais, à sua aquisição e tratamento e, em especial, às características e potenciais do uso de imagens de profundidade no reconhecimento da língua de sinais.

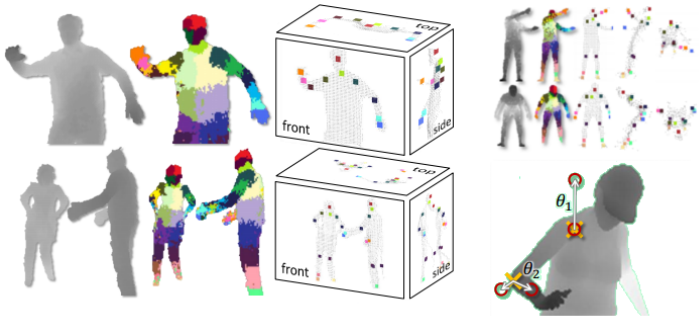
(a) *Image Stitching*: Ampliação de imagens por múltiplas vistas [27].



(b) Reconstrução e animação de faces *3D* [28].



(c) Reconhecimento e detecção de pessoas em imagens de intensidade [29].



(d) Reconhecimento de diferentes poses do corpo humano em tempo-real [30].

Figura 2.1: Exemplos de aplicações diversas com sistemas baseados em visão.

2.1 Imagens Digitais

"Imagens" são representações da informação visual presente no mundo. A formação de imagens, a partir de câmeras e sensores, remete ao estudo da geometria projetiva. Neste contexto, o princípio geral de qualquer câmera é utilizar um conjunto de lentes e receptores sensíveis que permita focalizar e captar projeções perspectivas (transformações projetivas) das informações do mundo real no plano de uma imagem [31].

Estruturalmente, uma "imagem digital" [32] pode ser vista como uma matriz bidimensional de números, onde cada célula desta matriz determina um *pixel* – elemento de figura (do Inglês, *picture element*). Cada *pixel* de uma imagem carrega uma parcela discreta de informação referente à cena representada. Assim a imagem digital caracteriza-se pela amostragem (número de *pixel*s)

e quantização (reserva de *bits* para representação do *pixel*). Neste contexto, a "resolução" de uma imagem é uma métrica sobre a capacidade da câmera ou do sensor em distinguir claramente os objetos em cena. Embora haja contra-indicações pelos padrões internacionais [33], o termo resolução se popularizou como expressão do tamanho de uma imagem digital em valores absolutos de *pixel*s do número de colunas e linhas da imagem [34]. Sob estas circunstâncias, a resolução pode ser aplicada para determinar a quantidade de detalhes em uma imagem. Na prática, uma imagem de alta resolução implica em um maior número de *pixel*s e, portanto, uma quantidade maior de informação sendo armazenada. De outra forma, uma imagem de baixa resolução pode indicar um conjunto menor de *pixel*s e uma representação pobre dos elementos de uma cena.

De um ponto de vista qualitativo, uma imagem digital é classificada pela natureza da informação que o *pixel* codifica. Para os objetivos desta dissertação, duas variantes são consideradas: imagens de intensidade luminosa e imagens de profundidade.

2.1.1 Imagens de Intensidade

Uma "imagem de intensidade" é definida como uma matriz bidimensional em que cada *pixel* codifica um valor de função para a intensidade luminosa capturada nos elementos de uma cena; cada valor, por sua vez, depende da iluminação ($i(x, y)$) e das propriedades de reflectância na superfície dos objetos ($r(x, y)$) [35]. Sendo M o número de linhas e N o número de colunas da matriz de uma imagem de intensidade, esta pode ser modelada matematicamente por uma função em duas variáveis, da seguinte forma:

$$f(x, y) = i(x, y)r(x, y), \tag{2.1}$$

onde $0 \leq x < M$ e $0 \leq y < N$. Matricialmente, sua representação é:

$$f(x, y) = \begin{pmatrix} f(0,0) & f(0,1) & \cdots & f(0,N-1) \\ f(1,0) & f(1,1) & \cdots & f(1,N-1) \\ \vdots & \vdots & \ddots & \vdots \\ f(M-1,0) & f(M-1,1) & \cdots & f(M-1,N-1) \end{pmatrix}. \tag{2.2}$$

Com base neste modelo, a função f pode ser interpretada sob diversas formulações matemáticas – os chamados "espaços de cores". Vários espaços de cores foram propostos, incluindo o RGB, HSV, YCrCb, YUV, escalas de cinza, entre outros [36]. A definição do espaço de cor pode ter uma grande influência nos algoritmos baseados puramente em imagens de intensidade. Um espaço que permita distinguir eficazmente as componentes de crominância das componentes de luminância são normalmente desejáveis, uma vez que tornam os algoritmos que o utilizam mais robustos e menos suscetíveis às variações de iluminação do ambiente onde a imagem fora adquirida [37].

2.1.2 Imagens de Profundidade

Para se recuperar a geometria *3D* de um objeto, faz-se necessário mais informações do que aquilo que é projetado no plano de uma imagem de intensidade. Há a necessidade de se estimar com maior ou menor precisão a distância dos elementos da cena para o sistema de referência do sensor de captura que gerou a imagem.

Tabela 2.1: Características e tendências do uso de imagens de intensidade e profundidade.

Imagens de Intensidade	**Imagens de Profundidade**
Menor dimensão do problema (2D)	Maior dimensão do problema (*3D*)
Menor complexidade em tempo e espaço	Maior complexidade em tempo e espaço[(1)]
Maior dependência da iluminação ambiente	Menor dependência da iluminação ambiente
Segmentação pelo espaço de cores	Segmentação pela diferença de profundidades

[(1)] A atribuição de uma maior complexidade de tempo e espaço verificada quanto ao uso de imagens de profundidade se deve à maior dimensão dos objetos inferidos na cena.

O termo "imagem de profundidade" ou "mapa de profundidade" é usado, sem perda de generalidade, para descrever uma imagem digital onde seu elemento mínimo, ao invés de conter um valor de intensidade (cor), carrega uma informação de distância em relação a um plano de coordenadas (usualmente o sistema de coordenadas discretas do dispositivo sensor). Sendo $d(i,j)$ o valor de profundidade (distância) do *pixel* (i,j) da imagem ao dispositivo sensor, M o número de linhas, N o número de colunas da matriz de uma imagem de profundidade,esta pode ser descrita matricialmente como:

$$d(x,y) = \begin{pmatrix} d(0,0) & d(0,1) & \cdots & d(0,N-1) \\ d(1,0) & d(1,1) & \cdots & d(1,N-1) \\ \vdots & \vdots & \ddots & \vdots \\ d(M-1,0) & d(M-1,1) & \cdots & d(M-1,N-1) \end{pmatrix}. \tag{2.3}$$

Assim, uma imagem de profundidade refere-se, normalmente, à imagem produzida pelo dispositivo sensor que contém informação espacial da distância dos objetos em cena.

Outros termos são também encontrados na literatura [2, 32, 38] em referência a este tipo de imagens tais como "mapas xyz", "nuvem de pontos *3D*" (do Inglês, 3D *data point cloud*), "perfis de superfície" (do Inglês, *surface mesh*) e "imagens *2.5D*". Em geral, as diferenças entre os termos dizem respeito a uma maior ou menor estrutura de organização da malha de pontos tridimensionais extraída.

As imagens de intensidade e as imagens de profundidade apresentam distintas propriedades e interpretações na representação de objetos do mundo real. Imagens de intensidade refletem a cor e intensidade luminosa dos objetos com coordenadas em *pixel*s relativas e restritas a um plano. Devido a esta restrição sobre o plano bidimensional, as imagens de intensidade são também, por vezes, denominadas imagens bidimensionais (imagens *2D*). Por outro lado, as imagens de profundidade permitem inferir a disposição espacial (*3D*) dos objetos contribuindo para distingui-los entre os diversos componentes em uma cena.

Sob o ponto de vista da natureza da informação codificada e suas características, o uso das imagens possui diferentes implicações para os algoritmos de sistemas baseados em visão. A Tabela 2.1 reporta comparativamente as principais tendências usualmente observadas na literatura com respeito ao uso de imagens de intensidade e profundidade.

Por fim, vale ressaltar que o uso de um tipo de imagem não inviabiliza a aplicação do outro. Na verdade, muitos trabalhos sugerem a aplicação conjunta [39] ou a fusão da informação [40, 41] destas imagens para a obtenção de maior acurácia e robustez em seus sistemas. Um dos focos deste trabalho está na análise e aplicação da informação de imagens de profundidade no

reconhecimento de gestos. Por isto, a próxima seção descreve resumidamente os principais métodos de aquisição destas imagens.

2.2 Métodos de Aquisição de Dados de Profundidade

Atualmente, a aquisição de imagens de intensidade é uma tarefa simples, amplamente estudada e acessível, dado o acentuado desenvolvimento das câmeras digitais. Muitas vezes é possível ter acesso a uma imagem de alta qualidade, utilizando até mesmo as câmeras comuns embutidas dos dispositivos portáteis, como celulares e *tablets*.

Em contrapartida, com uma busca cada vez maior pela imersão dos usuários no mundo digital, há uma demanda crescente por dispositivos de captura e representação da informação tridimensional (*3D*). Esta tendência motiva a evolução de diferentes sistemas baseados em visão, que, por sua vez, consideram os recentes sensores *RGB-D*, tais como o *Microsoft® Kinect* [5], promissores instrumentos na obtenção de imagens de profundidade.

2.2.1 Taxonomia dos Métodos de Aquisição

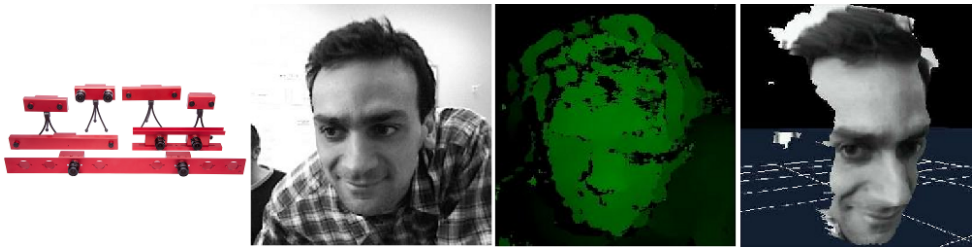
Em geral, os dispositivos sensores e câmeras digitais são classificados quanto sua natureza em: (i) "passivos", que buscam uma representação sem interferência direta no objeto; ou (ii) "ativos", que se aproximam fisicamente ou emitem energia sobre o objeto, retirando informações com base na resposta obtida [42].

Métodos Passivos: As técnicas passivas obtêm informação sobre a superfície de um objeto analisando imagens de intensidades capturadas por um ou mais dispositivos de sensoreamento óptico. Isto é possível graças a um conjunto de características visuais observadas nas imagens. Os principais métodos são listados a seguir:

- "estereoscopia ou triangulação passiva" [2, 31, 43, 44] (Figura 2.2a);
- "forma a partir de sombreamento" [45];
- "forma a partir de silhuetas e texturas" [46] (Figura 2.2b);
- "forma a partir de movimentos" [47];
- "focagem e desfocagem ativa" [48] (Figura 2.2c).

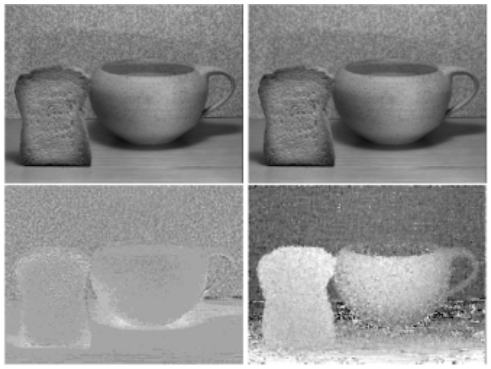
Dispositivos passivos normalmente possuem tecnologia de custo reduzido, mas utilizam softwares em alto nível para configuração do sistema; o que aumenta a complexidade do processo de aquisição e encarecem o custo do método como um todo.

Métodos ativos: Ao contrário dos métodos passivos, que exigem um intenso processamento computacional, as técnicas ativas normalmente controlam a incidência de radiação luminosa sobre os objetos por um sistema de projeção, procurando recuperar diretamente a informação *3D*. Neste quesito, o *laser* costuma ser a opção mais utilizada como radiação luminosa, pois é monocromático, estruturado (luz coerente) e facilmente direcionável (feixe estreito e concentrado).



(a) Visão estereoscópica: Videre Design system© [44].

(b) Modelo Deformável usando informações de silhueta e textura [46].



(c) Forma por focagem e desfocagem ativa [48].

(d) Estereoscopia aliada a um sistema ativo de projeção de padrões [28].

Figura 2.2: Aquisição de mapas de profundidade por métodos ativos e passivos.

A técnica ativa mais praticada, e com melhor custo-benefício em aplicações de curta distância, é a "iluminação estruturada", ou ainda "triangulação ativa" [28, 49] (Figura 2.2d). Esta é também a estratégia aplicada na extração de mapas de profundidade pelo *Microsoft® Kinect* [5], estudado na Subseção 2.2.2. Outros métodos ativos menos populares incluem os "radares baseados a *laser*" (*LADAR*) [50, 51], com aplicações diversas em navegação robótica; e a "interferometria de *Moiré*" [52]. Em geral, os digitalizadores ativos são capazes de extrair imagens com maior resolução, porém possuem um custo mais elevado. Além disso, a extração de modelos completos pode ser um processo dispendioso, pois requer várias iterações com diferentes posicionamentos do dispositivo sensor.

Uma outra possível classificação diferencia os digitalizadores que efetuam contato com a cena/objeto daqueles que não o fazem. Um método que efetua contato não precisa fisicamente entrar em contato com o objeto, mas apenas se aproximar de forma suficiente até que a geometria *3D* do objeto possa ser extraída. Alguns métodos que efetuam contato são capazes de recuperar toda a geometria *3D*, pois não são limitados a uma posição fixa do dispositivo sensor. Além

disso, possuem a vantagem de extrair dados em precisões submilimétricas (até em nível atômico – abaixo de $10^{-10}m$) [53]. No entanto, devido à evolução da tecnologia de aquisição, é comum que os métodos que efetuam contato levem desvantagem frente aos demais, pois são dispendiosos, podem comprometer a estrutura do objeto, e muitas vezes falham ao extrair uma superfície de amostragem igualmente densa [54, 55].

2.2.2 Sensor *RGB-D Microsoft® Kinect*

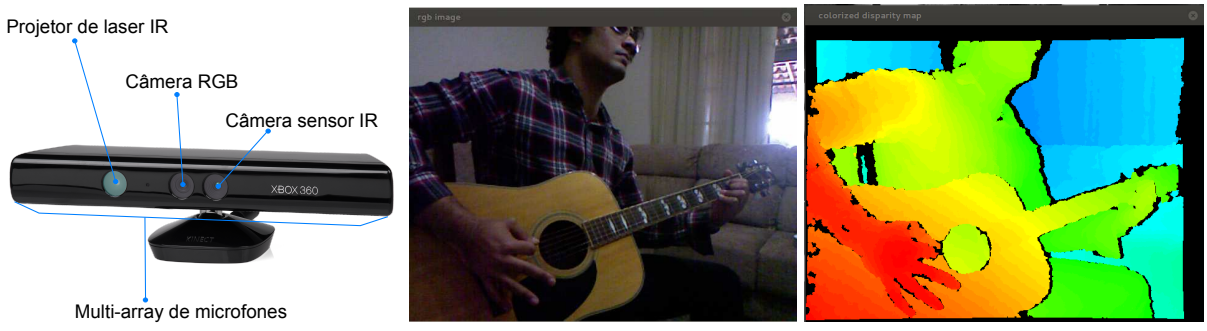
Neste trabalho, toda a aquisição de dados é realizada pelo sensor *Microsoft® Kinect* (Figura 2.3). Este produto foi desenvolvido com base em um sensor *3D* da empresa *PrimeSense* [56], e, em 2010, tornou-se um dos principais periféricos utilizados para o console de videogame *XBox 360©*. O propósito original do dispositivo é fornecer um fluxo rápido e constante de informações – cor, profundidade e áudio – por uma conexão *USB*, viabilizando, assim, uma interface de interação natural (*NUI*) do usuário com o videogame.

O *Kinect* provê dois canais de vídeo: (i) imagens *RGB* com resolução *VGA* 640x480 e taxa de quadros de 30 *FPS*; e (ii) imagens de profundidade com resolução *VGA* 640x480, 11 bits de codificação de profundidade por *pixel* e taxa de quadros de 30 *FPS*. A extração de imagens de profundidade é considerada um método ativo de aquisição por triangulação ativa, porém sem efetuar contato com objetos da cena, conforme a taxonomia apresentada.

Para realizar o processamento da imagem de profundidade, o sensor utiliza o laser infra-vermelho (inofensivo para o olho humano) projetando uma imagem feita por padrões de difração pseudoaleatórios estáticos (um holograma gerado computacionalmente) sobre a cena. Em seguida, um sensor *CMOS* de infra-vermelho (*IR*), alinhado horizontalmente com o projetor *IR*, reconhece a marcação de cada feixe de projeção no ambiente da cena e, utilizando triangulação ativa, o sistema é capaz de calcular a disparidade entre pontos da imagem capturada e a originalmente projetada. A reconstrução da localização de cada ponto *3D* em cena é dada por uma relação inversamente proporcional desta disparidade com os parâmetros da câmera [6].

Um resumo das especificações técnicas do sensor *Kinect* é apresentado abaixo [18, 30]:

- Câmera *RGB* de resolução *VGA* 640x480 *pixel*s, com *8-bits* por canal e filtro de cores Bayesiano. Opera a uma frequência de aquisição média de 30 *FPS*.



(a) *Hardware* do sensor *Kinect*. (b) Imagem RGB VGA 640x480. (c) Mapa de profundidade em gradiente de cores.

Figura 2.3: Composição e amostragem do sensor *Kinect*.

- O Sensor de profundidade, composto por projetor e câmera *IR*, permite a captura de imagens de profundidade em baixa ou alta intensidade de luz ambiente.
- O sensor monocromático utilizado no *stream* de vídeo é aplicado em resolução *VGA* 640x480 *pixel*s, com valor de 11-*bits* para profundidade em cada *pixel*, o que provê 2048 níveis de sensibilidade. Opera a uma frequência de aquisição média de 30 *FPS*.
- O alcance prático do sensor de profundidade é de $1{,}2$ até $3{,}5$ metros de distância, para uma área total de possíveis $6m^2$.
- Campo de visão angular em $57°$ horizontalmente e $43°$ verticalmente.
- Pivô motorizado é capaz de realizar movimentos verticais do sensor (*tilting*), passível de ajustes por software, em até $27°$ pra cima ou para baixo.
- A uma distância mínima teórica de 80cm, é possível cobrir um campo de visão de 87cm horizontalmente e de 63cm verticalmente, com uma resolução possível de $1{,}3$mm por *pixel* em distância.
- O conjunto de *softwares* da *API* de desenvolvimento *Microsoft® Kinect SDK* [57] é capaz de identificar até 6 usuários em uma cena, sendo que para 2 destes usuários rastreiam-se até 20 juntas do corpo de forma independente.
- O *array* de microfones é composto de 4 cápsulas de microfones e opera com cada canal processando áudio de *16-bit* em uma taxa de amostragem de 16kHz.

Uma limitação da tecnologia de aquisição do sensor *Kinect* é que este apenas obtém a imagem de profundidade sob a vista do sensor. Isto implica que apenas uma "casca" da malha de pontos *3D* dos objetos é obtida, e não seu modelo completo. Muitas vezes, no entanto, o modelo incompleto já contém informação suficiente para a aplicação de um algoritmo de visão. Outros trabalhos exploram o uso de mais de um *Kinect* e o registro de múltiplas vistas para a obtenção de modelos completos dos objetos em cena [58].

Além da *API* oficial, *Microsoft® Kinect SDK* [57] de desenvolvimento de sistemas para *PC*, a empresa *PrimeSense* disponibiliza uma versão compatível de *drivers* para sistemas Windows e Linux. Estes *drivers* garantem acesso físico à informação de cor, profundidade e áudio do sensor, e, juntamente com a *API* de código aberto *OpenNI* [59] e o *middleware* proprietário *NiTE* [56] fornecem a interface para que o programador as processe. De forma geral, tanto a *API Kinect SDK* quanto este conjunto de ferramentas possuem compatibilidade de funções. Porém, por apresentar um maior vínculo com o desenvolvimento de soluções de código aberto, os protótipos desenvolvidos neste trabalho utilizam o conjunto *PrimeSense*, *OpenNI* e *NiTE*.

As pesquisas com imagens providas pelo sensor *Kinect* são particularmente recentes, visto que até o momento não se encontravam, com facilidade de custo e configuração, equipamentos que viabilizassem a aquisição de imagens de profundidade para aplicações em tempo-real. Na próxima seção é apresentado um dos escopos de aplicação deste sensor: o reconhecimento da Língua de Sinais. Este problema constitui um grande desafio para estes sensores visto que o reconhecimento deve ser preciso ao distinguir um grande conjunto de sinais, e seu processamento deve ser prático suficiente, influenciando minimamente a taxa de aquisição dos quadros (manter a intuição de tempo-real).

2.3 Reconhecimento de Gestos no Contexto da Língua de Sinais

Um "gesto" é aqui definido como o processo pelo qual uma pessoa utiliza posição e articulação de partes do seu corpo (dedos, mãos, braços, cabeça, face) para expressar uma ideia e transmitir uma mensagem que é reconhecida por um receptor. Portanto, para que o gesto possua significado é preciso que algo ou alguém o conheça inicialmente (aprenda), para em seguida reconhecê-lo (identificação e classificação) em cena.

Esta seção aborda a contextualização do uso de gestos para a representação da língua de sinais. Neste escopo, são apresentadas as características das línguas de sinais, sua importância para a comunidade de surdos e deficientes auditivos, e alguns de seus desafios quanto ao aprendizado e interpretação por sistemas baseados em visão.

2.3.1 A Língua de Sinais

O reconhecimento de gestos por sistemas computacionais possui aplicações em variadas áreas como a língua de sinais, projeto de técnicas para identificação forense, monitoramento médico de pacientes, detector de mentiras, navegação e manipulação de ambientes virtuais [3]. Dentre estas áreas de aplicação, a língua de sinais é especialmente importante, não apenas por trazer um grande desafio para a interação humano-computador, mas, principalmente, por seu grande poder de expressão para a comunidade surda, portando também parte de sua cultura e legado.

É comum confundir os termos linguagem de sinais e língua de sinais, porém estes são termos semanticamente distintos. Enquanto se define linguagem de sinais como um complemento da comunicação oral (aumentando seu poder de expressão), a língua de sinais deve ser reconhecida como uma língua humana, obedecendo a um padrão de linguística próprio e com autonomia de criação, convenção e expressão. Além disso, a língua de sinais possui léxico, sintaxe, semântica e pragmática próprios. Não existe uma língua de sinais universal e, mesmo dentro de uma mesma língua de sinais, é possível encontrar diferentes dialetos, característica e influência da cultura de diferentes povos e regiões [60].

Em todo seu poder de expressão, os gestos de uma língua de sinais podem ser estáticos (com posição fixa sem considerar o tempo de encenação) ou dinâmicos (considerando-se o tempo de encenação). Ao reconhecer um gesto, é preciso ainda perceber a interação entre diferentes partes do corpo, e não apenas as mãos. Isto ocorre até mesmo para transmissão de mensagens que, tidas como simples de se expressar na comunicação oral, tornam-se complexas quando representadas em sinais (Figura 2.4).

Historicamente, embora muitas pesquisas tenham sido realizadas, poucos sistemas tecnológicos surgiram no intuito de auxílio da comunidade surda. Entre estes, o telefone de texto (*TDD*) (Figura 2.5) se tornou significativo ao longo do desenvolvimento das tecnologias de comunicação à distância. Este telefone especial possui um teclado portátil que permite a comunicação com um interlocutor que também possua o mesmo aparelho. A comunicação se dá por meio da leitura e escrita, ao invés da audição e fala. O aparelho consiste de um pequeno visor, que comporta apenas uma linha por transmissão, e dois receptáculos nos quais se inserem os bocais de um telefone comum (por onde normalmente se fala e escuta). O texto nestes aparelhos é escrito em Português, sem influência da língua de sinais [62].

(a) Mão direita em M, com a palma virada para a esquerda e com os dedos apontando para cima tocando o lado direito da testa;

(b) Braços cruzados em frente à cabeça, impondo as duas mãos fechadas e com as palmas para trás;

(c) Braços paralelos ao lado do corpo, mantendo as duas mãos fechadas e com as palmas para trás.

Figura 2.4: Posturas do gesto referente à expressão "derrame cerebral" em *Libras* [61].

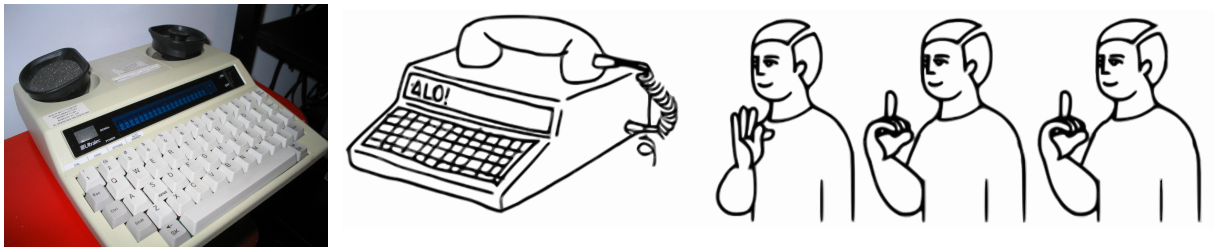


Figura 2.5: Telefone de texto (*TDD*). Adaptado de [34, 62].

Hoje, com o crescimento do mercado de *smartphones*, *tablets* e outros recursos digitais, a comunicação de surdos e deficientes auditivos se tornou bem mais ampla. Não há dúvidas dos efeitos positivos que o uso destes dispositivos tem produzido para a inserção e interação do surdo na sociedade. Acredita-se, no entanto, que esta colaboração (dispensando o uso de sinais) possa ter também seu impacto negativo, dado que pouco estimula o aprendizado e a aplicação da língua de sinais no dia-a-dia. Diante disto, há também que se ter o cuidado em preservar, ensinar e disseminar o uso da língua de sinais para a comunicação interpessoal. Esta é com certeza uma das principais motivações no desenvolvimento de sistemas computacionais para o reconhecimento da língua de sinais, como no caso deste trabalho.

A seguir são apresentadas, em particular, duas diferentes línguas de sinais utilizadas neste trabalho: a *American Sign Language* (*ASL*) e a Língua Brasileira de Sinais (*Libras*), ambas com origem comum na Língua de Sinais Francesa (*LSF*), mas estendidas e utilizadas em culturas bem distintas.

American Sign Language

Apesar de nunca se ter feito um censo da quantidade de falantes, a "Língua de Sinais Americana" (do Inglês, *American Sign Language – ASL*) é uma língua predominante entre a comunidade de surdos dos Estados Unidos e com uso internacional e intercontinental, majoritário entre os povos das Américas de origem anglo-saxônica [34]. Tomando como exemplo o seu correspondente oral (o Inglês), a *ASL* é uma língua franca tida como uma segunda língua para

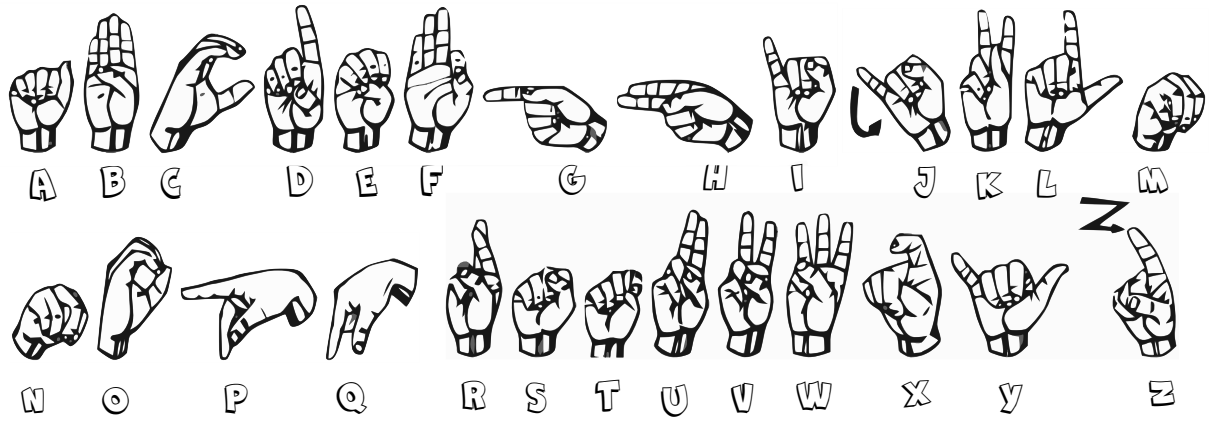


Figura 2.6: Alfabeto Manual da *ASL*. Adaptado de [34].

a maioria da comunidade surda pelo mundo. Um dos grandes fatores de sucesso da *ASL*, nos Estados Unidos, é o atendimento e apoio da sociedade desde o nascimento para as crianças que apresentem o prejuízo auditivo. Este apoio acontece nos hospitais, na preparação dos pais, nas escolas e por programas governamentais. Acredita-se que com um diagnóstico precoce e uma pronta tomada de ações, a cultura da língua de sinais é mais facilmente assimilada e integrada na sociedade.

O alfabeto manual é parte integrante de uma língua de sinais. Alguns dos gestos representados em um alfabeto manual lembram geralmente a forma de sua respectiva letra escrita na língua oral. O seu uso prático é associado ao doletrar (*i.e.*, soletração manual ou datilologia) de palavras para as quais algum dos interlocutores não as conheça ou não existam representações na língua utilizada. É usada, adicionalmente, para enfatizar, esclarecer ou aprender a língua de sinais. O usuário da língua doletra palavras em um ritmo próprio, utilizando sua mão dominante – apesar de que existem línguas, como a Língua de Sinais Britânica (*BSL*), que utilizam ambas as mãos na gesticulação do alfabeto manual.

A Figura 2.6 apresenta o alfabeto manual para a *ASL*, um conjunto de 26 posturas de mão representando as respectivas letras da língua oral inglesa. Nem todos os sinais do alfabeto da *ASL* são estáticos. Para os propósitos de reconhecimento deste trabalho, os gestos dinâmicos (letras 'J' e 'Z') são convertidos em posturas estáticas definidas pela captura da sua última posição na sequência do gesto.

Língua Brasileira de Sinais

A "Língua Brasileira de Sinais" (*Libras*) é a língua predominantemente usada entre os surdos nos centros urbanos brasileiros. A *Libras* teve sua oficialização postergada por muitos anos ao longo da história. Em 1857, foi fundada a primeira escola para surdos no Brasil, o atual Instituto Nacional da Educação de Surdos (*INES*). Naquele tempo, sob influência do congresso de Milão em 1880, preferiu-se adotar no Brasil a oralização (leitura labial e fala) dos surdos, quando já se desenvolvia e firmava uma língua própria de sinais [63, 64]. Somente no fim do século XX a comunidade se unia e se fortaleciam os movimentos pela oficialização da *Libras*. A luta obteve resultados com proposição de um projeto de lei para regulamentação do tema. No ano de 2002, a *Libras* foi oficialmente reconhecida e aceita como segunda língua oficial brasileira, através da Lei 10.436, de 24 de abril de 2002 [34, 65].

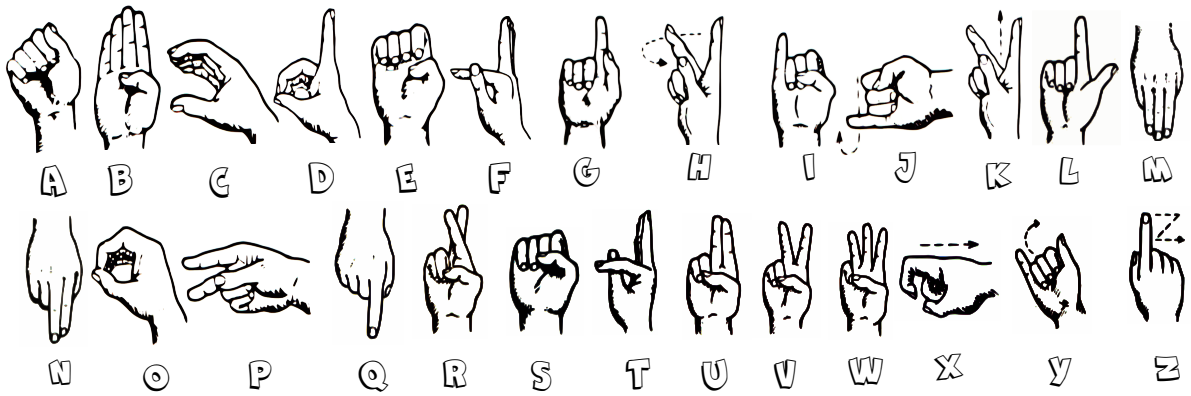


Figura 2.7: Alfabeto Manual da *Libras*. Adaptado de [66]

Em 2000, o censo demográfico contou $5{,}75$ milhões de pessoas que apresentam dificuldades auditivas ou surdez no Brasil, das quais 796.344 eram jovens com até 24 anos. No censo escolar de 2003, havia apenas 344 surdos cursando faculdades brasileiras [8].

A Figura 2.7 apresenta o alfabeto manual da *Libras* utilizado neste trabalho. O alfabeto possui 26 posturas de mão representando as respectivas letras da língua oral portuguesa. Nem todos os sinais do alfabeto da *Libras* são estáticos. Para os propósitos de reconhecimento deste trabalho, os gestos dinâmicos (letras 'H', 'J', 'X' e 'Z') são convertidos em posturas estáticas definidas pela captura da sua última posição na sequência do gesto.

2.3.2 Complexidade dos Sistemas de Reconhecimento Baseados em Visão

Uma abordagem conhecida e eficaz para o problema de reconhecimento da língua de sinais é utilizar marcadores ou dispositivos sensores anexados ao usuário [11, 12]. Normalmente, o uso destas tecnologias acessórias possibilita que objetos sejam rastreados eficientemente e com acurácia satisfatória na obtenção dos modelos cinemáticos. No entanto, seu uso pode também implicar em sistemas de hardware de elevado custo e configuração, apresentando fatores que dificultam a interação natural do usuário com o sistema.

Nesse caso, além de acurácia e eficiência, é desejável planejar um sistema de baixo custo, facilidade de configuração e capacidade pervasiva de uso. Assim, os sistemas baseados puramente em visão (sem o uso de marcadores) procuram tornar mais natural a interface com o usuário ao passo que buscam extrair os elementos mecânicos dos gestos apenas da informação visual da cena [4, 14].

Qualquer que seja a escolha de implementação para um sistema de reconhecimento de sinais baseado em visão, esta normalmente apresenta os seguintes componentes:

- **Extração de elementos em um quadro**: A Extração de elementos é bem sucedida quando se adquire e segmenta as informações essenciais para que o algoritmo de reconhecimento tome decisões:

  – A aquisição de imagens depende de características específicas do dispositivo sensor e é acompanhada de erro instrumental, dados incompletos, latência, frequência de amostragem, entre outros. Quando se utiliza imagens de profundidade é preciso considerar ainda a resolução a nível de profundidade e se adaptar o sistema de

posicionamento referencial (*3D*) para um espaço regular de métricas (a unidade de distância – tal como metros – não deve ser utilizada com coordenadas, em *pixel*s, do plano da imagem).

  - Após adquirir-se as imagens, deve-se segmentar a área de interesse ao reconhecimento. A segmentação de imagens é um procedimento usual em visão computacional e seu objetivo é separar objetos de interesse na cena, reduzindo a complexidade do problema e, ao mesmo tempo, contribuindo para a acurácia da solução [27]. O uso de imagens de profundidade têm se mostrado especialmente interessante no processo de segmentação, uma vez que permite identificar os objetos por sua localização tridimensional e possui menor dependência com fatores de intensidade luminosa do ambiente [6, 39, 40].

- **Base de dados**: Para se reconhecer, é preciso antes conhecer. O sistema de reconhecimento deve coletar previamente um conjunto suficiente de dados com amostras consistentes de cada gesto que se deseja reconhecer. A cada gesto é atribuída uma classe para qual o algoritmo deverá conhecer. A diversidade utilizada na coleta de amostra é algo desejável, mas um algoritmo robusto deve ser capaz de contornar a hipótese em que existam apenas amostras semelhantes ou em número reduzido. A extração de amostra para o banco pode ser feita com modelos sintéticos criados computacionalmente ou diretamente do dispositivo de captura [67].
- **Processo de Classificação**: O conceito de reconhecimento não se confunde com verificação. Reconhecer é uma tarefa mais complexa que verificar. Na verificação é fornecida uma amostra de teste juntamente com a identidade prévia para a qual o sistema deve responder se a amostra corresponde ou não. No reconhecimento não se tem a identidade prévia de uma amostra de teste, sendo necessário estimá-la [68]. Classificar é o procedimento necessário para se determinar a identidade de uma amostra segundo uma base de dados previamente definida.

O desempenho do sistema está atrelado a uma boa escolha de seus componentes. Devido à possibilidade de cada sistema utilizar um processo próprio de classificação, extração e formação da base de dados, é difícil estabelecer parâmetros que permitam a comparação direta entre diferentes soluções. Uma forma mais geral possível para se tentar avaliar e comparar o desempenho dos diferentes sistemas de classificação é considerar as métricas de acurácia e eficiência, definidas aqui nos seguintes termos:

Acurácia

Um dado algoritmo está correto se para cada instância do problema, a resposta produzida está correta [69]. A classificação é um mecanismo de se estimar a identidade de um objeto. Assim, a menos que se prove que um dado algoritmo de reconhecimento sempre obtém a identidade correta de uma amostra de teste qualquer, não há que se falar em correção algorítmica para os diferente métodos de classificação.

Existem dois resultados possíveis da instância de um problema de classificação:

- A amostra de teste foi identificada de forma correta. Isto é, o algoritmo de classificação realizou uma busca na base de dados e atribuiu a classe correta de um gesto à amostra de entrada.

Tabela 2.2: Descrição e comparação assintótica entre funções de custo de pior caso.

Notação	Descrição	Exemplos [(1)]
$\mathrm{O}(1)$	Constante	Acesso direto a um *pixel* de imagem.
$\mathrm{O}(\log n)$	Logarítmica	(Capítulo 3) Consulta em uma *K-D Tree*.
$\mathrm{O}(n)$	Linear	(Capítulo 4) Segmentação de pontos em uma imagem.
$\mathrm{O}(n \log n)$	Quase Linear	(Capítulo 3) Custo de uma iteração do algoritmo *ICP* proposto.
$\mathrm{O}(n^2)$	Quadrática	(Capítulo 5) Testes de "validação cruzada" sobre a base de dados.

[(1)] O conjunto de exemplos descritos na tabela possui valor informativo da complexidade de tempo de alguns procedimentos da metodologia implementada, encontrados ao longo do texto.

- A amostra de teste foi identificada, mas de forma incorreta. Ou seja, a atribuição da classe da amostra apresentou-se incompatível com sua verdadeira identidade.

A Acurácia é introduzida como uma forma de relativizar a correção dos algoritmos de reconhecimento. Define-se aqui acurácia como a medida de desempenho da fração de correção estimada para um algoritmo de classificação. Por sua vez, o valor de acurácia (ACR) de um sistema é computado como a razão entre a quantidade de atribuições corretamente identificadas ($acertos$) e a quantidade de experimentos realizados ($experimentos$):

$$\text{ACR } (\%) = \frac{acertos}{experimentos} \text{ x } 100 \tag{2.4}$$

Além de possibilitar uma métrica de comparação entre algoritmos, a acurácia é um valor de confiança no uso de sistemas. A fração definida é também um valor estimado, uma vez que é praticamente impossível simular todas as possíveis condições de amostras de entrada para o problema. Neste caso, ao medir a acurácia do sistema como um todo, é importante realizar uma quantidade suficiente de experimentos e procurar variar as condições de entrada em cada experimento.

Eficiência

Além de se avaliar aspectos de correção, um projetista deve se preocupar com a eficiência de um algoritmo. Esta preocupação justifica-se ao tentar predizer se um sistema projetado será capaz ou não de realizar uma dada tarefa no tempo adequado e com os recursos disponíveis.

A análise de complexidade de algoritmos tem entre seus objetivos prover mecanismos de avaliação do desempenho de um algoritmo quanto à sua complexidade de tempo e espaço [69]. Para isso, a teoria define um conjunto de notações matemáticas que avaliam o comportamento assintótico dos algoritmos em função de suas entradas. O uso de notações assintóticas permite comparar o desempenho de diferentes algoritmos que buscam a solução de um problema em comum, sem necessariamente considerar os recursos de máquina disponíveis.

Convencionaram-se várias métricas assintóticas cujo uso depende de como se deseja avaliar um algoritmo. A mais popularmente utilizada é a notação de limite assintótico superior O (*big O*), pois esta mede a complexidade de pior caso de um algoritmo. Assim, falar que um algoritmo possui complexidade de tempo em $\mathrm{O}(n^2)$ significa dizer que o número máximo de operações não passa de uma função quadrática da ordem de grandeza da entrada (Tabela 2.2).

Na implementação de algoritmos que utilizam imagens de profundidade, a complexidade de tempo toma como uma de suas variáveis de entrada (ordem de grandeza) o nível de resolução das imagens adquiridas, isto é, a quantidade de pontos *3D* obtidos da cena.

Quando se utiliza sistemas de reconhecimento em tempo-real, a complexidade assintótica não é tão prática, pois espera-se que um usuário interaja com o sistema ao mesmo tempo em que este provenha os resultados. Para estes sistemas, deve-se avaliar também o "tempo de processamento" de cada instância do problema de reconhecimento e medir a sua "frequência média de processamento de quadros" (*FPS*). Embora tenha pouca relevância teórica, estas outras métricas de comparação permitem estabelecer o quão prático o sistema se comporta quanto ao uso de interfaces de interação natural com o usuário (*NUI*).

O próprio conceito de tempo-real é algo discutível na literatura. Em geral, os sistemas interativos, baseados em visão, que operam em tempo-real devem reconhecer os gestos continuamente de modo que, ao resolver uma instância do problema, o usuário não perceba a latência para a resolução da instância seguinte [70]. Neste caso, um sistema que opere na frequência de 30 *FPS* (na mesma frequência de aquisição de imagens de profundidade do *Kinect*) deve processar 30 imagens a cada segundo, ou seja, deve realizar o reconhecimento de cada instância de imagem com tempo de processamento de ≈ 33ms, algo difícil mesmo com as máquinas atuais.

2.4 Discussão

Este capítulo apresentou os principais elementos constituintes de um sistema baseado em visão para o reconhecimento de gestos das línguas de sinais. A definição dos conceitos elencados constitui parte da visão geral do problema e escopo deste trabalho, resumidos a seguir.

2.4.1 Escopo da Dissertação

O reconhecimento da língua de sinais é considerado a categoria mais complexa no domínio de reconhecimento de gestos [10]. Isto porque é preciso lidar com um grande conjunto de posturas estáticas e dinâmicas, que podem ser muito similares em forma e incluir não só as mãos, mas também expressões da face, do tronco e dos braços. Devido à complexidade identificada, esta dissertação é limitada ao reconhecimento das posturas estáticas do alfabeto manual da *ASL* e da *Libras*, conforme as Figuras 2.6 e 2.7.

Para resolver o problema neste escopo, propõe-se o uso apenas de imagens de profundidade adquiridas por um sensor *Kinect*, utilizando-se da *API OpenNI* exclusivamente para aquisição dos dados brutos, sem qualquer outro tipo de processamento desta biblioteca. Em conformidade com os conceitos desenvolvidos, o sistema proposto deve adquirir e processar imagens de entrada do sensor e atribuí-las corretamente a uma das 26 possíveis classes de representação dos alfabetos manuais estudados.

O capítulo seguinte introduz uma visão técnica do problema estudado, avalia os principais métodos de classificação propostos na literatura, e introduz os conceitos necessários para a construção da proposta.

Capítulo 3

Revisão Teórica

Este capítulo aborda uma revisão do estado da arte em soluções para o reconhecimento da língua de sinais e introduz os fundamentos técnicos relevantes para a construção da proposta. A Seção 3.1 apresenta um levantamento de alguns dos principais trabalhos associados a algoritmos de classificação com aplicação no reconhecimento de padrões. A Seção 3.2 introduz pontualmente o algoritmo *Iterative Closest Point* (*ICP*), uma técnica utilizada no alinhamento de pontos *3D* aplicada na proposta para viabilizar a comparação direta de padrões em pares de imagens de profundidade.

3.1 Estratégias Gerais para o Reconhecimento de Posturas

Embora o reconhecimento da língua de sinais seja um tópico estudado em pesquisas de visão computacional [10, 17], a maioria dos trabalhos recentes não abordam uma solução completa para o problema quando se considera acurácia, eficiência e interface natural com o usuário (*NUI*). Neste sentido, o uso de imagens contendo informação de profundidade permite não só uma rápida segmentação dos objetos, como naturalmente é utilizada, mas também uma melhoria considerável quanto à acurácia do algoritmo de reconhecimento.

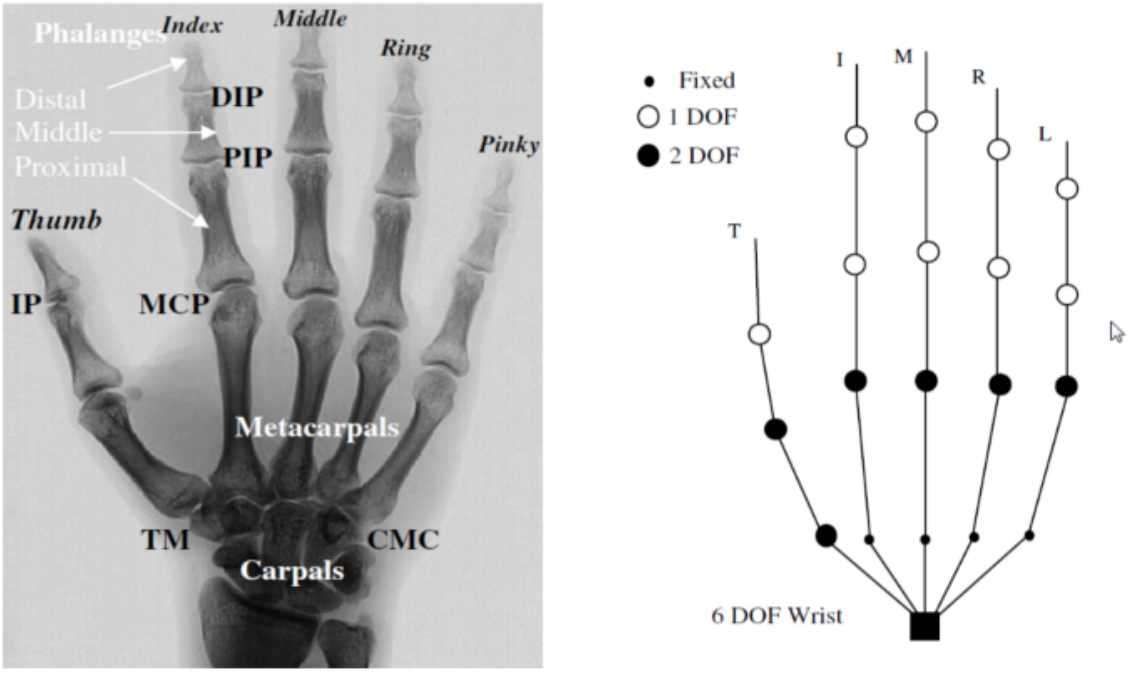
Esta seção descreve uma breve compilação de técnicas voltadas ao reconhecimento de padrões em imagens. Com vistas a uma comparação mais direta com a proposta desta dissertação, procura-se identificar as principais estratégias que aplicam dados de profundidade para o reconhecimento de posturas manuais e de gestos da língua de sinais.

3.1.1 Estrutura dos Gestos Manuais

O estudo do reconhecimento de posturas é uma área interdisciplinar que visa a análise do uso das mãos e outras partes do corpo para fins comunicativos. Na literatura, foram encontrados trabalhos para o reconhecimento automatizado de gestos sobre diferentes perspectivas.

Muitos trabalhos analisam o corpo humano completo a partir de sua decomposição em juntas estruturais (pontos-chaves de articulação); e viabilizam não só o reconhecimento de gestos com o corpo, mas também o rastreamento quadro-a-quadro destas juntas em contextos de tempo-real [30, 71, 72, 73, 74]. Estes trabalhos possuem aplicações diversas, como em jogos ou realidade virtual, porém permitem apenas o reconhecimento de gestos simples, uma vez que as juntas menores – como a de articulação dos dedos das mãos – não são plenamente estimadas ou reconhecidas.

De fato, o estudo do reconhecimento de gestos manuais da língua de sinais implica uma tarefa de análise mais complexa quando comparada a das macroestruturas do corpo humano [70]. Um modelo cinemático completo da mão (Figura 3.1) possui em geral 27 "graus de liberdade" (do Inglês, *degree of freedom – DOF*), interpretados a partir de uma estrutura hierárquica com 15 juntas móveis e 4 fixas. Esta formação permite criar um grande número de posturas e combinações que justificam o tratamento de subproblemas mais difíceis, como a ocorrência de ambiguidades no reconhecimento ou a oclusão de segmentos da imagem.



(a) Modelo anatômico

(b) Modelo cinemático

Figura 3.1: Modelo anatômico e cinemático de mãos para a estimação de posturas *3D* [70].

A dinâmica dos movimentos manuais possíveis permite construir uma estratégia inicial para o reconhecimento de sinais. Em [75], os autores obtêm um fluxo de imagens de profundidade que lhes permite analisar a morfologia, posição e orientação das mãos em diferentes gestos estáticos e dinâmicos da "Língua de Sinais Japonesa" (*JSL*). O treinamento é definido em função das restrições de posturas cinemáticas dos sinais; por exemplo, para um gesto de punho fechado, o ângulo entre dedos vizinhos deve ser mínimo. Embora seja possível definir um conjunto de restrições próximo ao ótimo em acurácia para conjuntos pequenos de gestos, ainda é uma tarefa difícil escalonar classificadores baseados em restrição dos movimentos em cenários maiores.

3.1.2 Aplicações para Informação de Profundidade

Os trabalhos relacionados neste capítulo utilizam os dados da imagem de profundidade em diferentes tarefas e atribuições como: na (i) "segmentação de objetos", para localização e extração de segmentos de interesse nas imagens; e na (ii) "construção de descritores", para obtenção de características derivadas da inferência de atributos nas imagens.

Segmentação de Objetos

No atual estado da arte, é unânime a aplicação de imagens de profundidade para tarefas relacionadas a segmentação de objetos. Dentre estas tarefas, é possível citar: (a) extração do fundo de cenas [6, 39, 75, 76]; (b) agrupamento (do Inglês, *clustering*) e marcação (do Inglês, *labeling*) de segmentos [30, 40]; e (c) localização de *pixel*s pertencentes às mãos em imagens [6, 15, 76, 77, 78].

De fato, a informação de profundidade simplifica o trabalho de se estimar a posição espacial dos objetos em cena. Além disso, os atuais sensores de profundidade, como o *Kinect*, permitem a aquisição de dados mesmo para ambientes com pouca iluminação. Isto mostra que as imagens de profundidade têm sido melhor aproveitadas quando comparadas ao uso isolado de imagens de intensidade na segmentação de gestos.

Em aplicações onde se espera que o usuário fique contra o campo de visão do sensor e mantenha as mãos à frente do seu corpo, é comum o emprego de uma estratégia simples de segmentação das mãos, denominada limiarização por profundidade (do Inglês, *depth thresholding*) [6, 15, 39, 75, 77, 78, 79]. O método de limiarização permite identificar as mãos como pontos contidos entre valores de limiar próximo e distante, calculados pela coordenada Z (profundidade) do centróide das mãos, que por sua vez pode ser pré-determinado para o usuário ou computado como o ponto mais próximo do sensor à cena. Nesta técnica, uma variação consiste em utilizar valores de limiar não apenas para a profundidade, mas para os outros eixos coordenados do espaço, definindo formas como uma caixa ou esfera para gesticulação. Com isto é possível eliminar grande parte dos ruídos e erros instrumentais do sensor na extração dos dados [17].

Sem o uso de dados de profundidade, a abordagem mais comum de segmentação para mãos se baseia em mapas de cor de pele [3, 37, 70]. Estas técnicas apresentam desempenho reduzido com variações da luz ambiente ou quando muitos objetos com cor de pele são observados em cena.

Utilizando imagens de intensidade e profundidade, Oikonomidis *et al.* [39] combina o método de limiarização com um algoritmo de detecção por cor de pele para atingir melhores resultados na etapa de segmentação. Nesta técnica, a maior área detectada como pele é tomada como a representação da mão, e é então dilatada de forma conservativa para torná-la evidente. O autor considera, por fim, toda a região previamente calculada que está sob um raio de distância fixo de $25cm$ (Figura 3.2). Um estudo com esta abordagem foi conduzido neste trabalho (Capítulo 6), porém sua aplicação não foi considerada para execução da metodologia proposta (Capítulo 4).

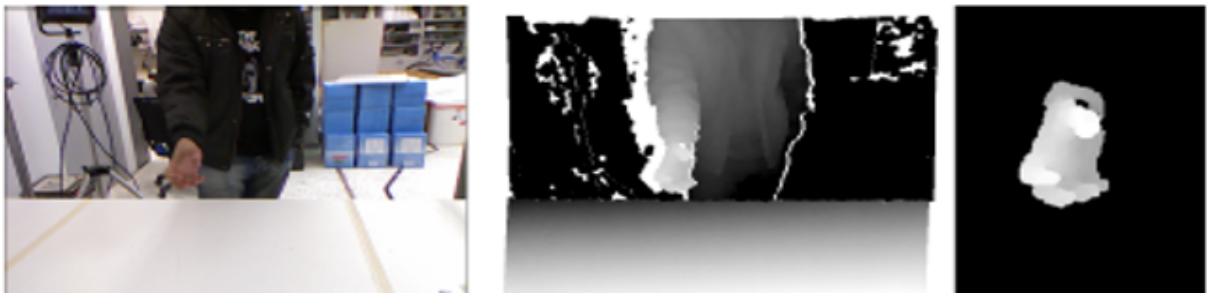


Figura 3.2: Segmentação de mãos a partir de imagens de profundidade e intensidade [39].

Construção de Descritores

A aquisição e segmentação são passos iniciais para a construção de classificadores. Neste sentido, um segundo aspecto determinante é a forma de tratamento dos dados adquiridos, isto é, como utilizar as imagens de profundidade para extrair elementos descritivos. Em geral, a extração de descritores deve ser feita de forma rápida para sistemas de interfaces naturais com o usuário (*NUI*). Um grande desafio neste caso é definir descritores que permitam aumentar a acurácia sem comprometer a eficiência do sistema.

Uma primeira abordagem é não utilizar nenhum processamento adicional para construção de descritores ou realizar apenas uma subamostragem do conjunto bruto de dados da imagem [23, 68, 80]. No entanto, isto não significa que poucas características são assimiladas pelo sistema; pelo contrário: devido à dificuldade, muitas vezes, de se extrair características significativas de forma explícita, dados brutos da imagem são tomados como entrada e as características são selecionadas implícita e automaticamente pelo classificador. Assim, sistemas que adotam esta escolha possuem uma rápida etapa de aquisição e pré-processamento, deixando o maior custo para a etapa de reconhecimento.

Na grande parte dos trabalhos [11, 76, 79, 81, 82], a extração de característica está associada a estimar o posicionamento de juntas ou pela Análise de Componentes Principais (do Inglês, *Principal Component Analysis – PCA*) das regiões hierárquicas, conforme o modelo da Figura 3.1. Nestes casos, a própria detecção e rastreamento destas características estruturais são tarefas não triviais exigindo consideráveis recursos e processamento de máquina [70].

Outros trabalhos sugerem a formação de descritores a partir de características que não dependam da estrutura anatômica das mãos, como o uso de filtros por funções matemáticas, silhuetas, bordas, cantos, sombras ou textura [15, 17, 77, 83].

3.1.3 Classificação por Treinamento e Aprendizagem

Em geral, problemas de classificação que exijam a interpretação de uma grande quantidade de dados ou combinações são tratados por algoritmos de aprendizagem. Esta abordagem inclui estratégias como:

- Modelos Ocultos de Markov (do Inglês, *Hidden Markov Models – HMM*) [6, 14, 84];
- Máquina de Vetores Suporte (do Inglês, *Support Vector Machines – SVM*) [76, 82, 85];
- Redes Neurais Artificiais (do Inglês, *Artificial Neural Networks – ANN*) [11, 76, 81, 85, 86];
- K-Vizinhos mais Próximos (do Inglês, *K-Nearest Neighbors – K-NN*) [15];
- Maximização da Margem de Vizinhança Média (do Inglês, *Average Neighborhood Margin Maximization – ANMM*) [79, 87].

Nestes algoritmos, um conjunto de amostra é utilizado para extrair características específicas (*features*) das imagens e aprender padrões ou tendências de classificação. De maneira geral, exige-se um extenso treinamento *offline* dos classificadores que, uma vez treinados, produzem estimações rápidas, mesmo em contextos voltados a tempo-real. Como desvantagem, a definição de características específicas para extração deve ser bem planejada, com o risco de que quando má realizada poder levar a uma baixa acurácia do sistema treinado. Além disso, a inclusão de

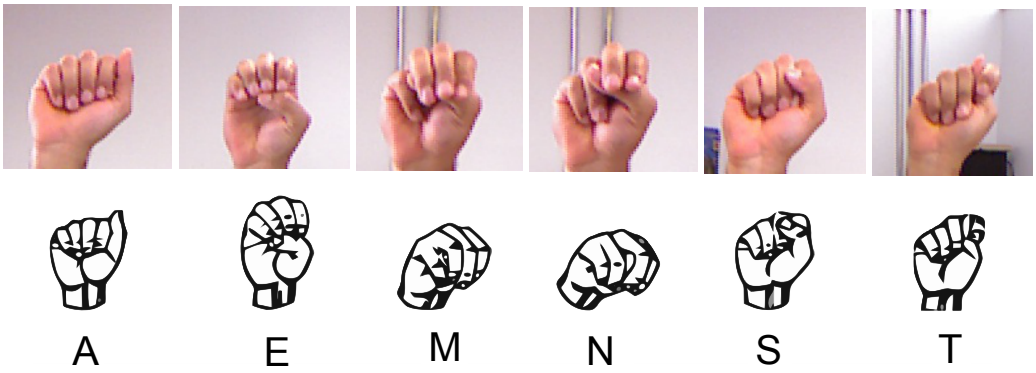


Figura 3.3: Ilustração de classes ambíguas no alfabeto da *ASL*. As letras são representadas por punhos fechados e se diferem apenas pela posição do polegar, levando à maiores níveis de confusão. Adaptado de [77].

novas classes ou a diversificação de gestos para um sistema de reconhecimento pode exigir um novo treinamento e a redefinição das características.

No escopo de aprendizagem para o reconhecimento de sinais, uma estratégia recente aplica o conceito de Floresta de Decisão Aleatória (do Inglês, *Random Decision Forest – RDF*) [76, 77]. As propostas relacionadas permitem classificar imagens de profundidade em tempo-real (30 *FPS* com o sensor *Kinect*) porém reportam apenas um valor de confiança para cada classe de gesto, o que nem sempre permite distinguir duas classes com um vetor de características próximo. Pugeault e Bowden [77] constroem um sistema iterativo que utiliza *RDF* na classificação semi-automática do alfabeto da *ASL*. A ideia é que letras reconhecidas de forma ambígua possam ser decididas pela interação do sistema com o usuário. Embora a acurácia média do sistema não tenha atingido um bom resultado (75%), os experimentos conduzidos apresentaram um conjunto interessante de classes de ambiguidade para o alfabeto manual da *ASL*, com destaque para os conjuntos de letras $\{A, E, M, N, S, T\}$ e $\{U, R\}$ (Figura 3.3).

Alguns trabalhos sugerem o estudo simultâneo de diferentes algoritmos de aprendizagem de forma a verificar qual melhor se adapta ao contexto de reconhecimento da língua de sinais:

- Em [85], os autores comparam o uso de *SVM*s e *ANN*s para o reconhecimento de posturas estáticas da *Libras* e concluem que ambas as técnicas são compatíveis em termos de acurácia ($94{,}69\%$ e $94{,}82\%$, respectivamente) e eficiência, porém o treinamento e seleção de *SVM*s foi mais simples do que lidar com possíveis mínimos locais e a longa fase de treinamento das *ANN*s. Os autores concluem que ao contrário das *ANN*s, *SVM*s parecem não sofrer com o aumento da dimensionalidade do problema.
- A autora em [15] propõe o uso da ferramenta Weka [88], uma coleção de algoritmos de aprendizagem de máquina que permite uma rápida adequação, comparação e flexibilidade dos métodos a diferentes problemas de mineração de dados (do Inglês, *data mining*). O texto compara o uso de variações das técnicas *K-NN* e *RDF*, implementadas pelo próprio Weka, para o reconhecimento do alfabeto manual da Língua Gestual Portuguesa. Observou-se que classificadores com *RDF* são mais rápidos que aqueles que utilizam *K-NN*, embora em termos de acurácia tenha se observado uma pequena vantagem para o uso da *K-NN* (96%) em relação a *RDF* (94%).

Neste mesmo trabalho, a autora propõe ainda um sistema de interação e datilologia onde se identifica o problema da coarticulação entre as letras, isto é, como definir o período de transição entre o reconhecimento de duas letras consecutivas ao soletrar uma palavra. Quando este problema era considerado, a acurácia do sistema caía consideravelmente para uma faixa entre 15% e 21%, independentemente do algoritmo de aprendizagem escolhido. O motivo justificado é de que a técnica proposta para o tratamento da coarticulação permitia a previsão consecutiva de letras duplicadas e o reconhecimento de falsos positivos durante a transição dos gestos.

3.1.4 Classificação por Casamento de Modelos

Conforme visto na seção anterior, a grande maioria dos classificadores propostos utilizam os métodos de treinamento e aprendizagem para atingir o reconhecimento de gestos na Língua de Sinais. No entanto, a estratégia de Casamento de Modelos (*Template Matching*) aparece com frequência, principalmente quando se aborda o reconhecimento de posturas estáticas [6, 23, 24, 78, 83].

Os algoritmos de classificação por Casamento de Modelos buscam a correspondência de uma amostra de teste com um modelo previamente adquirido. Características gerais destes algoritmos incluem:

- os dados de entrada podem ser imagens inteiras, parte de imagens ou a transformação de imagens;
- a classificação pode é feita a partir dos dados brutos ou de características explicitamente extraídas;
- Não existe uma etapa de treinamento;
- o par teste e modelo precisam ser normalizados sobre uma mesma perspectiva de comparação;
- é preciso estabelecer métricas para a comparação direta do par de teste e modelo.

Da necessidade em se ter uma perspectiva única de comparação entre as imagens da amostra de teste e do modelo, este trabalho utiliza o algoritmo *Iterative Closest Point* (*ICP*), um método para o alinhamento (registro) de duas superfícies tridimensionais correspondentes. Uma das hipóteses deste trabalho é que a aplicação do *ICP* não só permita o alinhamento das amostras de teste com modelos de casamento em um mesmo espaço, mas também estabeleça métricas comparativas que indiquem o nível de proximidade de um par alinhado.

Besl e McKay [22] propuseram originalmente o *ICP* como um procedimento para alinhar imagens de um mesmo corpo rígido tomadas por diferentes vistas a partir de uma câmera. Neste primeiro trabalho os autores também observaram que, apesar do *ICP* ser um lento processo de alinhamento baseado em iterações, o método poderia ser aplicado para verificar a congruência de duas formas geométricas distintas. No entanto, desde então, poucas pesquisas visando o reconhecimento de gestos têm investigado esta abordagem de congruência de formas, receando-se comprometer a eficiência dos sistemas.

Um conjunto mais notável de contribuições com o *ICP* pode ser emprestado de pesquisas que exploram aplicações biometricas. *Amor et al.* [80] constroem um clássico modelo *probe-and-gallery* que lhes permite reconhecer de forma satisfatória imagens de profundidade de faces

sob pontos de vista arbitrários utilizando o casamento de formas por *ICP*. Em [68], preocupados com a eficiência no contexto biométrico, os autores propõem uma indexação espacial por *voxel* aplicada durante a etapa de registro da galeria do banco de dados e utilizam-na, posteriormente, para computar, em tempo constante, os pares de pontos próximos em uma iteração do *ICP*. Ambos os trabalhos usam apenas a saída de valor quadrático médio da distância dos pontos para identificar a congruência entre as formas. Além disto, estudos do *ICP* visando reconhecimento biométrico frequentemente o aplicam à partes do corpo com maior inércia quando comparadas com os vários graus de liberdade (*DOF*) no contexto de gestos manuais, logo é mais simples alcançar valores altos para acurácia (Tabela 3.1).

Em um trabalho mais recente, *Trindade et al.* [23] desenvolveu um sistema para reconhecer o alfabeto manual da Língua de Sinais Portuguesa. Os autores conduziram experimentos com a aquisição de dados de profundidade e usaram-nos para o casamento de formas aplicando uma abordagem do algoritmo *ICP* simples, sem variantes. Afirma-se no trabalho que, por causa da pouca informação adquirida com o sensor *Kinect*, a implementação padrão proposta para o *ICP* não lhes permitiu corresponder sinais dentro de uma mesma classe. No entanto, não foram apresentados resultados concretos desta análise, nem realizada uma completa descrição dos experimentos.

3.1.5 Sumário Comparativo

A Tabela 3.1 apresenta os principais trabalhos citados neste capítulo. Ela traz um sumário das estratégias abordadas e o contexto próprio de reconhecimento ao qual foram aplicadas. O sumário também traz o valor de acurácia e eficiência reportado nos respectivos experimentos, quando disponíveis.

Como mencionado no Capítulo 2, é muito difícil obter valores de acurácia e eficiência que demonstrem uma real comparação entre técnicas de reconhecimento. Dessa forma, a ideia do sumário apresentado não é construir um ranqueamento para as melhores soluções, mas manter uma visão geral e atual do que se está utilizando hoje em trabalhos de reconhecimento e fornecer algum subsídio que permita a comparação com a proposta deste trabalho.

3.2 Alinhamento de Imagens de Profundidade por Iterações de Pontos Correspondentes

Imagens de profundidade podem ser adquiridas por diferentes perspectivas de acordo com a posição e campo de visão do dispositivo sensor. Isto dificulta o trabalho de reconhecimento, pois até mesmo imagens de uma mesma classe de postura apresentarão representações distintas, caso nenhum tipo de processamento adicional seja realizado.

O "alinhamento" ou "registro" pareado é o processo pelo qual duas imagens, adquiridas e representadas localmente (cada uma com seu próprio sistema referencial), são alinhadas sob um mesmo sistema de referência global. Neste sentido, o objetivo principal do alinhamento é recuperar o movimento rígido dado por uma transformação Euclidiana – com componentes de rotação e translação – que correlacione duas representações distintas de um mesmo objeto.

O alinhamento possui aplicações em diversos domínios e seu uso pode ser direcionado para imagens de intensidade ou profundidade, como por exemplo: construção de panoramas por ampliação de imagens de intensidade (*image stitching*) [27], navegação e robótica [89],

Tabela 3.1: Tabela comparativa relacionando os principais trabalhos levantados.

Referência	Aplicação	Estratégias	Acurácia[1]	Velocidade[2 3]	Profundidade
Almeida [6]	Alfabeto da *PSL*	Casamento de Modelos + Posição de Juntas	100.00%	29 *FPS*	✓
Amor *et al.* [80]	Biometria (Face)	Casamento de Modelos + *ICP*	97.25%	N/A	✓
Bowden *et al.* [14]	Palavras da *BSL*	*HMM* + Silhuetas e Orientação	97.67%	25 *FPS*	
Fujimura e Liu [75]	Palavras da *JSL*	Tabela de Restrições	N/A	N/A	✓
Keskin *et al.* [76]	Dígitos da *ASL*	*ANN* + Posição de Juntas	98.81%	30 *FPS*	✓
		SVM + Posição de Juntas	99.90%		
Lamar *et al.* [81]	Alfabeto da *JSL*	*ANN* + *PCA*	89.06%	N/A	
Liu e Fujimura [83]	Gestos Simples	Casamento de Modelos + Silhuetas	N/A	N/A	✓
Liwicki e Everingham [84]	Alfabeto da *BSL*	*HMM* + Orientação das Mãos	84.10%	N/A	
Pugeault e Bowden [77]	Alfabeto da *ASL*	*RDF* + Filtros de Gabor	75.00%	30 *FPS*	✓
Shotton *et al.* [30]	Juntas do Corpo	*RDF* + Rótulos e Orientação	73.10%	30 *FPS*	✓
Sousa [15]	Alfabeto da *PSL*	*K-NN* + Bordas	96.00%	N/A	✓
		RDF + Bordas	94.00%		
Souza *et al.* [85]	Alfabeto da *Libras*	*SVM*	94.69%	N/A	
		ANN	94.82%		
Trindade *et al.* [23]	Alfabeto da *ASL*	Casamento de Modelos + *ICP*	N/A	N/A	✓
Uebersax *et al.* [79]	Alfabeto da *ASL*	*ANMM* + Orientação das Mãos	89.60%	16 *FPS*	✓
Van den Bergh *et al.* [41]	Controle Robótico	*ANMM* + Haarlet	N/A	N/A	✓
Yan e Bowyer [68]	Biometria (Face)	Casamento de Modelos + *ICP*	94.10%	N/A	✓
	Biometria (Orelha)		97.30%		

(1) Valores da "acurácia" reportada para o reconhecimento não devem ser interpretados de forma absoluta, pois dependem da metodologia de realização dos testes utilizada em cada trabalho.

(2) Valores da "velocidade" reportada para o reconhecimento não devem ser interpretados de forma absoluta, pois dependem dos recursos de máquina disponíveis para realização dos testes em cada trabalho.

(3) Os valores da "velocidade" reportada foram retirados dos trabalhos com base em todo o sistema de reconhecimento e a partir da etapa mais dispendiosa. Por exemplo, considerando apenas a etapa de classificação em [30], o trabalho reporta velocidade de 200 *FPS* nos experimentos em uma máquina dedicada (*XBox 360*©), porém o valor inscrito na tabela para o sistema é de apenas 30 *FPS*, dominado pela velocidade de aquisição das imagens utilizando o *Kinect*.

reconstrução de modelos tridimensionais [90], diagnóstico médico por registro de imagens [91], entre outros.

No contexto de reconhecimento deste trabalho, o alinhamento é essencial uma vez que viabiliza a comparação direta entre imagens na arquitetura de Casamento de Modelos. A aplicação do alinhamento ao reconhecimento se difere de outros domínios devido a exigência de se encontrar mais regiões de sobreposição (*overlap*) entre a imagem do modelo e a imagem de avaliação. Esta seção descreve o algoritmo *Iterative Closest Point* (*ICP*), uma técnica de registro fino aplicado a pares de imagens de profundidade.

3.2.1 Visão Geral do Algoritmo *ICP*

Algoritmos de alinhamento refinado ou de registro fino são aplicados quando uma estimativa da transformação rígida é previamente conhecida. Esta estimativa é fornecida como entrada para um método iterativo, que a converge para uma solução mais precisa ao fim do processo.

Vários métodos de registro fino foram propostos: (i) *Iterative Closest Point* (*ICP*) e suas variantes [22, 25, 92, 93, 94, 95]; (ii) casamento utilizando "corpos de distâncias sinaladas" (*signed distance fields*) [96]; e (iii) "algoritmos genéticos" [97]. No entanto, os algoritmos baseados na técnica *ICP* são os mais aceitos e utilizados, principalmente por sua simplicidade conceitual, facilidade de implementação, eficiência e acurácia.

O objetivo do método é obter uma solução de qualidade para o movimento rígido entre duas coleções de pontos, minimizando iterativamente a função de custo das distâncias computadas entre pontos correspondentes selecionados sobre as duas superfícies (Figura 3.4).

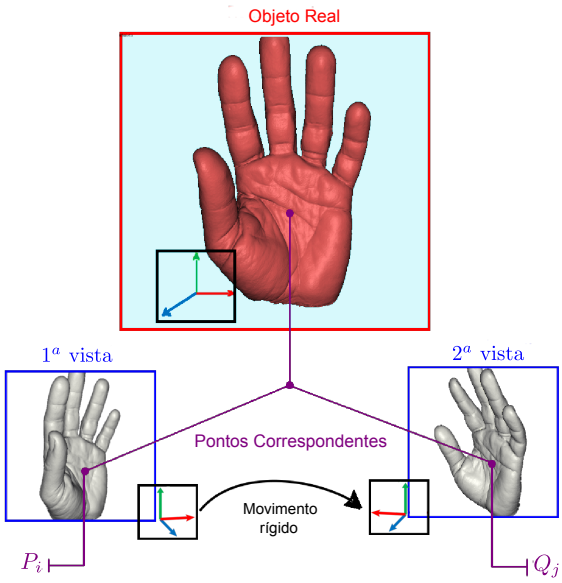


Figura 3.4: Ilustração da aquisição do movimento rígido a partir de pontos correspondentes.

Como uma técnica de alinhamento refinado, o algoritmo *ICP* supõe que uma boa estimativa inicial do movimento rígido seja provida. Dessa forma, é possível tanto aumentar a velocidade de convergência quanto atingir o mínimo global para a função de custo. Por outro lado, quando uma estimativa razoável não é provida, o método pode convergir incorretamente para um mínimo local e falhar na obtenção do alinhamento. As técnicas de alinhamento aproximado [38, 98, 99, 100] são frequentemente usadas para encontrar essa estimativa inicial. Na metodologia deste trabalho, uma estimativa inicial é tomada simplesmente a partir do vetor translação entre os centróides das duas superfícies.

Partindo-se de uma estimativa T – com componentes de rotação R e translação t – para a transformação rígida, em uma dada iteração, cada ponto na primeira imagem, $p_i \in P$, é levado a um ponto $T(p_i)$ referente ao mesmo sistema de coordenadas da segunda imagem. Dessa forma, ao longo de uma iteração o método deve buscar por um conjunto de correspondências na segunda imagem, $q_j \in Q$, que minimize a função de custo das distâncias entre $T(p_i)$ e q_j:

$$F = \frac{1}{L} \sum_{i=1}^{L} \left\| \mathbf{q_j} - [R\mathbf{p_i} + \mathbf{t}] \right\|^2 , \tag{3.1}$$

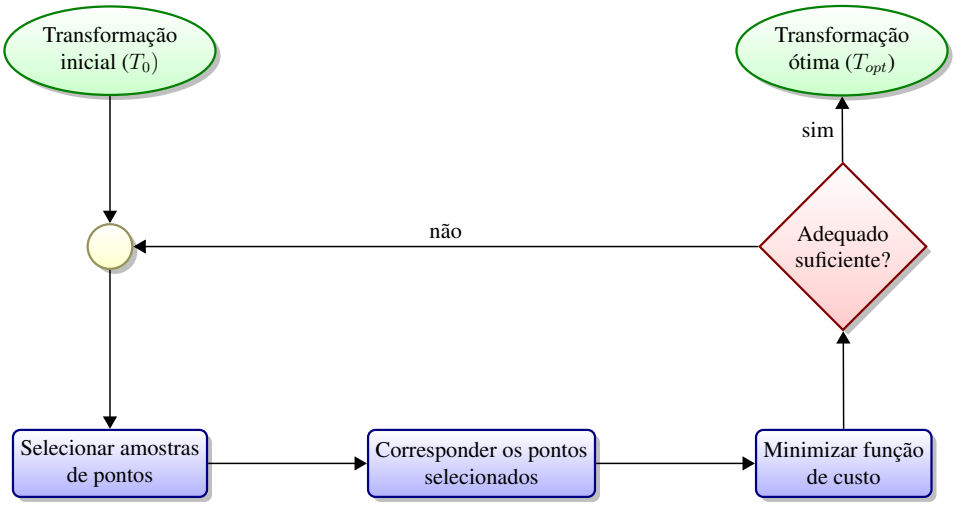


Figura 3.5: Fluxograma de etapas do alinhamento *ICP*.

com:

$$j = \underset{k,\, q_k \in Q}{\arg\min} \left\| \mathbf{q_k} - [R\mathbf{p_i} + \mathbf{t}] \right\|, \tag{3.2}$$

onde L é o número de correspondências selecionadas, p_i e q_j são pares de pontos correspondentes selecionados das imagens P e Q respectivamente, R é a matriz de rotação e $\mathbf{t}$ é o vetor translação, os três últimos obtidos pela estimativa da transformação rígida T, adquirida na última iteração.

O método como um todo se resume em iterações consecutivas das etapas apresentadas pelo fluxograma da Figura 3.5. A seguir, serão detalhados separadamente cada um destes procedimentos, esboçando-os pela formulação original do algoritmo, proposta em [22], e introduzindo as variantes utilizadas na implementação deste trabalho. O conjunto de variantes implementadas para o ICP segue do trabalho de Rusinkiewicz e Levoy [25], e visa atingir um alinhamento de rápida convergência, permitindo o seu uso no contexto de tempo-real.

3.2.2 Seleção de Amostras

Na Equação (3.1), o parâmetro L indica a quantidade de pontos que devem ser selecionados a cada iteração do algoritmo. Neste sentido, é evidente que o tempo de processamento do método de alinhamento possui uma relação estreita com esta quantidade de pontos. Em sua formulação original, o algoritmo *ICP* utiliza todos os pontos $p_i \in P$ disponíveis de uma das imagens do alinhamento. Dessa forma, com o aumento do nível de resolução das imagens, adotar a seleção de todos os pontos ao efetuar correspondências torna-se impraticável para a maior parte das aplicações.

Com o objetivo de reduzir o tempo de processamento em cada iteração do alinhamento, propõe-se realizar uma subamostragem dos pontos selecionados. Em [25], os autores garantem bons resultados para a transformação Euclidiana obtida com subamostragens de apenas 2% do total de pontos, indicando ser válida tal otimização.

Embora o uso de amostras reduzidas possa conduzir a uma menor acurácia do registro, é possível superar esta limitação por meio da seleção de pontos característicos em cada amostra. Do ponto de vista qualitativo, uma amostra pode ser composta de três formas [101]:

1. “Pontos uniformemente distribuídos”: A escolha de pontos da amostra é uma distribuição uniforme sobre a área de pontos da imagem. Esta escolha permite que a amostra contenha uma representação mais global do que a imagem representa.

2. “Pontos escolhidos aleatoriamente”: A escolha de pontos é feita aleatoriamente a partir do total de pontos. O procedimento aleatório permite que, eventualmente, se obtenha pontos específicos e característicos da imagem.

3. “Pontos de controle”: A escolha de pontos é feita com respeito a particularidades ou características específicas da imagem, como cantos, descontinuidades e pontos sobre arestas. Escolher pontos de controle diretamente não é uma tarefa fácil e exige algum tipo de processamento extra sobre a imagem, embora se obtenha melhores resultados.

Uma vez que há uma redução significativa do montante de pontos selecionados, a definição da estratégia de amostragem é essencial para garantir um registro correto. Visando uma boa acurácia para o algoritmo de alinhamento, o ideal é a escolha da amostra a partir de pontos de controle que estejam uniformemente distribuídos sobre a superfície da imagem.

Como variante utilizada neste trabalho, o algoritmo *ICP* aplica a subamostragem de pontos distribuídos uniformemente sobre o espaço de vetores normais disponíveis, descrita em [25]. Da combinação dos itens “1” e “3” enumerados acima, este método de seleção permite estabelecer a convergência em menos iterações e com menos tempo gasto em cada uma delas. Além disso, a definição de pontos de controle a partir das componentes normais permite um ganho significativo de robustez quando a imagem contém regiões de superfície aproximadamente uniforme mas com certas especificidades (Figura 3.6), como é o caso das imagens de posturas manuais.

As imagens de profundidade adquiridas pelo sensor *Kinect* não são estruturadas, isto é, a única informação disponível é a dos valores das coordenadas dos pontos *3D*. Para aplicar a seleção de amostras apresentada, é preciso estimar a componente normal de cada ponto da imagem adquirida. A normal $\mathbf{n_i}$ de um ponto $p_i \in P$ pode ser estimada por um problema de minimização da função de custo:

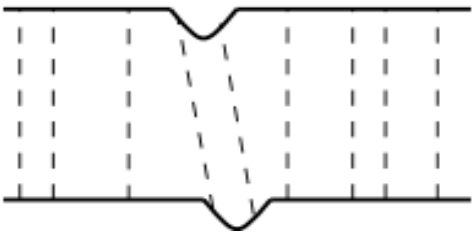
$$G = \sum_{v_j \in \eta} \frac{(\mathbf{n_i} \mathbf{v_j})^2}{\|\mathbf{n_i}\|^2}, \tag{3.3}$$

onde v_j são pontos selecionados sobre a vizinhança de pontos η, mais próxima de p_i. Neste caso, é possível minimizar a função G a partir da Análise das Componentes Principais (*PCA*), onde o cálculo da matriz de covariância é avaliada por:

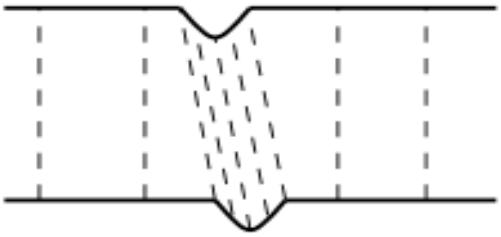
$$cov(p_i, \eta) = \sum_{v_j \in \eta} (\mathbf{v_j} - \mathbf{p_i})(\mathbf{v_j} - \mathbf{p_i})^T. \tag{3.4}$$

Assim, a componente normal n_i é computada pelo autovetor correspondente ao maior autovalor da matriz $cov(p_i, \eta)$. A qualidade da estimativa depende diretamente do número de pontos selecionados na vizinhança de p_i. Neste trabalho, uma vizinhança de oito pontos ($\|\eta\| = 8$) foi empiricamente considerada para estimação da componente normal.

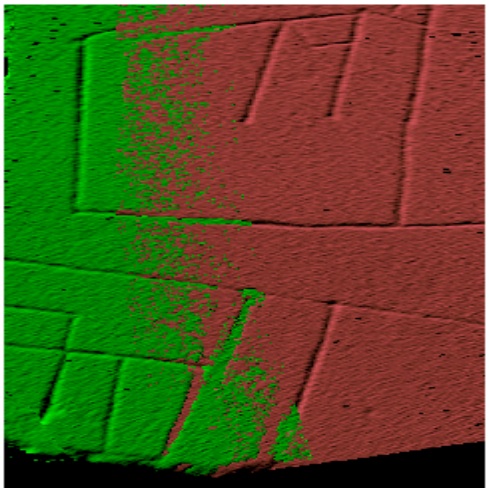
Por fim, ressalta-se que, conforme defendido em [102], a definição da amostragem de pontos retirados de ambas as imagens leva a uma maior acurácia da transformação final quando comparado a amostragem derivada de apenas uma das imagens. Esta foi também uma das estratégias adotadas na implementação do algoritmo *ICP* proposto.



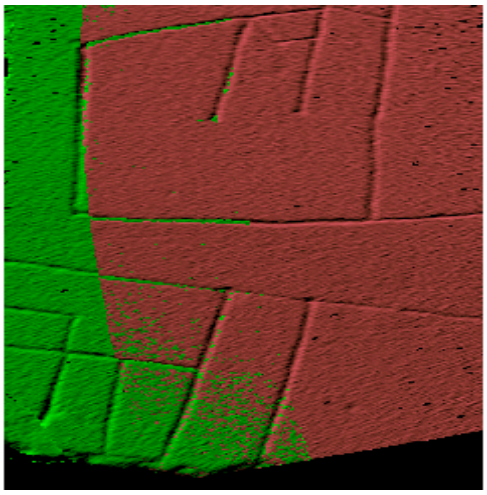
(a) Amostragem e correspondência de pontos de controle escolhidos aleatoriamente.



(b) Amostragem e correspondência de pontos escolhidos uniformemente sobre o espaço de possíveis normais à superfície.



(c) Alinhamento gerado com erros para pontos encontrados sobre *features* esparsos utilizando amostragem aleatória.



(d) A amostragem por distribuição uniforme da componente normal garante que pontos contendo *features* esparsos sejam escolhidos no registro.

Figura 3.6: Influência da escolha de amostragem sobre a acurácia do método ICP [25].

3.2.3 Correspondência entre Pontos Selecionados

O núcleo do algoritmo resume-se em encontrar pares de pontos correspondentes nas duas imagens e minimizar iterativamente o custo da soma das distâncias entre estes pares. Uma interpretação para a transformação rígida adquirida no processo de registro é de que esta estabelece um mapeamento entre os conjuntos de pontos alinhados. Ou seja, o ponto $p_i \in P$ está mapeado a um ponto p'_i, referente ao conjunto Q, dado por:

$$p'_i = T(\mathbf{p_i}) = R\mathbf{p_i} + \mathbf{t}. \tag{3.5}$$

Neste caso, a Equação (3.1) pode ser simplificada em:

$$F = \frac{1}{L}\sum_{i=1}^{L} \|\mathbf{q_j} - \mathbf{p'_i}\|^2, \tag{3.6}$$

e diz-se que $q_j \in Q$ é o par correspondente a p_i em P.

Problema do Vizinho mais Próximo

O problema do vizinho mais próximo [94] resume-se em encontrar para um dado ponto o seu vizinho ou conjunto de vizinhos mais próximo em uma coleção de pontos. A Equação (3.2) resume matematicamente a formulação do problema em termos da métrica de distância Euclidiana. O fator limitante desta formulação consiste em iterar sobre toda a coleção de pontos para se estimar um único vizinho mais próximo. Considerando o passo iterativo do registro *ICP*, é necessário estimar o vizinho mais próximo para todos os pontos da imagem sendo alinhada, o que levaria a uma complexidade quadrática de $O(|P||Q|)$.

Para reduzir a complexidade do problema do vizinho mais próximo, estruturas de dados de vizinhança mais próxima foram propostas. Estruturas, como as *K-D Tree*s [103], realizam um único pré-processamento sobre os pontos da coleção e permitem, em um momento posterior, responder a consultas dos k vizinhos mais próximos com relação a um ponto arbitrário dado. A vantagem destas estruturas é possuir um tempo de pré-processamento em $O(N)$ e um tempo de resposta médio para cada requisição em $O(\log N)$.

No caso de iterações do *ICP*, as *K-D Tree*s permitem realizar consultas para todas as correspondências possíveis entre as coleções P e Q com complexidade de tempo em $O(|P| \log |Q|)$, contribuindo para a eficiência do registro. Estas estruturas são úteis ainda para outras tarefas do algoritmo, como ao estimar a vizinhança η, dada pela Equação (3.3).

Rejeição de Pares Correspondentes

Muitos dos pares de correspondência tomados durante a execução do método *ICP* são *outliers*, isto é, são correspondências selecionadas erroneamente pelo método que não deveriam participar do processo de minimização. Um exemplo clássico são de pontos que existem sobre uma determinada vista do sensor de aquisição, mas não estão presentes em uma segunda vista correspondida.

A rejeição de pares de pontos correspondentes ocorre logo após se estabelecer a correspondência inicial e consiste em limitar a escolha dos pares a pontos que sejam *inliers* (correspondentes a uma mesma identidade do objeto e não meras associações). Uma primeira abordagem para rejeição consiste em limitar a distância máxima aferida entre correspondências.

Definição 1 (Distância Máxima de Rejeição – δ):
Sejam $P = \{p_i\}$ e $Q = \{q_i\}$, com $p_i, q_i \in \mathbb{R}^3$ dois conjuntos de pontos correspondentes mapeados por uma transformação rígida $T : \mathbb{R}^3 \to \mathbb{R}^3$.

Sejam ainda $D = \{d_i \,|d_i = \|T(\mathbf{p_i}) - \mathbf{q_i}\|\}$, o conjunto formado pelas distâncias entre as correspondências encontradas; $\overline{d}$, a média das distâncias em D; e σ_d, o desvio padrão das distâncias em D.

A "distância máxima de rejeição" (δ) é definida a cada iteração do método *ICP* como:

$$\delta = \overline{d} + 2{,}5\sigma_d \;, \tag{3.7}$$

Assim, um par de pontos correspondentes (p', q'), com $p' \in P$ e $q' \in Q$, será rejeitado quando:

$$\|T(\mathbf{p}') - \mathbf{q}'\| > \delta. \tag{3.8}$$

Embora a rejeição de pares não contribua significativamente para a velocidade de convergência do método, ela é uma propriedade essencial na obtenção de uma transformação rígida de qualidade. O valor de $2{,}5\sigma$ proposto na definição segue do resultado de experimentos conduzidos no trabalho em [104].

Neste trabalho, um segundo conjunto de restrições se impõe ao definir que pontos *inliers* devem apresentar compatibilidade quanto a componente normal à superfície. Sabe-se pelas definições da componente normal e de planos em $\mathbb{R}^3$ que o produto escalar do vetor normal pelo vetor diferença entre pontos contidos em um mesmo plano deve ser nulo. Assim define-se que um par de pontos correspondentes (p', q'), com $p' \in P$ e $q' \in Q$, será rejeitado se:

$$n_{q'} \cdot [T(\mathbf{p}') - \mathbf{q}'] > \varepsilon, \tag{3.9}$$

onde $n_{q'}$ é o vetor normal estimada para o ponto q', e ε é um valor limiar de implementação ($\varepsilon \approx 0$). O cálculo apresentado na Equação (3.9) é uma forma elegante de mostrar que os pontos p' e q' não possuem normais compatíveis.

3.2.4 Minimização da Função de Custo

Tendo-se um conjunto de pares de pontos correspondentes selecionados em duas imagens, o problema específico de estimar uma transformação rígida – rotação e translação – que leve pontos em um sistema de coordenadas O_1 a pontos correspondentes em um sistema de coordenadas O_2 é denominado "orientação absoluta" (do Inglês, *absolute orientation*) [105].

Diferentes fórmulas fechadas existem como solução a este problema [105, 106]. Estas soluções realizam operações de álgebra linear para minimizar a função de custo dada. A seguir é abreviada a solução por quatérnios unitários proposta em [105]:

Para minimizar a Equação (3.1), uma matriz simétrica $H(\Sigma_{pq})$, de dimensão 4×4, é construída:

$$H(\Sigma_{pq}) = \begin{bmatrix} tr(\Sigma_{pq}) & \Delta^T \\ \Delta & \Sigma_{pq} + \Sigma_{pq}^T - tr(\Sigma_{pq}) I_3 \end{bmatrix} \quad , \tag{3.10}$$

onde tr é a função "traço"; $\Delta = \begin{bmatrix} A_{23} A_{31} A_{12} \end{bmatrix}^T$ é computada a partir da matriz antissimétrica $A_{ij} = \Sigma_{p_i q_j} - \Sigma_{p_i q_j}^T$; Δ^T é a transposta de Δ; I_3 é a matriz identidade; e Σ_{pq} é a matriz de variância cruzada dos pontos p_i e q_i dada por:

$$\Sigma_{pq} = \frac{1}{N_p} \sum_{i=1}^{N_p} \begin{bmatrix} \mathbf{p}_i \mathbf{q}_i \end{bmatrix} - \mu_p \mu_q \quad , \tag{3.11}$$

com:

$$\mu_p = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{p}_i \quad , \tag{3.12}$$

$$\mu_q = \frac{1}{N_q} \sum_{i=1}^{N_q} \mathbf{q}_i \tag{3.13}$$

os centróides dos pontos p_i e q_i, respectivamente.

O autovetor unitário $\mathbf{h}_R = \begin{bmatrix} h_0 & h_1 & h_2 & h_3 \end{bmatrix}^T$ correspondente ao maior autovalor da matriz H é eleito como a nova rotação expressa em função de um quatérnio. A nova matriz de rotação R pode então ser recuperada, e o novo vetor translação t é calculado pelo vetor diferença entre os centróides, expressos em função da nova rotação:

$$R = \begin{bmatrix} h_0^2 + h_1^2 - h_2^2 - h_3^2 & 2(h_1h_2 - h_0h_3) & 2(h_1h_3 + h_0h_2) \\ 2(h_1h_2 + h_0h_3) & h_0^2 - h_1^2 + h_2^2 - h_3^2 & 2(h_2h_3 - h_0h_1) \\ 2(h_1h_3 - h_0h_2) & 2(h_2h_3 - h_0h_1) & h_0^2 - h_1^2 - h_2^2 + h_3^2 \end{bmatrix}, \tag{3.14}$$

$$t = \mu_q - R\mu_p. \tag{3.15}$$

O método executa várias iterações até a convergência para uma solução ótima, onde a função de custo é minimizada, isto é, está abaixo de um limiar, determinado experimentalmente.

Minimização Ponto-a-plano

Chen e Medioni [92] introduziram uma variante considerando a etapa de minimização do alinhamento *ICP*. Ao contrário da estratégia clássica de tomar distâncias "ponto-a-ponto", a proposta consiste em calcular a função de custo tomando distâncias entre pontos alinhados da primeira imagem a planos de tangência correspondentes da segunda imagem. Mais precisamente, o plano de tangência é calculado a partir do ponto de intersecção da projeção da componente normal do ponto alinhado da primeira imagem contra a superfície dada pela vizinhança mais próxima de pontos da segunda imagem.

Em geral, o algoritmo *ICP* "ponto-a-plano" converge em menos iterações que o método clássico [101]. Adicionalmente, como a distância é tomada em relação ao plano de tangência – e não em função de um ponto fixo – esta técnica minimiza o problema com *outliers*; e é robusta o suficiente contra grande parte das pertubações decorrentes de falta de informação das imagens de profundidade.

Como implementação utilizada, optou-se por realizar iterações com distâncias "ponto-a-ponto" somente nas 10 primeiras iterações do algoritmo; enquanto realiza-se iterações com distâncias "ponto-a-plano" para os restante das iterações no corpo do laço principal do procedimento. Conforme apontado em [25], esta abordagem híbrida permite estabilizar o processo de minimização (evitando mínimos locais), nas primeiras iterações, enquanto viabiliza uma rápida convergência da transformação, nas iterações seguintes.

3.3 Discussão

Este capítulo apresentou uma revisão teórica com as principais técnicas relevantes ao reconhecimento de posturas da língua de sinais. Inicialmente, levantou-se a complexidade do problema de representação de gestos manuais e como o uso de imagens de profundidade pode ser abordado em tal contexto. Em seguida, foram apresentadas as técnicas de classificação por treinamento e aprendizagem; mais populares no contexto de reconhecimento da língua de sinais. Foi introduzida também a abordagem por Casamento de Modelos; fundamental para o entendimento da proposta deste trabalho. Por fim, foi feito um comparativo geral entre as técnicas levantadas, servindo como referência compacta dos trabalhos similares mais atuais.

Da discussão levantada, nenhum dos trabalhos relacionados analisou sistematicamente o registro *ICP* como um possível procedimento para o reconhecimento de formas tridimensionais

das mãos. Além disso, poucos trabalhos abordam a eficiência do método em contexto de tempo-real, o qual exige otimizações quanto a simples implementação do algoritmo. Destes últimos, que propõem aprimoramentos para a eficiência do método, nenhum é diretamente aplicado ao reconhecimento da língua de sinais.

Como resumo da variante do algoritmo *ICP* descrita no capítulo, foram implementadas as seguintes características:

- iterações definidas pelo modelo selecionar-corresponder-minimizar;
- subamostragem de pontos distribuídos uniformemente sobre o espaço de vetores normais disponíveis;
- estrutura de dados de vizinhança mais próxima (*K-D Tree*);
- rejeição de pares correspondentes por distância máxima de rejeição e incompatibilidade de vetores normais;
- minimização da função de custo pela métrica de distância ponto-a-ponto nas primeiras iterações;
- minimização da função de custo pela métrica de distância ponto-a-plano no corpo do laço principal.

A complexidade geral da variante *ICP* com as otimizações propostas é da ordem de:

$$O(KL(\log N_P + \log N_Q)), \tag{3.16}$$

onde K é o número de iterações realizadas; L é o número de pares correspondentes amostrados em cada iteração; e N_p e N_q são os números de vértices das duas imagens alinhadas.

No próximo capítulo é apresentada a proposta do trabalho: o reconhecimento de posturas manuais do alfabeto da língua de sinais utilizando a estratégia de Casamento de Modelos junto ao alinhamento de pontos com o algoritmo *ICP* proposto.

Capítulo 4

Reconhecimento do Alfabeto Manual de Sinais

Neste capítulo descreve-se, em detalhes, o desenvolvimento da proposta do trabalho. Inicialmente, é apresentada a definição formal do problema abordado (Seção 4.1). Introduz-se, também, o esboço da solução de reconhecimento, com origem na aplicação conjunta do "algoritmo de registro" *ICP* e a estratégia de "Casamento de Modelos". Nas seções seguintes, são propostas melhorias a cada um destes elementos. Com respeito ao algoritmo *ICP* (Seção 4.2), a ideia é investigar parâmetros de entrada e possíveis métricas de saída que possam ser aplicados ou derivados do registro e permitam inferir similaridades entre um dado par de teste e modelo. Da parte do Casamento de Modelos (Seção 4.3), são propostos dois tipos de classificadores, que viabilizam o reconhecimento de letras dos alfabetos manuais, e permitem realizar um ajuste fino do equilíbrio entre a acurácia e a eficiência do sistema. Ao fim, as contribuições propostas são agregadas em uma metodologia única de implementação (Seção 4.4), definindo uma sequência lógica de estágios a serem executados pelo sistema de reconhecimento.

4.1 Formalização do Problema

O funcionamento básico para um sistema de reconhecimento do alfabeto manual de sinais pode ser decomposto no seguinte fluxo de ações:

- Uma imagem de profundidade P sem identificação conhecida, designada por "imagem de teste", é adquirida pelo sensor de captura.
- A imagem P é preparada para o reconhecimento a fim de que se tenha apenas um conjunto essencial de elementos, isto é, separa-se em P apenas a informação-chave das mãos que se espere corresponder à letra a ser identificada.
- O sistema deve possuir uma base de dados $\mathbb{Q}$, contendo amostras de imagens de modelo e de identificação conhecida, para cada classe do alfabeto manual analisado.
- Da imagem de teste P, é preciso extrair e comparar características que possam inferir similaridades contra um modelo qualquer Q, pertencente a base de dados $\mathbb{Q}$.
- O sistema deve prover um classificador, responsável por escolher a melhor classe (letra do alfabeto) de equivalência para a imagem de teste. Neste caso, mantendo-se fixa a imagem

P, deve-se aferir um conjunto de similaridades entre pares teste-modelo, aplicando-se, em seguida, uma técnica de classificação que rotule a melhor equivalência encontrada.

Tendo-se como exemplo de sistema o fluxo de ações elencado, o problema de reconhecimento abordado neste trabalho é definido formalmente como:

Definição 2 Reconhecimento de posturas estáticas do alfabeto da língua de sinais: Seja Σ um conjunto finito de classes (letras) de um alfabeto manual para o qual uma base prévia de conhecimento, $\mathbb{Q}$, tenha sido construída.

Dada uma imagem de teste P qualquer, o problema de "reconhecimento da postura estática do alfabeto da língua de sinais" consiste em identificar corretamente a letra $\mathfrak{C} \in \Sigma$ correspondente à P, entre as $|\Sigma|$ possíveis classes de equivalência amostradas por $\mathbb{Q}$.

4.1.1 Proposta Integrada para o Reconhecimento

Uma estratégia simples, porém eficaz, para o problema apontado consiste em aplicar uma arquitetura baseada em Casamento de Modelos. Intuitivamente, esta arquitetura foi apresentada na introdução desta seção, como o exemplo de fluxo de ações para um sistema de reconhecimento. Particularmente, dois passos de interesse são evidenciados: (i) "casamento" da imagem de teste com um conjunto de modelos representativos; e (ii) "comparação" pareada destas imagens para estimar o grau de proximidade em cada par. Assim, modelos de referência que provejam melhores similaridades para uma determinada "métrica de correspondência" são utilizados para identificar a classe à qual o dado de teste pertence. Embora não seja o melhor processo para o reconhecimento rápido de imagens, esta arquitetura permite uma completa e detalhada análise dos seus estágios [17].

Deste escopo, um procedimento natural para inferir similaridades consiste em comparar diretamente o par de imagens teste-modelo, desde de que estas estejam registradas sobre um mesmo sistema de coordenadas. O *Iterative Closest Point* [22] (*ICP*) é um algoritmo que realiza tal alinhamento e, se aprimorado, permite estabelecer correspondências entre o par verificado. Trindade *et al.* [23] tentaram aplicar o registro *ICP* neste mesmo contexto, porém descartaram-no alegando que a técnica não era adequada para recuperar boas métricas de similaridade durante a classificação.

Este trabalho propõe uma investigação conjunta do algoritmo *ICP* à estratégia de Casamento de Modelos para solucionar o problema de reconhecimento de posturas estáticas do alfabeto manual.

4.2 Aprimoramentos do Registro *ICP* para o Casamento de Formas

Uma observação quanto ao reconhecimento de padrões por Casamento de Modelos é que deve existir um mecanismo que estabeleça correspondências entre as imagens avaliadas. O processo de alinhamento surge, assim, como um passo natural, uma vez que se espera que as imagens estejam sob um mesmo sistema de referência para comparação. No entanto, apenas o alinhamento referencial não possui a função precípua de aferir similaridades, sendo preciso inferí-las ou estimá-las.

Uma primeira contribuição deste trabalho consiste no aprimoramento do alinhamento *ICP*. Especialmente à descrição do algoritmo proposto no capítulo anterior, incluem-se como melhorias:

(i) o gerenciamento dos parâmetros de entrada (Subseção 4.2.1);

(ii) a inferência de similaridades computadas ao longo de uma única instância de execução do registro (Subseção 4.2.2).

A hipótese é que estas propriedades contribuam diretamente para o desempenho da acurácia ou da eficiência do casamento de padrões com o algoritmo *ICP*.

4.2.1 Parâmetros de Instância

Como forma de alterar o comportamento de uma instância de alinhamento, são propostas variáveis de entrada que permitem: (a) restringir a duração e complexidade das iterações (elementos iterativos); ou (b) modificar a qualidade da transformação linear obtida pelo registro (modificadores da transformação obtida).

Elementos Iterativos

O custo computacional do algoritmo *ICP* é um dos problemas principais de sua aplicação no contexto de sistemas em tempo-real. Este fato convalida-se pela análise da complexidade do algoritmo, apresentada na Equação (3.16). Neste caso, quando algum dos parâmetros, K ou L, assume valores grandes, o tempo de execução das instâncias de registro pode se tornar o principal gargalo destes sistemas. Existem, portanto, dois tipos de elementos iterativos na proposta de implementação que claramente influenciam a eficiência do algoritmo:

1. **Número máximo de iterações (K)**:

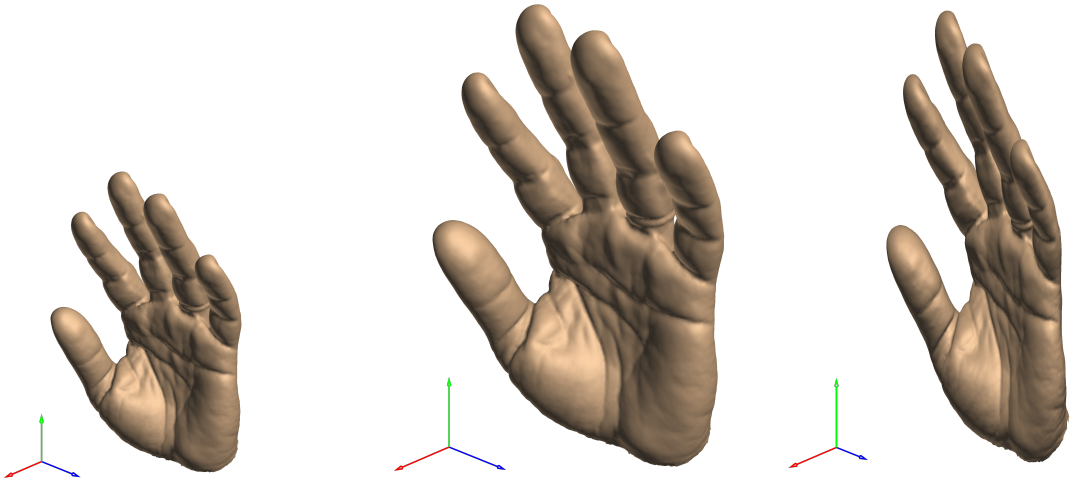
   Em geral, o número de iterações aplicadas em uma instância do alinhamento é variável, e condicionado pela convergência da função de custo ao mínimo local mais próximo.

   Adicionalmente, o alinhamento de imagens que não representam um mesmo gesto tende a consumir mais iterações durante a otimização, constituindo assim um pior caso do algoritmo (formas não equivalentes são impróprias ao alinhamento). Esta é uma condição sensível, principalmente, quando aplicada ao Casamento de Modelos, pois nesta estratégia espera-se que somente uma classe de imagens de modelos seja compatível com a imagem de teste. Ou seja, comparações com classes não equivalentes – o que ocorre na maior parte da estratégia – são possíveis geradores de pior caso do alinhamento.

   Sob tais circunstâncias, a interrupção forçada do número de iterações do registro (parâmetro K da complexidade do algoritmo) é justificável para a proposta do trabalho, contribuindo para a velocidade final do reconhecimento.

2. **Número máximo de pontos selecionados (L)**:

   Em um algoritmo *ICP* básico não há preocupação de se ter um passo de seleção de amostras, conforme descrito no fluxograma de registro (Figura 3.5). Visto de outro modo, pode-se dizer que o passo de seleção de amostras, para uma implementação fiel do algoritmo [22], consiste na atribuição cega de todos os pontos possíveis das imagens sendo alinhadas.



(a) Modelo original (b) Escala uniforme ($s = 1{,}5$) (c) Escala não–uniforme ($sx = 0{,}8$; $sy = 1{,}5$; $sz = 1{,}3$)

Figura 4.1: Ilustração da aplicação de diferentes parâmetros de escalonamento a um mesmo modelo *3D*.

A restrição do número de pontos selecionados para correspondência permite reduzir a quantidade de operações algébricas necessárias. Tal restrição define, portanto, uma subamostragem do problema de orientação absoluta original em cada iteração.

A metodologia implementada inclui um parâmetro que permite definir o número máximo de pontos escolhidos (parâmetro L da complexidade do algoritmo), reduzindo o tamanho do conjunto de dados tratado em cada iteração.

***Resultados Esperados*:** De ambos os elementos iterativos, espera-se que a redução de seus valores apresente um ganho quantitativo na velocidade do reconhecimento sem afetar, entretanto, o potencial de acurácia do casamento com aplicação do algoritmo *ICP*.

Modificadores da Transformação Obtida

A implementação básica do algoritmo *ICP* permite recuperar apenas a transformação rígida com parâmetros de rotação e translação entre as duas formas geométricas alinhadas. Neste contexto, vale lembrar que uma transformação rígida preserva a distância entre pontos e, como tal, o tamanho e a forma das imagens [90]. Entretanto, sistemas de reconhecimento devem ser capazes também de aferir similaridades entre objetos de distintas formas e tamanhos.

Diante desses requisitos, propõe-se o uso adicional de modificadores ao procedimento *ICP*, responsáveis por computar parâmetros aproximados de escalonamento entre as formas durante o passo de minimização [105]. Três tipos de transformações são investigadas (Figura 4.1):

- **Movimento rígido sem recuperação de escala entre as imagens**: Forma usual apresentada de obtenção da transformação Euclidiana.
- **Movimento com escala uniforme dos eixos das componentes principais**: Analisa-se o eixo das componentes principais do conjunto de pontos da imagem alinhada e permite-se escaloná-los, igualmente, caso isto contribua para minimizar a função de custo.

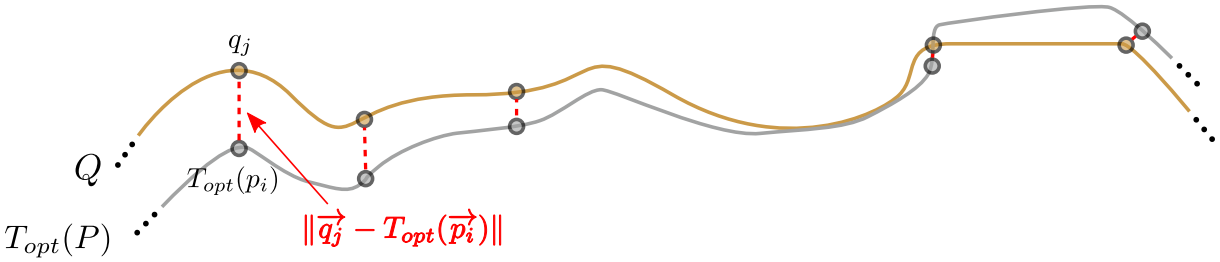


Figura 4.2: Ilustração do cálculo das distâncias ponto-a-ponto (em vermelho) entre pontos correspondentes de duas curvas, P e Q, alinhadas por uma transformação T_{opt}.

- **Movimento com escala não-uniforme dos eixos das componentes principais**: Analisa-se o eixo das componentes principais do conjunto de pontos da imagem alinhada e permite-se escaloná-los, de modo desigual, caso isto contribua para minimizar a função de custo.

***Resultados Esperados*:** Espera-se que a aplicação das variantes de escala sobre o eixo das componentes principais conduza a uma pequena melhora de acurácia, uma vez que seu uso pode viabilizar a comparação de diferentes formas e tamanhos de mãos das imagens adquiridas.

4.2.2 Métricas para Inferência de Similaridades

A definição de métricas de correspondência surge da necessidade de se comparar diferentes alinhamentos da imagem de teste com os modelos da base de dados. Para o reconhecimento da língua de sinais, a escolha de um tipo de métrica de comparação irá definir se a classe da amostra poderá ou não ser corretamente identificada.

A primeira métrica de erro ponto-a-ponto é a mais conhecida e é normalmente implementada na maioria dos trabalhos para se avaliar a correspondência relativa a dois conjuntos de pontos. Em sequência, a métrica de erro ponto-a-plano é também comum entre os trabalhos com o algoritmo *ICP* mas, sendo raramente associada a tarefa de reconhecimento, é aplicada mais como função de custo para convergência do registro.

Valor Quadrático Médio para Distância Ponto-a-ponto

Esta métrica tem sido empregada em um número de trabalhos relacionados [23, 25, 68, 80]. O cálculo da raiz da média quadrática (do Inglês, *Root Mean Square Error* – (*RMS error*), para distância ponto-a-ponto é realizado após a última iteração do alinhamento e consiste do valor de raíz quadrada da média aritmética das distâncias quadráticas entre pontos correspondentes (Figura 4.2).

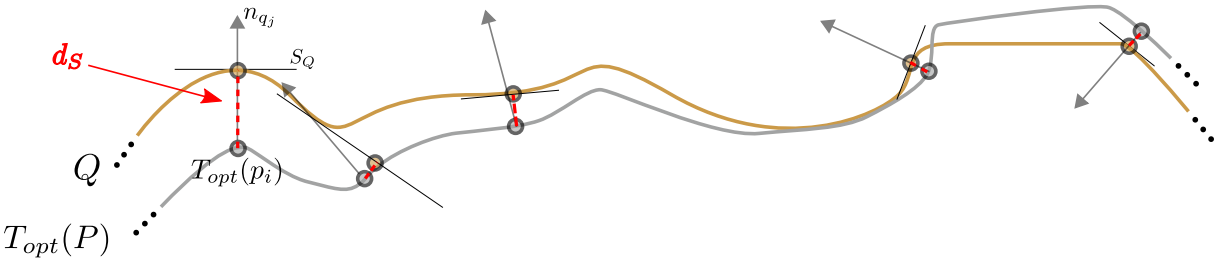


Figura 4.3: Ilustração do cálculo das distâncias ponto-a-plano (em vermelho) entre pontos e planos correspondentes de duas curvas, P e Q, alinhadas por uma transformação T_{opt}.

Definição 3 (Erro *RMS* ponto-a-ponto – M_1): Sejam $P = \{p_i\}$ e $Q = \{q_j\}$, com $p_i, q_i \in \mathbb{R}^3$, dois conjuntos de pontos correspondentes mapeados por uma transformação rígida $T_{opt} : \mathbb{R}^3 \to \mathbb{R}^3$, adquirida no último passo de minimização do algoritmo *ICP*.

Da Equação (3.1), tem-se:

$$\boldsymbol{M}_1 = \sqrt{F} = \sqrt{\frac{1}{L}\sum_{i=1}^{L} \|\mathbf{q_j} - T_{opt}(\mathbf{p_i})\|^2}, \tag{4.1}$$

onde $j = \arg\min_{k,\, q_k \in Q} \|\mathbf{q_k} - [R\mathbf{p_i} + \mathbf{t}]\|$ e L é o número de correspondências selecionadas ao fim do registro.

A métrica $\boldsymbol{M}_1$ é, portanto, uma métrica de mínimo, de forma que quanto menor o erro *RMS* ponto-a-ponto computado, melhor qualificada estará a classe de equivalência do modelo em relação à imagem de teste.

Valor Quadrático Médio para Distância Ponto-a-plano

O uso da função de custo baseado em distâncias ponto-a-plano é indicado para alinhar formas predominantemente planas que possuam algumas irregularidades em suas superfícies [92]. Além disto, esta técnica permite uma redução do número de passos de minimização, possibilitando ganhos no tempo de execução do alinhamento. Poucos trabalhos, no entanto, aplicaram esta métrica para avaliar a similaridade entre duas formas. O erro *RMS* para distâncias ponto-a-plano é computado após a última iteração do alinhamento e consiste da raíz quadrada da média aritmética das distâncias quadráticas entre pontos e superfícies correspondentes (Figura 4.3).

Definição 4 (Erro *RMS* ponto-a-plano – M_2): Sejam $P = \{p_i\}$ e $Q = \{q_j\}$, com $p_i, q_i \in \mathbb{R}^3$, dois conjuntos de pontos correspondentes mapeados por uma transformação rígida $T_{opt} : \mathbb{R}^3 \to \mathbb{R}^3$, adquirida no último passo de minimização do algoritmo *ICP*.

O "erro *RMS* ponto-a-plano" é calculado por:

$$\boldsymbol{M}_2 = \sqrt{\frac{1}{L}\sum_{i=1}^{L} d_s^2(T_{opt}(p_i), S_Q)}, \tag{4.2}$$

onde $S_Q = \{s|\mathbf{n_{q_j}} \cdot [\mathbf{q_j} - \mathbf{s}] = 0\}$ é o plano de tangência a imagem Q; $\mathbf{n_{q_j}}$ é a normal a superfície de Q por q_j; $q_j = T_{opt}(l_i) \cap Q$ é o ponto de interseção da reta $T_{opt}(l_i)$ a superfície de Q; $l_i = \{a|(\mathbf{p_i} - \mathbf{a}) \times \mathbf{n_{p_i}} = 0\}$ é a reta normal a superfície P por p_i; e d_s é a distância sinalada de um ponto a um plano.

A métrica $\boldsymbol{M}_2$ é, portanto, uma métrica de mínimo, de forma que, quanto menor o erro *RMS* ponto-a-plano computado, melhor qualificada estará a classe de equivalência do modelo em relação à imagem de teste.

As métricas de valores quadráticos médios atendem primariamente ao objetivo de minimização da função de custo do alinhamento de formas. Assim, o uso destes parâmetros como inferência de similaridades pode não ser o mais apropriado para o reconhecimento de posturas estáticas da língua de sinais. Este trabalho propõe, como contribuição, um segundo conjunto de métricas para avaliar correspondências: a (i) "norma da matriz de transformação residual", que visa avaliar a convergência da transformação; e o (ii) "limiar de distância máxima", empregado como medida da distância máxima de proximidade entre as formas.

Norma da Matriz de Transformação Residual

A norma da matriz de transformação utilizada neste trabalho é relativa a norma de Frobenius, que é uma extensão natural da norma de vetores lineares aplicada a matrizes. A norma relativa a transformação residual de alinhamento é computada após a última iteração e permite quantificar o quanto este último passo de minimização caminhou na direção do mínimo local.

Definição 5 (Norma Residual do Registro – $\boldsymbol{M}_3$): Sejam $T^{(K)} = T_{opt}$, $T^{(K-1)}$, $T_{\text{res}} \in \mathbb{R}_{4\times4}$, matrizes de transformação linear, tais que $T^{(K)}$ e $T^{(K-1)}$ foram obtidas, respectivamente, na última e penúltima iteração do algoritmo, e $T^{(K)} = T_{\text{res}}(T^{(K-1)})$.

Neste caso, a "norma residual do registro" é calculada por:

$$\boldsymbol{M}_3 = \sqrt{\sum_{i=1}^{4}\sum_{j=1}^{4} |(T_{\text{res}} - I_4)_{ij}|^2}, \tag{4.3}$$

onde $I_4 \in \mathbb{R}_{4\times4}$ é a matriz identidade.

A métrica $\boldsymbol{M}_3$ é, portanto, uma métrica de mínimo, de forma que, quanto menor a norma residual do registro computado, mais estáveis foram os últimos passos de convergência, e melhor qualificada estará a classe de equivalência do modelo em relação à imagem de teste.

Limiar de Distância Máxima

Na implementação *ICP* proposta, o limiar de distância máxima viabiliza a triagem de pares de pontos selecionados para correspondência, e seu valor é reajustado durante o alinhamento, conforme a Equação (3.7). Supondo que, a cada iteração, os dois conjuntos estejam cada vez

mais próximos, o valor de distância máxima de rejeição é reduzida progressivamente de acordo com a média das distâncias computadas.

Definição 6 (Limiar Dist-Max – M_4): O valor da métrica "limiar dist-max" é calculado por:

$$M_4 = \delta_{opt}, \tag{4.4}$$

onde δ_{opt} é a "distância máxima de rejeição" (Definição 1) obtida no último passo de minimização do algoritmo *ICP*.

Em termos práticos, esta é também uma métrica de mínimo, ou seja, espera-se que quanto menor a estimativa final do limiar de distância máxima, melhor qualificada estará a classe de equivalência do modelo em relação a imagem de teste.

Análise da Região de Sobreposição

A métrica M_3 apresenta uma avaliação da convergência da transformação obtida, enquanto que M_4 retoma uma medida de proximidade em distância entre as formas. Em ambos os casos, no entanto, estimam-se apenas características locais do movimento rígido, podendo tornar o uso destas métricas instáveis para modelos parcialmente completos.

Um conjunto de métricas mais robustas é proposto com base em características estruturais das amostras. Para estas métricas, é avaliado o "maior subconjunto comum entre pontos" (do Inglês, *Largest Common Pointset – LCP*), expresso na seguinte definição (Figura 4.4):

Definição 7 (Maior Subconjunto Comum entre Pontos e pontuação *LCP*): Sejam $P = \{p_i\}$ e $Q = \{q_j\}$, com $p_i, q_i \in \mathbb{R}^3$, dois conjuntos de pontos correspondentes mapeados pela "transformação rígida" $T_{opt} : \mathbb{R}^3 \to \mathbb{R}^3$, adquirida no último passo de minimização do algoritmo *ICP*. Seja, ainda, δ_{opt} a "distância máxima de rejeição" obtida sob a mesma condição.

O "maior subconjunto comum entre pontos" (*LCP*$_{\mathrm{P}}$), relacionado à T_{opt} sobre congruência δ_{opt} com respeito à P, é definido pelo subconjunto $P_{\max} \subseteq P$ de maior cardinalidade possível, tal que $\forall p_i \in P_{\max}, \quad \|T_{opt}\,(\mathbf{p_i}) - \mathbf{q_i}\| \leq \delta_{opt}$.

Neste caso, define-se a *pontuação LCP*$_{\mathrm{P}}$ (γ_{P}) como a fração entre a cardinalidade do maior subconjunto comum entre pontos e o total de pontos em P, isto é:

$$\gamma_P = \frac{|P_{\max}|}{|P|}. \tag{4.5}$$

Da mesma forma, definem-se *LCP*$_{\mathrm{Q}}$ e γ_{Q}, relacionados à transformação $(T_{opt})^{-1}$ (a inversa de T_{opt}) sobre congruência δ_{opt}, com respeito à Q.

A *pontuação LCP*$_{\mathrm{P}}$ (γ_P) relativa a T_{opt} é a fração de pontos em P correspondente a região de sobreposição (*overlap*) de seu alinhamento em Q. Neste caso, é fato que: quando há uma fração considerável para a região de sobreposição entre as imagens, a transformação T_{opt} apresenta necessariamente uma elevada *pontuação LCP*. Por outro lado, um valor alto de γ_P significa que a maioria dos pontos em $T_{opt}\,(P)$ foi correspondida em Q por uma distância de erro de no

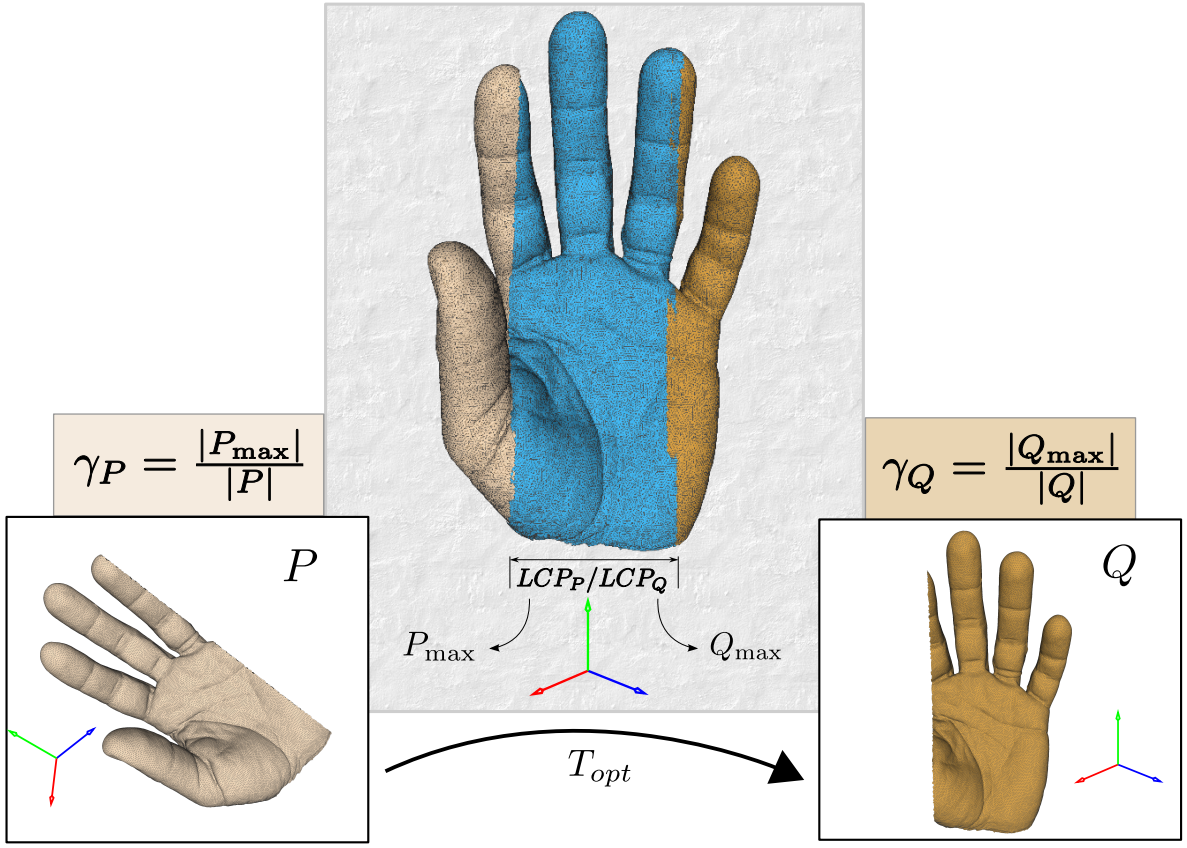


Figura 4.4: Ilustração do cálculo do Maior Subconjunto Comum entre Pontos (em azul), LCP_P e LCP_Q, e da Pontuação *LCP*, γ_P e γ_Q, para dois modelos, P e Q, alinhados por uma transformação T_{opt}.

máximo δ_{opt}. Se a maioria dos pontos satisfaz o critério de limiar de distância, a transformação T_{opt} obtida será também robusta para as imagens extraídas, mesmo em moderada presença de ruídos ou *outliers*, visto que estes normalmente não interferem em toda a estrutura das formas.

Com base nos conceitos abordados, são propostas três métricas: (a) "pontuação *LCP* mínima"; (b) "pontuação *LCP* máxima"; e (c) "pontuação *LCP* média".

Definição 8 (Pontuação *LCP* Mínima – M_5): Da Definição 7, a métrica de "pontuação *LCP* mínima" é calculada por:

$$M_5 = \min(\gamma_P, \gamma_Q). \tag{4.6}$$

A pontuação *LCP* mínima indica a menor fração possível de estimativa da região de sobreposição (visão pessimista da correlação entre imagens). No entanto, M_5 é uma métrica de máximo, de forma que, quanto maior for a pior suposição com relação a área de sobreposição, melhor

qualificada estará a classe de equivalência do modelo em relação a imagem de teste.

Definição 9 (Pontuação *LCP* Máxima – M_6): Da Definição 7, a métrica de "pontuação *LCP* máxima" é calculada por:

$$M_6 = \max(\gamma_{\mathrm{P}}, \gamma_{\mathrm{Q}}). \tag{4.7}$$

A pontuação *LCP* máxima indica a maior fração possível de estimativa da região de sobreposição (visão otimista da correlação entre imagens). Assim, M_6 é uma métrica de máximo, de forma que, quanto maior for a melhor suposição com relação a área de sobreposição, melhor qualificada estará a classe de equivalência do modelo em relação a imagem de teste.

Definição 10 (Pontuação *LCP* Média – M_7): Das Definições 8 e 9, a métrica de "pontuação *LCP* média" é calculada por:

$$M_7 = \frac{M_5 + M_6}{2}. \tag{4.8}$$

A métrica M_7 é o valor médio entre as *pontuações LCP* mínima e máxima. Esta também é uma métrica de máximo e espera-se, a partir dos conceito de valor médio, que um parâmetro de similaridade mais representativo seja obtido.

***Resultados Esperados*:** As métricas utilizadas na inferência de similaridades são mecanismos de avaliar a equivalência entre duas imagens *3D*. Dos valores extraídos para as métricas apresentadas, espera-se que um sistema de reconhecimento baseado no alinhamento *ICP* possa identificar corretamente todo o conjunto de posturas estáticas dos alfabetos manuais da *ASL* e da *Libras*.

4.3 Classificadores Baseados em Casamento de Modelos

Um aspecto de interesse no projeto de sistemas baseados na estratégia de Casamento de Modelos é a análise da quantidade de comparações teste-modelo necessárias para realizar a classificação. Métodos de força bruta, que procuram por toda a base de modelos, são muitas vezes proibitivos para cenários de reconhecimento em tempo-real. Neste sentido, deseja-se saber o quanto é possível melhorar a eficiência da estratégia sem comprometer a sua acurácia.

Utilizando o registro *ICP* em conjunto ao Casamento de Modelos, a metodologia geral consiste da aplicação de M instâncias (proporcional ao número de modelos) do algortimo de alinhamento. Da Equação (3.16), a complexidade em tempo total deste processo é de:

$$O(MKL(\log N_P + \log N_Q)). \tag{4.9}$$

Em um cenário básico de reconhecimento do alfabeto manual, uma imagem de teste P possui valores de similaridades computados contra um subconjunto de modelos Q_i, com $i = 1, \ldots, M$,

Algoritmo 1 Melhor-Ajuste.

Entrada: P: Instância de imagem de teste;
$\mathbb{Q}$: Base de modelos;
$\boldsymbol{M}$: Métrica de correspondência.
Saída: $\mathfrak{C}$: Classe reconhecida.

1. $\mathfrak{C} \leftarrow \emptyset$, MelhorAjuste $\leftarrow 0$
2. **para toda** (imagem Q $\in \mathbb{Q}$) **faça**
3. $v \leftarrow$ AVALIARMÉTRICAICP(P, Q, $\boldsymbol{M}$)
4. **se** ($\mathfrak{C} = \emptyset$) **ou** (MelhorAjuste $\otimes\, v$)[1] **então**
5. $\mathfrak{C} \leftarrow$ Q.classe, MelhorAjuste $\leftarrow v$
6. **fim se**
7. **fim para**
8. **retorne** $\mathfrak{C}$;

uniformemente distribuídos pelo número de classes avaliadas. Assim, o classificador deve predizer a qual letra (classe) do alfabeto a imagem de teste deverá ser rotulada. Para realizar esta tarefa, uma técnica de classificação deverá encontrar uma interpretação consistente para as métricas extraídas.

Este trabalho propõe, como contribuições, a adequação de um algoritmo de força bruta ao algoritmo *ICP* (*Melhor-Ajuste*) (Seção 4.3.1), e a técnica de *Ajuste Aproximado por K-Baldes* (Seção 4.3.2) para o estágio de classificação. Enquanto o Melhor-Ajuste é indicado para atestar a acurácia do casamento de padrões *3D*, o Ajuste Aproximado por K-Baldes possui um forte viés para melhoria da eficiência da metodologia implementada.

4.3.1 Melhor-Ajuste

Uma primeira abordagem de classificação considera, para cada instância de reconhecimento, o conjunto completo de modelos inscritos na base de dados e executa o alinhamento *ICP* entre todos os pares teste-modelo possíveis. O Algoritmo 1 apresenta o esboço de implementação da estratégia.

A técnica descrita poderia ser relacionada a uma forma de classificação 1-NN [107], onde uma amostra de teste é reconhecida pela classe do modelo vizinho com vetor característico mais próximo. No entanto, no contexto do alinhamento *ICP*, não existe propriamente um vetor característico para uma imagem por si só, uma vez que as similaridades são computadas aos pares. Logo, não se pode utilizar o contexto de vizinho mais próximo e é necessária a busca completa pela similaridade mais próxima. Para propósitos de identificação, esta primeira estratégia de classificação será denominada técnica de "Melhor-Ajuste".

4.3.2 Ajuste Aproximado por K-Baldes

Uma das desvantagens óbvias da técnica de *Melhor-Ajuste* é o seu lento processo de comparação de uma imagem de teste contra toda a base modelos. Isto é, anteriormente ao registro *ICP*,

[1] O operador $\otimes$ utilizado no Algoritmo 1 e no Algoritmo 2 deve ser substituído adequadamente pelo operador relacional '$<$' quando a métrica $\boldsymbol{M}$ for uma métrica de mínimo, e pelo operador relacional '$>$' quando for uma métrica de máximo.

Algoritmo 2 Ajuste Aproximado por K-Baldes.

Entrada: P: Instância de imagem de teste;
$\mathbb{Q}$: Base de modelos;
$\boldsymbol{M}$: Métrica de correspondência;
K: Tamanho de balde por classe.
Saída: $\mathfrak{C}$: Classe reconhecida.

1. $\mathbb{U} \leftarrow \emptyset$ *{Armazena amostras aleatórias das* 26 *classes}*
2. $B[26] \leftarrow \{\emptyset\}$ *{Listas de baldes com* K *valores de similaridades por classe}*
3. $V[26] \leftarrow \{0\}$ *{Valor médio de similaridade por classe}*
4. $\mathbb{U} \leftarrow$ GERARSUBCONJUNTO($\mathbb{Q}, K$)
5. **para toda** (imagem U $\in \mathbb{U}$) **faça**
6. B[U.classe] $\leftarrow B$[U.classe] $\cup$ AVALIARMÉTRICAICP(P, U, $\boldsymbol{M}$)
7. **fim para**
8. $\mathfrak{C} \leftarrow \emptyset$, MelhorMédia $\leftarrow 0$
9. **para toda** (classe $c \in$ {A-Z}) **faça**
10. $V[c] \leftarrow$ CALCULARVALORMÉDIO($B[c]$);
11. **se** ($\mathfrak{C} = \emptyset$) **ou** (MelhorMédia $\otimes\, V[c]$)[1] **então**
12. $\mathfrak{C} \leftarrow c$, MelhorMédia $\leftarrow V[c]$
13. **fim se**
14. **fim para**
15. **retorne** $\mathfrak{C}$;

não há valores de métricas computadas entre a instância de entrada com os modelos armazenados. Com vistas a uma maior eficiência da metodologia implementada, este trabalho propõe a técnica de *Ajuste Aproximado por K-Baldes* (*KB-Ajuste*).

Um esboço do procedimento é listado no Algoritmo 2. Nesta técnica, um subconjunto da base de modelos é selecionado aleatoriamente com distribuição uniforme, escolhendo-se K amostras de cada classe de reconhecimento (linha 4). Em seguida, são executadas $\|U\|$ instâncias do alinhamento *ICP* (linhas 5-7), com o objetivo de recuperar valores de similaridade para um tipo de métrica escolhida previamente. O último cálculo consiste em obter o valor médio de similaridade para cada classe (linhas 8-14). O processo de reconhecimento termina, na linha 15, onde a classe com o melhor valor médio é encontrada.

Como a técnica *KB-Ajuste* depende de um procedimento aleatório e estatístico, um conjunto específico de experimentos foi conduzido no Capítulo 5 para análisar sua acurácia e eficiência.

4.4 Metodologia Implementada

O paradigma aplicado como metodologia deste trabalho consiste da estratégia de Casamento de Modelos onde a classificação é realizada comparando-se um dado caso de teste contra um conjunto de modelos de imagens de profundidade. O diagrama da metodologia implementada é apresentado na Figura 4.5. O processo de reconhecimento é dividido em três estágios, cada um deles com seus próprios requisitos e provendo os respectivos elementos de saída a ser utilizados nos estágios subsequentes.

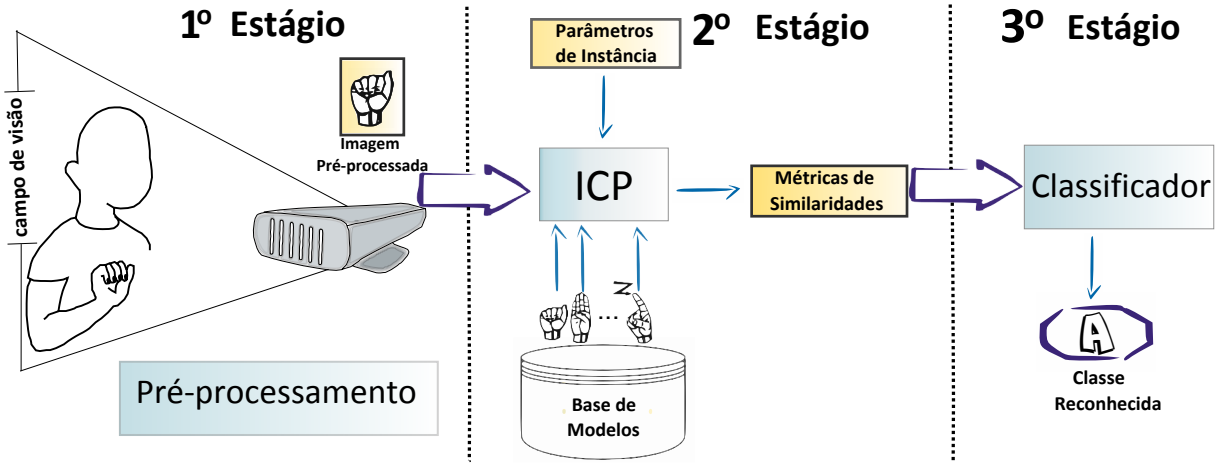


Figura 4.5: Diagrama da metodologia de reconhecimento implementada.

Uma sequência inicial de passos de pré-processamento é executada para preparar as imagens de profundidades adquiridas pelo dispositivo sensor (1º Estágio). Em seguida, o procedimento *ICP* alinha a imagem de teste pré-processada contra um subconjunto de modelos da base de dados, escolhidos uniformemente entre as classes de posturas do alfabeto manual (2º Estágio). Então, os subprodutos do alinhamento *ICP* são utilizados como métricas de avaliação para o esquema de classificação proposto (3º Estágio). As subseções seguintes descrevem cada um destes estágios, em detalhes.

4.4.1 Passos de Pré-processamento

A aquisição de imagens na metodologia implementada é sempre realizada com o auxílio de: (i) um sensor *Kinect* padrão [5], (ii) a *SDK OpenNI* [59], e o (iii) *middleware NiTE* [56]. O sensor se mantém em uma posição fixa a meia-altura do corpo e apenas o fluxo de imagens de profundidade é utilizado para aplicar as técnicas propostas.

A segmentação de imagens sempre foi um passo crucial para o reconhecimento da língua de sinais [17]. Utilizando a localização espacial de objetos, algoritmos baseados em dados de profundidade simplificam essa difícil tarefa, fornecendo resultados robustos e precisos. Como a segmentação não é o foco principal desta pesquisa, a metodologia implementada utiliza uma abordagem comum de "limiarização por profundidade" (Capítulo 2), que define uma caixa fixa regular no espaço da cena, por onde todos os pontos contidos são definidos como parte do segmento da mão gesticulante.

As imagens segmentadas são adquiridas em uma janela de 128×128 *pixel*s, onde cada *pixel* representa a distância de profundidade variando entre 70cm a 110cm do dispositivo sensor. Para normalizar a aquisição e obter uma representação de imagem mais uniforme, solicita-se que o usuário do sistema gesticule, tentando preencher a maior área possível do quadro segmentado.

O denominado "sistema de coordenadas de profundidade" [59] é a representação de dados nativa do sensor *Kinect*. Neste sistema, as coordenadas tridimensionais adquiridas diretamente do sensor são dadas em *pixel*s para os eixos-x,y e em milímetros para a informação de profundidade (eixo-z). Tal representação, no entanto, não é apropriada para a obtenção do movimento rígido pelo alinhamento *ICP*. Por isso, é preciso a conversão de coordenadas para um "sistema de coordenadas global", mais próximo do mundo real, onde se aplicam as propriedades cartesianas

imparcialmente aos três eixos do espaço. Embora a *SDK OpenNI* forneça funções de conversão que, utilizando parâmetros intrísecos da câmera, permita a conversão de coordenadas para o mundo real, verificou-se experimentalmente que o uso destas funções requer cálculos complexos, limitando a eficiência do reconhecimento. Neste sentido, é proposta uma função mais simples de conversão que recupera a extensão máxima do campo de visão do sensor sobre os eixos horizontal e vertical (Figura 4.5), e realiza um escalonamento da imagem, conforme a Equação (4.10). Uma vez que todas as melhorias propostas ao algoritmo *ICP* derivam de propriedades da geometria Euclidiana, espera-se que o uso desta conversão torne o espaço métrico da aquisição dos dados mais compatível e consistente.

$$\begin{pmatrix} x_{\text{global}} = x_{\text{pixel}} \times \frac{\text{CampoVisãoHorizontal}}{\text{ResoluçãoHorizontalPixels}} \\ y_{\text{global}} = y_{\text{pixel}} \times \frac{\text{CampoVisãoVertical}}{\text{ResoluçãoVerticalPixels}} \\ z_{\text{global}} = z_{\text{profundidade}} \end{pmatrix} \tag{4.10}$$

A saída do estágio de pré-processamento é uma mapa de profundidade de até 128×128 (16K) pontos *3D*, contendo informação da mão gesticulante, e da qual todos os passos de reconhecimento serão aplicados. As imagens da base de modelos são pré-processadas da mesma maneira que as imagens de teste na metodologia implementada. Nenhum passo adicional de pré-processamento é feito.

4.4.2 Processamento *ICP*

Este estágio é responsável por aplicar o alinhamento *ICP* entre a imagem de teste, pré-processada, e a base de modelos, recuperando as métricas de correspondência necessárias para o estágio de classificação. A parte mais significativa em termos de tempo de processamento é gasta neste estágio, por conseguinte, a maioria das questões quanto a eficiência do sistema são também tratadas aqui.

Para executar o alinhamento, é necessário definir: (i) um conjunto de parâmetros de instância (Seção 4.2.1); e (ii) uma métrica de correspondência apropriada (Seção 4.2.2) para se acompanhar durante o registro. O propósito principal de selecionar os parâmetros de instância é melhorar a eficiência geral, limitando a complexidade do algoritmo *ICP*. Em contraste, a escolha certa da métrica de correspondência conduzirá a melhores resultados quanto à acurácia.

O banco de modelos utilizado foi construído com 20 amostras de cada uma das 26 posturas estáticas do alfabeto manual. Os alfabetos para a *ASL* e a *Libras* foram analisados independentemente, de forma a se ter um conjunto total de 520 amostras coletadas para cada alfabeto. As imagens foram adquiridas em diferentes condições de luz ambiente, com algumas possíveis variantes de postura, e a partir de um único usuário (Figura 4.6). Mesmo se tendo adquirido variantes de imagens quanto a diferentes condições, comprovou-se que a luz ambiente pouco influencia na aquisição de dados de profundidade com o sensor *Kinect*. Por outro lado, a diversificação da base de dados, em termos de variantes das posturas, foi aplicado por pequenas rotações das posturas padrões das Figuras 2.6 e 2.7.

Finalmente, a aplicação das imagens do banco de modelos proposto deve considerar as técnicas específicas de classificação que serão utilizadas no estágio seguinte. A classificação por *Melhor-Ajuste* encontrará o alinhamento da imagem de teste contra todas as amostras no

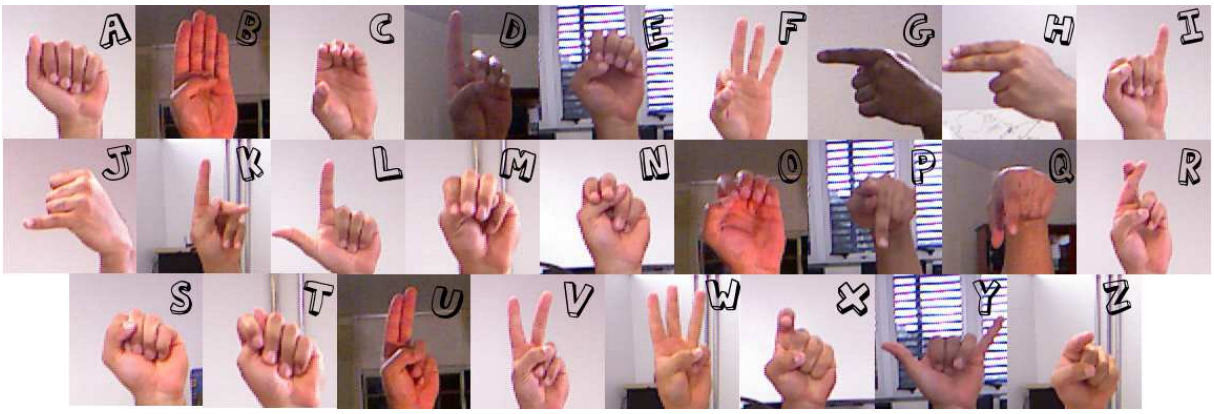


Figura 4.6: Subconjunto ilustrativo de posturas e ambientes de aquisição utilizados na formação do banco de modelos do alfabeto da *ASL*.

banco de modelos, ao passo que a classificação por KB-*Ajuste* tratará elegantemente apenas uma partição destas amostras, melhorando a eficiência do reconhecimento.

4.4.3 Classificação

Classificação é o último estágio na metodologia implementada. A etapa consiste basicamente em reunir todas as inferências computadas e indicar a melhor classe de representação para a imagem de teste. O estágio é computado eficientemente na arquitetura de Casamento de Modelos, sendo ainda mais rápido que o estágio de pré-processamento. Em relação aos algoritmos de classificação propostos, são executadas da ordem de $O(M)$ operações simples para se escolher a melhor classe de equivalência, sendo M um parâmetro da quantidade de modelos alinhados no estágio anterior de processamento *ICP*.

Em termos da acurácia, a técnica *Melhor-Ajuste* (Seção 4.3.1) apresenta os melhores resultados, uma vez que se procura por todo o banco de modelos. Tal propriedade é justificada na hipótese de que o modelo mais próximo encontrado deve ter o melhor valor de similaridade possível com a imagem de teste. Dessa forma, quanto mais amostras o banco de modelos possuir, melhores serão as chances da metodologia implementada encontrar a classe correta de reconhecimento.

Em casos onde a eficiência é um critério essencial, a técnica KB-*Ajuste* (Seção 4.3.2) poderá ser empregada. Ela permite reduzir o número das amostras necessárias para o alinhamento e procura manter o nível da acurácia. No contexto desta técnica, uma análise específica é feita, indicando o melhor equilíbrio entre acurácia e eficiência enquanto se varia o valores de tamanho dos baldes (Capítulo 5).

4.5 Discussão

Neste capítulo, foram introduzidas as principais contribuições deste trabalho visando o aprimoramento do algoritmo *ICP* e da estratégia de Casamento de Modelos aplicados ao reconhecimento de posturas estáticas dos alfabetos manuais. Até onde foi investigado, nenhuma pesquisa conseguiu aplicar o algoritmo *ICP* de forma satisfatória neste contexto. Assim, a motivação das propostas levantadas é de implementar uma metodologia de alto desempenho, em acurácia

e eficiência, voltada para ambientes de reconhecimento em tempo-real. O próximo capítulo abrange o conjunto completo de cenários experimentais que atestam os resultados das propostas formuladas em relação aos trabalhos relacionados.

Capítulo 5

Experimentos e Resultados

Este capítulo apresenta o conjunto de resultados analisados para avaliação da proposta deste trabalho. A Seção 5.1 descreve o ambiente experimental aplicado na verificação dos elementos de acurácia (Seção 5.2) e eficiência (Seção 5.3) da metodologia implementada. Em sequência, a Seção 5.4 realiza um pequeno estudo sobre o impacto do intercâmbio dos alfabetos manuais estudados sobre o sistema de reconhecimento proposto.

5.1 Ambiente Experimental

A metodologia implementada neste trabalho é avaliada sob aspectos de (i) "acurácia": a porcentagem de identificações corretas para um dado cenário de teste; e (ii) "eficiência": tempo de execução do algoritmo *ICP* e taxa média de reconhecimento dos quadros. Para esse propósito, foram conduzidas simulações *offline* e experimentos *online* a partir de diferentes configurações de elementos da proposta. Enquanto as simulações *offline* foram aplicadas para computar a acurácia do sistema e tempo de execução do algoritmo *ICP*, os experimentos *online* foram utilizados para o cálculo da taxa média de reconhecimento.

Como ambiente experimental para verificação de acurácia, este trabalho utiliza variantes ao conceito de validação cruzada. A validação cruzada (do Inglês, *cross-validation*) é uma técnica estatística importante considerada na avaliação de classificadores. Ela consiste em se separar partições ou conjuntos específicos da base de dados, de forma que todo o cenário de avaliação (imagens de teste e modelo) seja formado a partir de uma mesma origem [15]. Com base na implementação dos classificadores propostos, as seguintes variantes são aplicadas:

- ***Leave-one-out*** (utilizado com a classificação *Melhor-Ajuste*): É uma validação cruzada, tomando-se M partições, sendo M o número de modelos da base de dados. Neste caso, toda a base de dados é percorrida, mantendo-se uma instância à parte por vez (imagem de teste) e comparando-a, em seguida, contra todas as outras instâncias da base (imagens de modelo). Este procedimento é atrativo por dois motivos: primeiro porque a maior quantidade de informação possível é testada em cada caso, sendo ideal para aferir a acurácia; segundo, porque o procedimento é determinístico, ou seja, não envolve escolhas aleatórias em cada caso de uso. Dessa forma, este é o procedimento que melhor se alinha com a verificação da acurácia do classificador *Melhor-Ajuste* proposto.
- ***Bootstrap*** (utilizado com a classificação KB*-Ajuste*): É uma validação cruzada baseada no procedimento estatístico de amostragem com reposição. Anteriormente, cada instância

Tabela 5.1: Valores de linha de base para avaliação do desempenho da metodologia implementada.

Parâmetro Estudado	**Valor Padrão**
Número máximo de iterações permitidas (K):	10 iterações
Número máximo de pares correspondentes por iteração (L):	50 pontos
Modificadores da transformação obtida:	Movimento rígido sem escala
Métrica para inferência de similaridades:	*Pontuação* LCP *Média*
Técnica de classificação:	*Melhor-Ajuste*
Alfabeto manual:	*ASL*

de uma imagem de modelo era utilizada uma única vez em um dado caso de uso. A ideia do *bootstrap* é repetir diversos experimentos em cada aplicação. Em cada experimento, separa-se aleatoriamente uma partição da base de dados (imagens de modelo) aplicando-a na comparação cruzada contra as instâncias remanescentes da base (imagens de teste). As instâncias amostradas como modelos por um experimento anterior podem ser escolhidas novamente para compor outras partições nos experimentos seguintes. O valor de acurácia final é obtido do valor médio tomado entre todos os experimentos conduzidos. Dessa forma, este é o procedimento que melhor se alinha com a verificação da acurácia do classificador KB-*Ajuste* proposto.

No intuito de restringir o escopo da análise, e evitar uma explosão combinatorial no número de possibilidades, uma combinação de linha de base foi definida para os parâmetros estudados. Esses valores são apresentados na Tabela 5.1 e a razão específica da escolha destes valores é discutida ao longo do capítulo. Assim, sempre que houver omissão explícita nos resultados quanto ao valor de algum dos parâmetros propostos, esta tabela deverá ser consultada.

Vale recordar da descrição da metodologia proposta (Capítulo 4), que a base de dados é composta por 520 instâncias de imagens de profundidade por alfabeto estudado (*ASL* e *Libras*), totalizando um espaço amostral de 1040 instâncias. Por sua vez, para cada conjunto de 520 imagens inscritas, tem-se uma divisão uniforme de 20 delas para cada uma das 26 letras do alfabeto.

A apresentação dos resultados segue com os cenários de verificação da acurácia da metodologia implementada. Como disposto na Tabela 5.1, a maioria dos resultados dizem respeito a avaliação do alfabeto manual da *ASL*, de forma que os resultados para a *Libras* foram separados e analisados à parte (Seção 5.4).

5.2 Verificação da Acurácia da Metodologia

Como se observa na Figura 5.1, o procedimento *ICP* proposto pode efetivamente identificar as 26 letras do alfabeto manual. Este resultado apresenta evidências de que o alinhamento pode ser propriamente estabelecido sob as circunstâncias de aquisição (sensor *Kinect*) e pré-processamento. Como discutido no Capítulo 3, o cálculo de uma estimativa grosseira inicial para a transformação (alinhamento aproximado), que se aplica anteriormente ao registro *ICP*, não foi necessário.

Figura 5.1: Exemplos de posturas com reconhecimento bem-sucedido a partir do registro *ICP* de imagens de teste (em escala de cinza) com modelos das 26 letras do alfabeto da ASL (em cores).

Os alinhamentos da Figura 5.1 também contestam os resultados de outro trabalho [23], que afirma que o algoritmo *ICP* falharia ao tentar alinhar certas classes da *ASL*, dado que o sensor *Kinect* não recupera a superfície completa de pontos para os modelos. De fato, o uso de um único *Kinect* permite apenas a aquisição de vistas parciais do objeto em um dado instante. Concernentemente à aquisição destas vistas, é possível ainda haver falta de informação dos pontos recuperados pela imagem de profundidade, isto é, o algoritmo interno do *Kinect* para cálculo de disparidades normalmente falha ao detectar a extensão de objetos que sejam paralelos ao eixo-z do sistema referencial de posicionamento do sensor (por exemplo, um dedo indicador apontando para as lentes). No entanto, contrariando a afirmação dos autores, uma implementação cuidadosa do algoritmo *ICP* permite sim recuperar o alinhamento para o propósito de reconhecimento dos alfabetos manuais. Na figura, mesmo imagens de teste incompletas – como das letra 'M', 'P' ou 'Z' em escala de cinza – puderam ser eficazmente alinhadas aos seus respectivos modelos.

Embora promissor, esses resultados não mostram valores quantitativos para aferir a acurácia do sistema com as contribuições propostas. Assim, uma análise mais detalhada da aplicação de toda a metodologia é apresentada antes de se indicar novas conclusões.

5.2.1 Tratamento de Ambiguidades

Um primeiro cenário de validação cruzada do tipo *leave-one-out* é proposto para se observar o conjunto de similaridades inferidas. A Figura 5.2 apresenta a matriz de confusão dos valores brutos de similaridade obtidos, dada a configuração de linha de base da Tabela 5.1. Nesta matriz de confusão, as classes são divididas em 676 blocos (26×26), onde cada bloco equivale à comparação de valores de similaridades entre duas letras do alfabeto da *ASL*. Por sua vez, cada bloco pode ser decomposto em uma matriz de 20×20 elementos, onde cada elemento implica no alinhamento direto de uma imagem de modelo específico (Q_i) com outro modelo específico (Q_j) da base de dados.

Nesta figura, os blocos marcados por quadrados, em verde sobre a diagonal secundária, representam os valores de similaridade para a métrica de "Pontuação *LCP* Média" de instâncias dentro de uma mesma classe. Logo, esses blocos possuem uma intensidade de cor mais clara quando comparados aos demais blocos em suas respectivas linhas.

De outra parte, os blocos restantes marcados por quadrados esparsos, em vermelho, ilustram os valores ambíguos de correspondência entre os alinhamentos. Estes blocos ainda apresentam

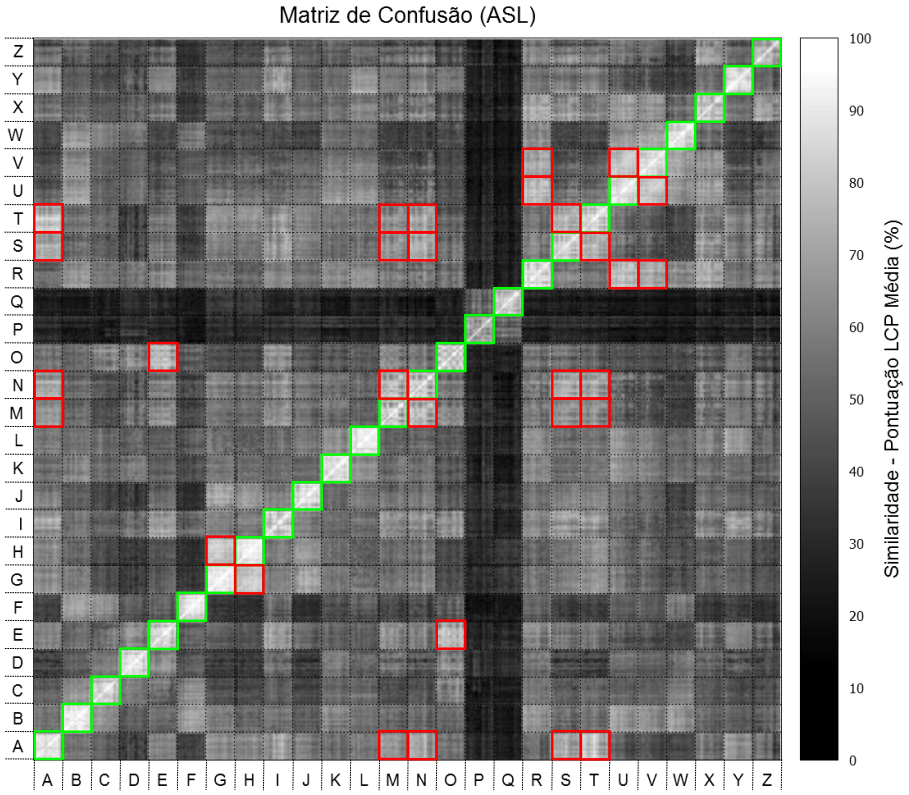


Figura 5.2: Matriz de confusão computada a partir do alinhamento de modelos da *ASL*. A figura traz uma das possíveis simulações aplicadas, utilizando os valores da Tabela 5.1 em ambiente de validação cruzada do tipo *leave-one-out*.

uma intensidade de cor clara, porém de menor grau que os blocos sobre a diagonal secundária. A Tabela 5.2 mostra a similaridade média computada para os elementos destes blocos ambíguos. Com poucas diferenças das ambiguidades sugeridas em [77], os conjuntos mais ambíguos de letras reportados foram: $\{A, M, N, S, T\}$; $\{E, O\}$; $\{G, H\}$; e $\{R, U, V\}$. Entretanto, como pode ser verificado na Tabela 5.2, mesmo nos piores casos, os valores médios computados para identificações corretas diferem em pelo menos 5 pontos percentuais dos valores para classes não-equivalentes. Tal diferença permite que o reconhecimento proposto seja bem-sucedido ao tratar estas ambiguidades.

Vale ressaltar da Figura 5.2, que as marcas escuras em forma de cruz denotam a comparação de alinhamentos incluindo posturas das letras 'P' e 'Q'. A justificativa para esta especificidade é que cada instância destas classes foi adquirida distintamente na metodologia proposta, com porções significantes de pontos pertencentes ao antebraço. Esta condição, no entanto, não foi aplicada propositalmente, mas sim em decorrência das limitações do método de segmentação utilizado (limiarização por profundidade). Tal peculiaridade fez com que estes modelos possuíssem baixa confusão quando alinhados a modelos de outras classes; porém, entende-se que a proposta permanece válida, uma vez que não é comprometida a acurácia final de reconhecimento do sistema.

Por fim, a análise feita nesta seção constitui apenas um dos cenários que foram simulados na geração dos resultados. Na prática, é possível construir uma matriz de confusão diferente com relação: à variação de qualquer parâmetro de instância do *ICP*; à derivação de similaridades por

Tabela 5.2: Valores médios de similaridades para a métrica "Pontuação *LCP* média" (em %) anotados para os conjuntos de blocos mais ambíguos da Figura 5.2.

%	A	M	N	S	T
A	87,72	64,77	68,50	71,32	77,13
M	64,77	78,26	73,06	64,63	63,13
N	68,50	73,06	78,08	71,05	68,97
S	71,32	64,63	71,05	82,73	74,10
T	77,13	63,13	68,97	74,10	83,86

%	E	O
E	83,60	75,14
O	75,14	80,74

%	G	H
G	90,06	79,62
H	79,62	89,38

%	R	U	V
R	85,74	76,61	70,90
U	76,61	86,63	79,35
V	70,90	79,35	86,02

quaisquer métricas de correspondência; ou ao uso das diferentes técnicas de classificação por KB-*Ajuste*. Na Seção 5.4 é apresentada uma análise similar a esta, procurando-se identificar as diferenças entre o conjunto de ambiguidades do alfabeto da *ASL* em relação ao alfabeto da *Libras*.

5.2.2 Avaliação do Reconhecimento

Os resultados apresentados até então mostram que o *ICP* pode gerar valores de similaridades entre os modelos das diferentes classes a partir das definições das métricas de correspondência propostas. Estes resultados são obtidos durante o 1º e 2º estágios da metodologia implementada (Figura 4.5). O papel do 3º estágio é tomar vantagem da diferença de percentual dos valores de similaridade computados para a classe correta em relação as demais. É esta diferença que o classificador implicitamente usa ao rotular uma dada imagem de teste em sua classe de representação. Em um próximo cenário de verificação, a técnica *Melhor-Ajuste* é aplicada em várias simulações de validação cruzada *leave-one-out*. Em um primeiro momento, deseja-se avaliar o comportamento da variação dos elementos iterativos e das métricas de correspondência para a acurácia final do reconhecimento.

Na Figura 5.3, a restrição do "número máximo de iterações" (K) do algoritmo *ICP* é considerada com respeito às diferentes métricas de correspondência. Uma primeira observação é que, para a maioria das métricas propostas, a acurácia reportada está fracamente associada ao valor de K. Nota-se ainda um maior grau de instabilidade do valor de acurácia para as métricas $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ e $\boldsymbol{M}_4$. Pelo reexame da definição destas métricas, percebe-se que todas estão diretamente relacionadas a convergência da função de custo minimizada. Neste caso, estas métricas não avaliam a estrutura global dos modelos alinhados, mas apenas erros de aproximação no registro das superfícies. Além disso, vale mencionar que a própria implementação do algoritmo *ICP* proposto é não-determinística, já que usa amostragens aleatórias de pontos uniformemente distribuídos sobre o espaço de vetores normais. Em observação ao gráfico apresentado, a rápida variação anotada anterior a 10ª iteração para as métricas $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, e $\boldsymbol{M}_4$ diz respeito à estabilização inicial do alinhamento *ICP* com iterações ponto-a-ponto, comentada na proposição do algoritmo no Capítulo 3.

A Figura 5.4 avalia, de forma semelhante, a variação do "número máximo de pontos selecionados" (L) em relação às métricas de correspondência. Percebe-se na figura que as métricas baseadas na pontuação *LCP* adquirem razoável estabilidade previamente às demais, com amostras pequenas de somente 6 pontos. Em todo caso, permanece uma análise similar de que o

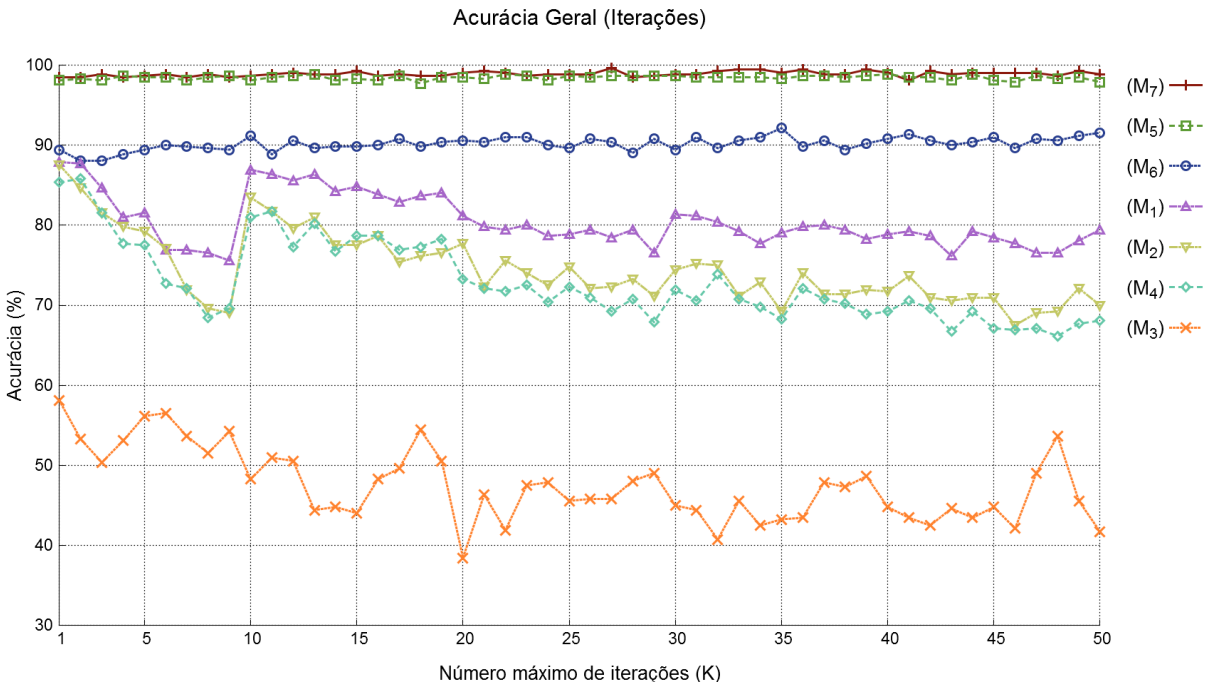


Figura 5.3: Desempenho da acurácia do sistema quanto ao "número máximo de iterações" (K) do registro *ICP*.

parâmetro L a partir de certo valor de amostragem ($L \geq 13$) pouco influencia na melhoria da acurácia.

De ambas as Figuras 5.3 e 5.4, percebe-se que a acurácia do reconhecimento é pouco beneficiada com um valor alto para os parâmetros iterativos do algoritmo *ICP*. Observa-se também que, em geral, as métricas de correspondência possuem certa independência com relação aos elementos de iteração. Em suma, os resultados mostram que a métrica $\boldsymbol{M}_3$ fornece os piores e mais flutuantes valores de acurácia, enquanto que as métricas $\boldsymbol{M}_5$ e $\boldsymbol{M}_7$ atingem uma taxa de acerto de quase 100%, com uma pequena vantagem a favor da métrica $\boldsymbol{M}_7$.

Tais observações acerca das métricas de correspondência são ainda confirmadas ao se investigar a acurácia geral dos "modificadores da transformação obtida" (Tabela 5.3). Com $99{,}04\%$ de acertos, a métrica $\boldsymbol{M}_7$ reconheceu com êxito 515 de 520 possíveis consultas de validação cruzada *leave-one-out*, errando apenas em 5 instâncias derivadas do alinhamento de modelos das letras 'A' e 'T'. Além disso, verificou-se que, ao menos para o ambiente experimental dado, os modificadores propostos não tiveram contribuições significativas no ganho de acurácia.

Com a meta de avaliar a acurácia quando aplicada com a técnica KB-*Ajuste*, um conjunto diferente de simulações de validação cruzada do tipo *bootstrap* foi conduzido. Para lidar com o fator aleatório introduzido pelas escolhas não-determinísticas do Algoritmo 2, foram realizados 100 experimentos distintos para cada possível K valor do tamanho de balde ($1..19$). Em cada experimento, para uma partição fixa de amostras dos baldes (imagens de modelo), a acurácia do algoritmo KB-*Ajuste* foi verificada tomando-se, como imagens de teste, todas as instâncias não selecionadas para escolha dos baldes.

A Figura 5.5 apresenta o valor de acurácia médio dos experimentos realizados. A primeira barra no eixo-x, rotulada por *Melhor-Ajuste*, representa um valor comparativo para a acurácia obtida, utilizando os parâmetros da linha de base (Tabela 5.1). Os resultados mostram que,

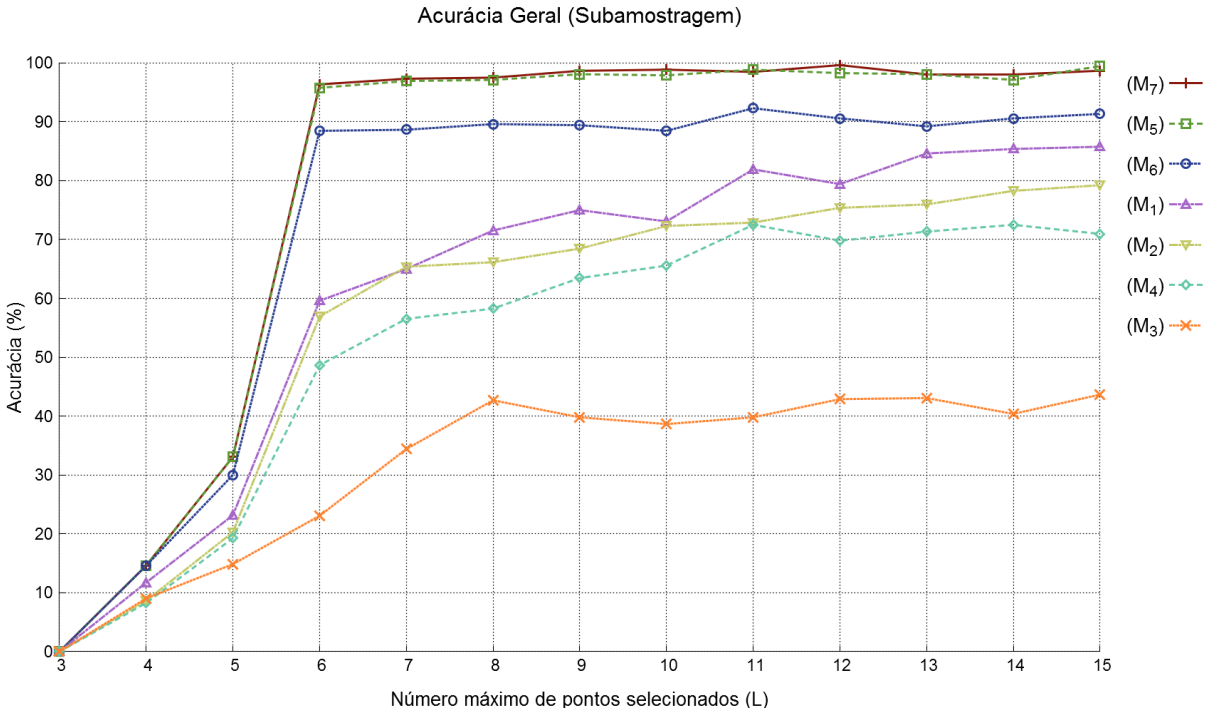


Figura 5.4: Desempenho da acurácia do sistema quanto ao "número máximo de pontos selecionados" (L) em cada iteração do registro *ICP*.

mesmo para o tamanho de balde $K = 1$, a técnica KB-*Ajuste* atinge um desempenho notável em relação às métricas propostas. Neste sentido, a análise estatística aplicada ao algoritmo, por meio do cálculo de valores médios, permite a construção de medidas representativas dos alinhamentos computados. Isto propicia um ganho significativo de estabilidade, mesmo para as métricas consideradas instáveis, como o caso da "Norma Residual de Alinhamento" ($\boldsymbol{M}_3$).

Os resultados de acurácia apresentados para a *ASL* permitem estabelecer uma ordem de classificação por acurácia obtida pelas métricas de correspondência estudadas, determinada do seguinte modo:

- (1ª) Pontuação *LCP* Média ($\boldsymbol{M}_7$);
- (2ª) Pontuação *LCP* Mínima ($\boldsymbol{M}_5$);
- (3ª) Pontuação *LCP* Máxima ($\boldsymbol{M}_6$);
- (4ª) Erro *RMS* Ponto-a-ponto ($\boldsymbol{M}_1$);
- (5ª) Erro *RMS* Ponto-a-plano ($\boldsymbol{M}_2$);
- (6ª) Limiar Dist-max ($\boldsymbol{M}_4$);
- (7ª) Norma Residual do Registro ($\boldsymbol{M}_3$).

A ordem acima obtida indica que as métricas baseadas no conceito de *Pontuação* LCP superam consideravelmente os resultados em acurácia das métricas baseadas em "valor quadrático médio" (*RMS*), comumente encontradas em pesquisas relacionadas ao *ICP* [23, 68, 80]. Entende-se, portanto, que as propostas de $\boldsymbol{M}_5$, $\boldsymbol{M}_6$ e $\boldsymbol{M}_7$ contribuam de forma significativa para o estado da arte no que tange ao reconhecimento de padrões em imagens de profundidade.

Tabela 5.3: Desempenho da acurácia do sistema quanto ao uso das métricas e modificadores no reconhecimento do alfabeto da *ASL*.

	Modificadores da Transformação		
Métricas de Correspondência	*sem modificador*	*escala uniforme*	*escala não-uniforme*
Pontuação LCP *Média* ($\boldsymbol{M}_7$)	98,85%	98,85%	99,04%
Pontuação LCP *Mínima* ($\boldsymbol{M}_5$)	97,88%	98,46%	98,46%
Pontuação LCP *Máxima* ($\boldsymbol{M}_6$)	89,81%	89,23%	89,81%
Erro RMS *Ponto-a-ponto* ($\boldsymbol{M}_1$)	86,73%	87,12%	86,35%
Erro RMS *Ponto-a-plano* ($\boldsymbol{M}_2$)	80,96%	80,96%	80,96%
Limiar Dist-max ($\boldsymbol{M}_4$)	81,15%	82,50%	81,15%
Norma Residual do Registro ($\boldsymbol{M}_3$)	28,08%	25,58%	24,81%

5.3 Verificação da Eficiência da Metodologia

De maneira global, a eficiência do algoritmo *ICP* proposto é dominada apenas pela escolha de seus elementos iterativos: "número máximo de iterações permitidas" (K) e "número máximo de pontos selecionados" (L). Isto é justificável uma vez que outros elementos de implementação, como a aplicação de modificadores ou inferência de métricas, são considerados uma extensão natural do processo de registro. A Figura 5.6 apresenta a correlação entre os elementos iterativos e o tempo médio de processamento de uma instância do alinhamento *ICP* em uma máquina de processador único de $2{,}4$ GHz. Verifica-se que o "número máximo de iterações permitidas" apresenta, em geral, maior influência em relação ao "número máximo de pontos selecionados". Logo, um aumento do valor de K gera um impacto maior sobre o tempo de processamento do alinhamento. Quando ambas variáveis são escolhidas com um valor alto, o tempo de execução do algoritmo *ICP* pode ser até 10 vezes mais lento (≈ 320ms). Ao averiguar a complexidade de tempo do Casamento de Modelos na Equação (4.9), quando um conjunto grande de instâncias de alinhamento são necessárias, esse cenário torna a aplicação da metodologia pouco prática em contextos de tempo-real. Em contraste, como os resultados da acurácia obtida (Figuras 5.3 e 5.4) mostraram não haver ganhos expressivos ao aumentar-se o valor destes parâmetros; valores de configuração mínimos, como os da Tabela 5.1, reduzem positivamente o tempo de processamento do registro (15ms), além de cumprir com os requisitos da acurácia do reconhecimento.

Outra importante análise acerca da velocidade de reconhecimento pode ser feita examinando-se a "frequência média de processamento de quadros" em uma implementação *online* da metodologia proposta. Os resultados indicados na Tabela 5.4 mostram que a técnica KB*-Ajuste* pode alcançar quase duas ordens de magnitude na velocidade de reconhecimento quando comparada a abordagem de *Melhor-Ajuste*. Em contrapartida, dos resultados para a acurácia do método, verifica-se que a técnica KB*-Ajuste* não prejudica substancialmente o desempenho do reconhecimento, mesmo para tamanhos de balde pequenos ($K > 4$). Ou seja, existe uma margem de ganho que favorece a aplicação da técnica de KB*-Ajuste* ao ponderar-se sobre a acurácia e a eficiência do sistema.

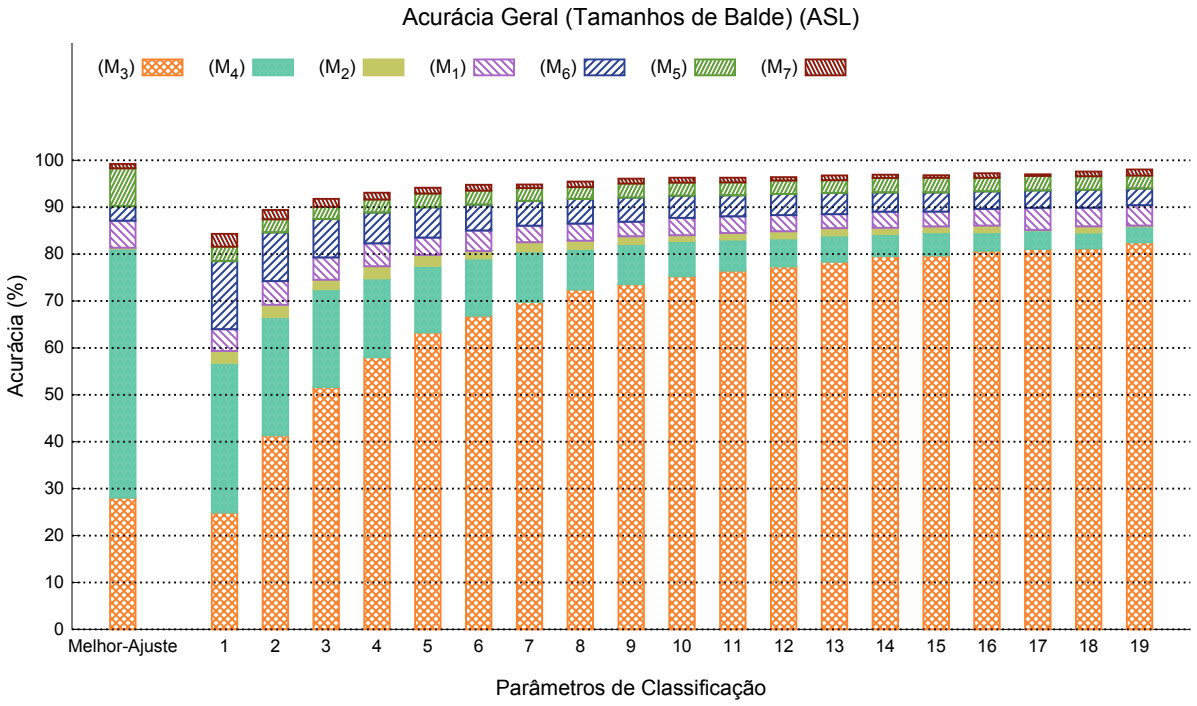


Figura 5.5: Acurácia média da técnica KB-*Ajuste* com tamanhos de balde variáveis para o alfabeto da *ASL*.

Dos resultados da Figura 5.5 e da Tabela 5.4, algumas das correlações entre acurácia e eficiência são resumidas a seguir:

- Dentre todos os resultados, a acurácia obteve seu melhor desempenho com a métrica de "Pontuação *LCP* Média"; portanto esta é a melhor escolha de métrica de correspondência para a metodologia implementada.
- Caso a acurácia seja uma meta essencial ao sistema, a classificação por *Melhor-Ajuste* atinge os maiores desempenhos em troca de um lento processo de reconhecimento ($0{,}20$ *FPS*): $99{,}04\%$ de identificações corretas nos cenários propostos.
- Caso a eficiência seja uma meta essencial ao sistema, a classificação KB-*Ajuste* com $K = 1$ supera qualquer outra técnica proposta para o estágio de classificação. Como

Tabela 5.4: Frequência média de processamento de quadros (*FPS*) anotada para diferentes parâmetros de classificação. Executado em uma máquina de processador único de $2{,}4$ GHz.

Parâmetro	Frequência	Parâmetro	Frequência	Parâmetro	Frequência	Parâmetro	Frequência
Melhor-Ajuste	0,20 *FPS*	**15**	0,41 *FPS*	**10**	0,97 *FPS*	**5**	3,70 *FPS*
19	0,27 *FPS*	**14**	0,48 *FPS*	**9**	1,33 *FPS*	**4**	4,53 *FPS*
18	0,29 *FPS*	**13**	0,55 *FPS*	**8**	1,68 *FPS*	**3**	5,33 *FPS*
17	0,33 *FPS*	**12**	0,67 *FPS*	**7**	2,20 *FPS*	**2**	6,29 *FPS*
16	0,36 *FPS*	**11**	0,81 *FPS*	**6**	2,91 *FPS*	**1**	7,41 *FPS*

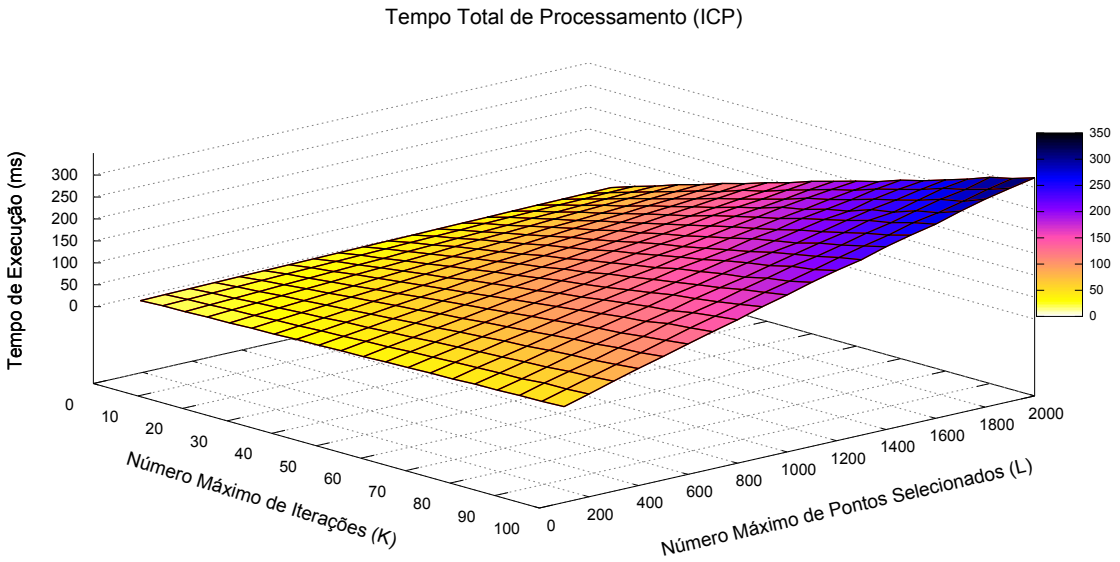


Figura 5.6: Tempo total de processamento do algoritmo *ICP* a partir da variação de seus elementos iterativos. Executado em uma máquina de processador único de $2{,}4$ GHz.

indicado pelos resultados, esta técnica mantém uma razoável acurácia ($84{,}31\%$) enquanto se atinge uma frequência média de reconhecimento de $7{,}41$ *FPS*.

- Um bom equilíbrio entre acurácia e eficiência é aplicar a técnica KB*-Ajuste* com $K = 5$: permite que níveis altos de acurácia ($94{,}16\%$) sejam obtidos, com uma frequência média de $3{,}70$ *FPS*.

O uso da metodologia implementada provou fornecer um sistema confiável de reconhecimento dos alfabetos manuais. Esta credibilidade pode ser avaliada pela comparação da acurácia obtida contra as acurácias dos trabalhos relacionados ao estado da arte (Tabela 3.1). Ao mesmo tempo, a metodologia possui uma desvantagem na velocidade de reconhecimento, resultado da aplicação da estratégia de Casamento de Modelos. A classificação por KB*-Ajuste* limita este problema ao reduzir a complexidade em espaço da base de dados; tornando possível a obtenção de bons resultados de acurácia e suporte a aplicações *online*.

5.4 Análise Comparativa do Reconhecimento entre Alfabetos

Esta seção abrange uma análise da metodologia implementada quanto à comparação dos resultados de reconhecimento para os alfabetos manuais da *ASL* e da *Libras*. Uma primeira observação, visto que ambos os alfabetos possuem 26 letras, é que a permuta destes não acarreta uma perda expressiva de desempenho quanto à eficiência do sistema. Neste caso, os resultados de verificação da eficiência da metodologia continuam valendo para o reconhecimento da *Libras*. Uma segunda ponderação a ser feita é que muitas das letras do alfabeto da *ASL* possuem representações idênticas na *Libras*. Desta forma, para abranger um maior número de resultados, optou-se por variar as posições de reconhecimento para algumas letras do alfabeto da *Libras*. Deve-se dizer, no entanto, que tais variações são aceitas oficialmente, tratando-se de diferentes

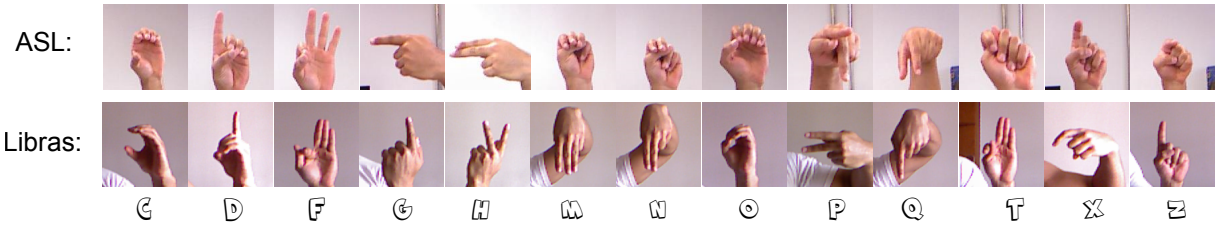


Figura 5.7: Conjunto de posturas distintas entre os alfabetos manuais da *ASL* e *Libras*.

dialetos de comunicação. Desta forma, pode-se dividir o conjunto comparativo de variações estudadas dos alfabetos nas seguintes categorias:

- 13 classes com representações equivalentes: $\{A, B, E, I, J, K, L, R, S, U, V, W, Y\}$;
- 4 classes com variações introduzidas: $\{C, D, O, Z\}$ (Figura 5.7);
- 9 classes distintas por origem do alfabeto: $\{F, G, H, M, N, P, Q, T, X\}$ (Figura 5.7).

Uma primeira análise proposta quanto ao alfabeto da *Libras* consiste na identificação de suas classes de reconhecimento ambíguas. A Figura 5.8 traz a matriz de confusão para as posturas da *Libras* em um cenário análogo ao proposto anteriormente para a *ASL*. Assim como para o alfabeto da *ASL*, a *Libras* apresenta um conjunto de classes para as quais se identifica valores de similaridade relativamente próximos: $\{A, S\}$; $\{B, F, T\}$; $\{C, D, O\}$; $\{M, N, Q\}$; e $\{R, U, V, Z\}$. A partir destes conjuntos, computaram-se os valores médio das similaridades entre blocos (Tabela 5.5), dos quais se fazem análises interessantes:

- Grande parte dos conjuntos ambíguos se referem a posturas que apresentam estruturas de mãos muito semelhantes. Em geral, estas posturas variam com relação a um único dedo que, ora está estendido ($\{A, B, D\}$), ora recolhido ($\{S, F, T, C, O\}$). No caso do conjunto $\{M, N, Q\}$, a forma da mão gesticulante é idêntica, variando-se apenas a quantidade de dedos que apontam para baixo. Analiticamente, estas classes ambíguas que se diferem por pequenas variações de um único dedo, acusam uma margem consistente de $\approx 10\%$ de diferença entre as médias de similaridade.

Tabela 5.5: Valores médios de similaridades para a métrica "Pontuação *LCP* média" (em %) anotados para os conjuntos de blocos mais ambíguos da Figura 5.8.

%	R	U	V	Z
R	87,66	82,37	75,60	78,63
U	82,37	91,45	82,80	78,68
V	75,60	82,80	90,11	75,33
Z	78,63	78,68	75,33	88,54

%	B	F	T
B	92,85	75,44	78,24
F	75,44	91,74	88,51
T	78,24	88,51	93,25

%	M	N	Q
M	87,65	81,71	73, 59
N	81,71	87,65	76,23
Q	73,59	76,23	82,97

%	C	D	O
C	88,21	69,90	77, 23
D	69,90	90,95	76,15
O	77,23	76,15	90,58

%	A	S
A	91,17	79,91
S	79,91	90,23

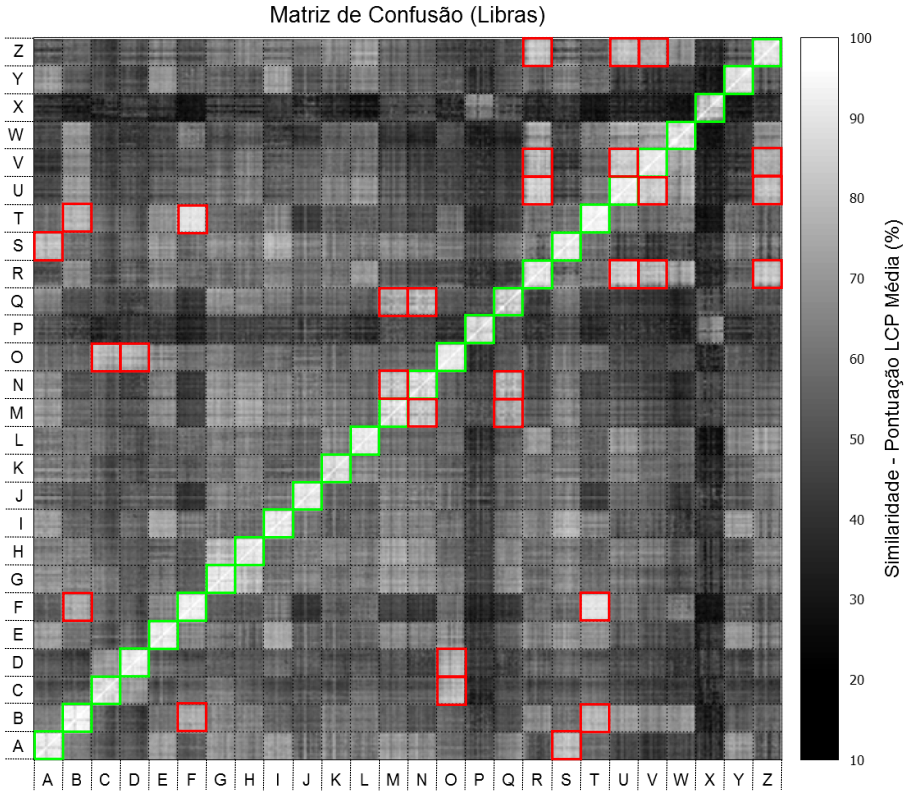


Figura 5.8: Matriz de confusão computada a partir do alinhamento de modelos da *Libras*. A figura traz uma das possíveis simulações aplicadas, utilizando os valores da Tabela 5.1 em ambiente de validação cruzada do tipo *leave-one-out*.

- Um caso peculiar se observa dos conjuntos $\{C, O\}$ e $\{D, O\}$. Da diferença dos valores médios computados, percebe-se que a classe 'C' é relativamente próxima à classe 'O' ($\approx 11\%$ de diferença) e que a classe 'D' também se aproxima à classe 'O' ($\approx 14\%$ de diferença); todavia as letras 'C' e 'D' estão afastadas uma da outra por uma margem de $\approx 20\%$.
- O pior caso de toda a análise de ambiguidades, incluindo-se a do alfabeto da *ASL* (Subseção 5.2.1), se deve ao conjunto $\{F, T\}$ de classes da *Libras*. As imagens de profundidade destas posturas são muito próximas, de forma que a diferença de valor médio das similaridades computadas é de apenas $\approx 3{,}5\%$.

As análises sobre o conjunto de ambiguidades propõem que a metodologia implementada é capaz de reconhecer, com certa margem de confiança, as letras dos alfabetos manuais estudados. Entretanto, verifica-se ainda uma carência de resultados que indiquem especificamente em quais classes o uso de um dos alfabetos é melhor aplicável que o outro. Desta forma, pode-se ter uma noção geral de qual alfabeto o sistema proposto melhor se adapta. A fim de construir tal análise com base no ambiente de validação cruzada *leave-one-out*, introduz-se o conceito de *média diferencial da confusão de 'l' relativa a 'Σ'* (ξ_l^{Σ}), definido por:

$$\xi_l^{\Sigma} = \mathrm{E}_m[\{l\}] - \mathrm{E}_m[\overline{\{l\}}], \tag{5.1}$$

onde:

- 'Σ' é um dos alfabetos manuais estudados: *ASL* ou *Libras*;
- 'l' é uma letra (classe) do alfabeto 'Σ';
- 'E$_m[\,]$' é o valor esperado de similaridade para um conjunto de alinhamentos, com respeito a métrica 'm';
- '$\{l\}$' é o conjunto de alinhamentos realizados com pares teste-modelo congruentes a uma mesma classe 'l' (ambas as imagens devem ser representações de 'l' por 'Σ');
- '$\overline{\{l\}}$' é um conjunto complementar a '$\{l\}$', formado por alinhamentos de pares teste-modelo; onde a imagem de teste represente a classe 'l', e a imagem de modelo represente uma classe diferente de 'l' do alfabeto 'Σ'.

A interpretação estatística do valor de ξ_l^Σ permite quantificar o afastamento da confusão de l em relação às demais letras do alfabeto Σ. Visualmente, o cálculo do valor ξ_l^Σ pode ser observado nas matrizes de confusão apresentadas (Figuras 5.2 e 5.8). Neste caso, cada valor ξ_l^Σ corresponde a um cálculo efetuado sobre uma das linha de blocos da matriz, computando-se a diferença entre as médias de similaridades do bloco da linha referente a diagonal principal (correspondente a letra l) e a média de similaridades dos demais blocos desta mesma linha. Deste ponto em diante, a "Pontuação *LCP* Média" ($\boldsymbol{M}_7$) será a métrica m utilizada no cálculo de ξ_l^Σ, dado que a análise da acurácia indicou melhores resultados para esta métrica.

Uma vez que se obtenha o valor ξ_l^Σ para cada letra de um alfabeto, é possível também analisar o conjunto de afastamento global da confusão deste alfabeto. A *média diferencial da confusão de* 'Σ' (ξ^Σ) é dada por:

$$\xi^\Sigma = \mathrm{E}[\{\xi_l^\Sigma\}], \tag{5.2}$$

onde:

- 'E$[\,]$' é o valor esperado estimado pela média aritmética de um conjunto de valores;
- '$\{\xi_l^\Sigma\}$' é o conjunto de "médias diferenciais da confusão" para todas as letras 'l' do alfabeto 'Σ'.

Considerando as definições das Equações (5.1) e (5.2), uma síntese de todos os afastamentos computados para os alfabetos manuais em estudo é apresentada na Figura 5.9. Da última barra, indicada por "global", obtemos uma relação das médias diferenciais para todo o alfabeto da *ASL* e da *Libras*. Esta barra mostra que, em geral, os valores obtidos para similaridades das letras identificadas em classes corretas se diferem em média em quase 35 pontos percentuais das similaridades encontradas para classificações incorretas. Interessante notar que, embora cada letra apresente um comportamento individual de suas médias diferenciais com relação a um dado alfabeto, a avaliação global dos alfabetos mostra que ambos estão bem próximos, observando-se apenas $0{,}46\%$ de vantagem para o alfabeto da *ASL*.

Individualmente a cada letra dos alfabetos, alguns fatos são comentados:

- Curiosamente as letras 'M', 'R' e 'Z' apresentaram os menores índices de afastamento médio, com máximas de $28{,}06\%$, $28{,}63\%$ e $30{,}19\%$, respectivamente. Isto mostra que

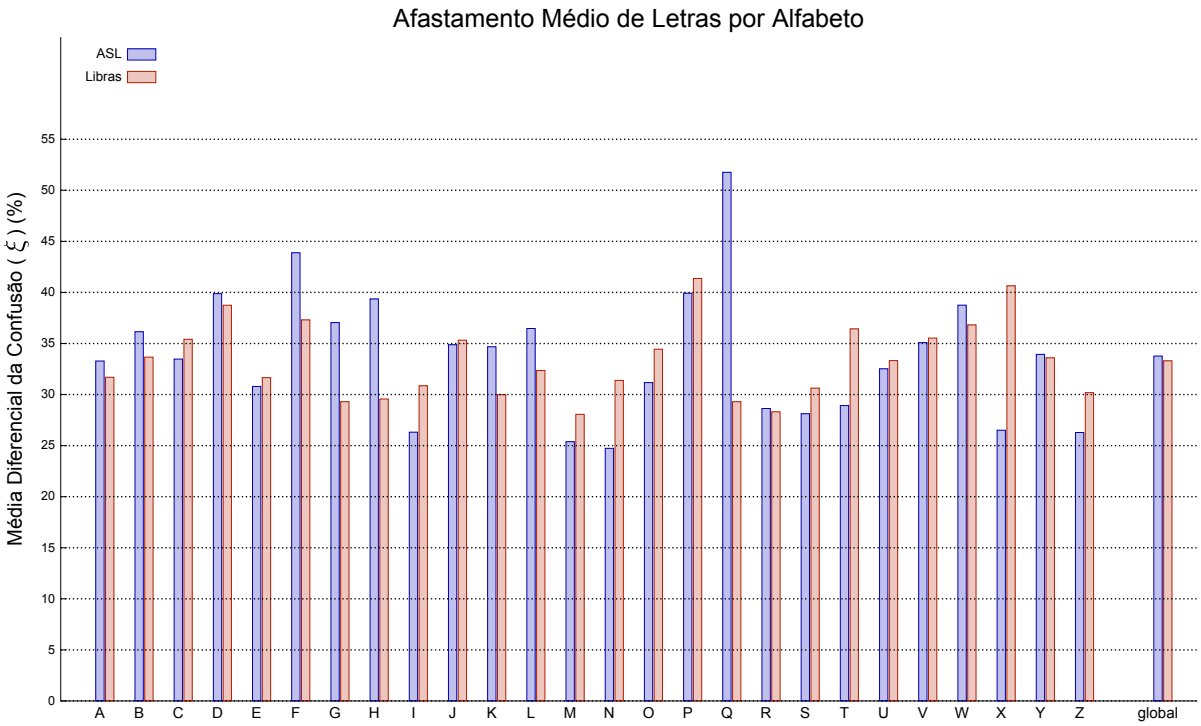


Figura 5.9: Nível de afastamento médio entre as 26 letras dos alfabetos *ASL* e *Libras*.

estas letras são as mais confundidas sob um aspecto geral de ambos os alfabetos manuais. Todavia, os valores percentuais do afastamento médio são altos, mostrando que a metodologia implementada não as tomaria como casos incertos de reconhecimento;

- A maior diferença de afastamento médio entre a *ASL* e a *Libras* se observa nas letras 'Q' e 'X'. A letra 'Q', com $22{,}47\%$ de vantagem para a *ASL*, se justifica no já comentado fato de que suas imagens de profundidade foram adquiridas com a captura de pontos significativos dos antebraços, alheios à representação das mãos. Teoricamente tal vantagem, para a *ASL*, também se observaria com relação às imagens de profundidade da letra 'P', cuja segmentação inclui pontos pertencentes ao antebraço. Porém, como a leitura do gráfico indica, as representações da letra 'P' capturadas na *Libras* são também *sui generis* em relação ao restante das classes, visto que suas imagens de profundidade possuem a componente do eixo principal dominante rotacionada por $90°$ em sentido anti-horário. Este último fato, por sua vez, é também o motivo observado na segunda maior diferença entre os alfabetos, indicada com respeito à letra 'X', apresentando $14{,}14\%$ de vantagem de afastamento para a *Libras*.
- Uma estatística que vale ser ressaltada se deve a partição das 26 letras que obtém vantagens de afastamento médio em cada alfabeto:
  - 12 letras para o alfabeto da *ASL*: $\{A, B, D, F, G, H, K, L, Q, R, W, Y\}$;
  - 14 letras para o alfabeto da *Libras*: $\{C, E, I, J, M, N, O, P, S, T, U, V, X, Z\}$.
- Enfim, como esperado, as letras com as maiores diferenças de afastamento entre os alfabetos, a partir da categorização apresentada no início desta seção, confirmam-se entre

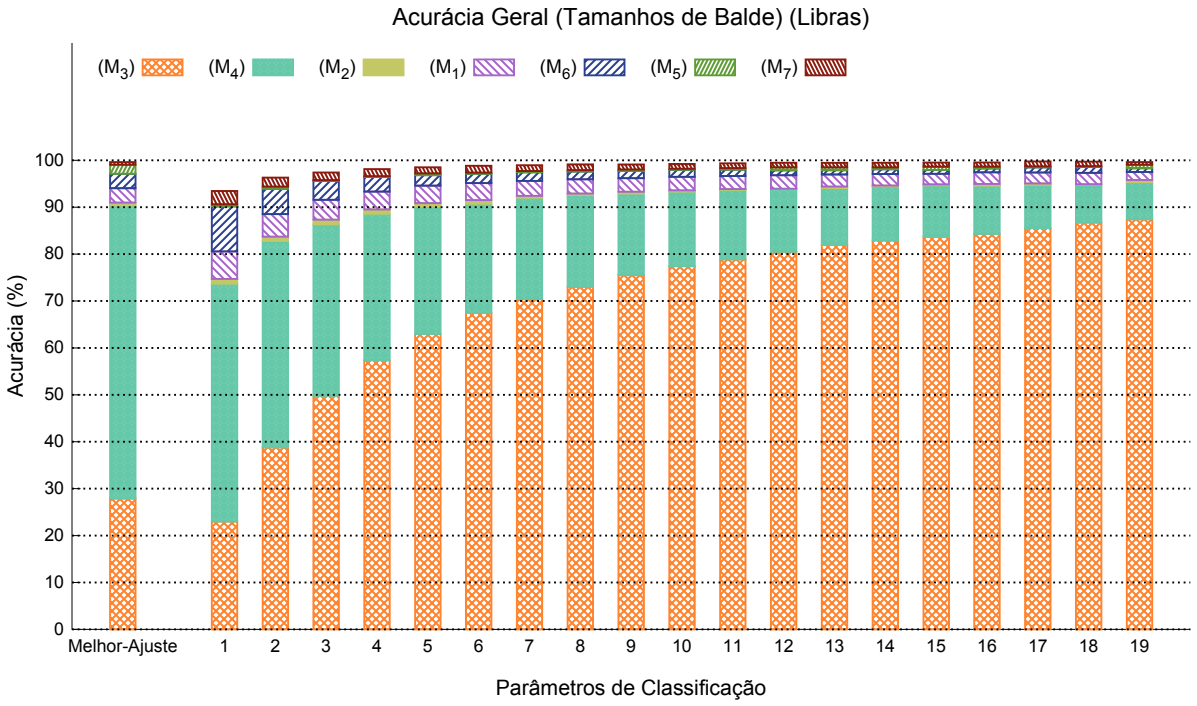


Figura 5.10: Acurácia média da técnica KB-*Ajuste* com tamanhos de balde variáveis para o alfabeto da *Libras*.

as que foram adquiridas com variações introduzidas e as que são distintas por origem do alfabeto (Figura 5.7).

Outro resultado, complementar à análise da *Libras*, diz respeito à acurácia do reconhecimento reportada para este alfabeto, apresentada na Figura 5.10. As condições para construção deste resultado foram as mesmas descritas no ambiente experimental para análise da *ASL*, isto é, com os dois tipos de cenário de validação cruzada: *leave-one-out* para o classificador *Melhor-Ajuste*; e *bootstrap* para as diferentes instâncias do classificador KB-*Ajuste*. Desta figura se extraem as seguintes características:

- Valores mais altos de acurácia são reportados. De fato, o reconhecimento por *Melhor-Ajuste* obteve uma taxa de acerto máximo de $99{,}62\%$ das classificações de validação cruzada, errando apenas 2 de 520 instâncias possíveis: divergências pela comparação de modelos das classes $\{M, N\}$ e $\{N, Q\}$.
- Em relação às métricas de correspondência, as leituras reiteram a ordem de classificação descrita na Seção 5.2.2; e mostram que, à exceção da "Norma Residual do Registro", todos os valores de acurácia para as métricas de correspondência apresentam-se em níveis superiores àqueles apontados para a *ASL*.
- Da análise dos parâmetros do algoritmo KB-*Ajuste*, o gráfico aponta um ganho significativo de acurácia relativamente à análise da *ASL*. Por exemplo, com máxima de $93{,}45\%$, a técnica KB-*Ajuste*, com $K = 1$, para o alfabeto da *Libras* supera a acurácia reportada por seu uso análogo no alfabeto da *ASL*, que apresenta $84{,}31\%$.

Por fim, da visão global dos resultados de acurácia para a *Libras*, vale ressaltar que a técnica KB-*Ajuste*, com $K = 5$, permanece sendo uma escolha balanceada entre acurácia (máxima de $98{,}51\%$) e eficiência ($3{,}70$ *FPS*) dos classificadores nos experimentos *online*.

5.5 Discussão

Este capítulo apresentou um amplo rol de experimentos realizados para avaliar a metodologia implementada. Os cenários de teste foram construídos visando a obtenção de resultados que atestam a acurácia e a eficiência do método com respeito aos diferentes elementos de implementação da proposta. Do conjunto de resultados esperados e hipóteses formuladas (Capítulo 4) para estes elementos, observou-se que:

- O aumento do "número máximo de iterações" (K) do registro *ICP* não influenciou na obtenção de níveis altos da acurácia do sistema (Figura 5.3); ao passo que uma restrição de K implica uma redução significativa no tempo total de processamento do registro (Figura 5.6).
- De modo semelhante, o aumento do "número máximo de pontos selecionados" (L) em cada iteração do registro *ICP* também não influenciou na acurácia do sistema (Figura 5.4); enquanto que uma subamostragem destes pontos reduz consideravelmente o tempo total de processamento do registro (Figura 5.6).
- A adição dos "modificadores da transformação obtida" propostos não levou a ganhos expressivos de acurácia (Tabela 5.3).
- A análise das "métricas de correspondência" aplicadas na inferência de similaridades apresentam diversas conclusões:
  - pelo exame dos níveis de confusão para os valores de similaridades extraídos (Figuras 5.2 e 5.8), observa-se que o algoritmo *ICP* permite distinguir marginalmente a equivalência ou não equivalência na comparação de dois modelos representantes das classes do alfabeto;
  - os resultados de acurácia que apresentam a comparação entre as métricas estudadas (Figuras 5.3, 5.4, 5.5 e 5.10; e Tabela 5.3) são coerentes ao estabelecer uma ordem de melhor aplicação, definida por: $\boldsymbol{M}_7$, $\boldsymbol{M}_5$, $\boldsymbol{M}_6$, $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_4$, $\boldsymbol{M}_3$.
  - quantitativamente, o melhor índice de acurácia anotado deve-se a métrica de "Pontuação *LCP* Média", com $99{,}62\%$ de taxa de acerto no reconhecimento de posturas do alfabeto manual da *Libras*.
- A técnica de classificação *Melhor-Ajuste* serviu a seu propósito de avaliação da acurácia para as diferentes configurações propostas. Apesar de apresentar os valores mais altos de acurácia (Figuras 5.5 e 5.10), a comparação a cada caso com todos os modelos da base de dados o distanciou de resultados práticos para a velocidade de reconhecimento no experimento *online* (Tabela 5.4).
- Por sua vez, a classificação KB-*Ajuste* realmente apresentou ganhos na velocidade de reconhecimento (Tabela 5.4), sem comprometer, entretanto, a acurácia da metodologia implementada (Figuras 5.5 e 5.10).

Em conjunto às verificações sobre a acurácia e eficiência da proposta, a Seção 5.4 introduziu resultados comparativos para os dois alfabetos estudados: *ASL* e *Libras*. Em termos das similaridades computadas e do nível de confusão reportado entre as classes, os dois alfabetos apresentaram comportamentos semelhantes (Figuras 5.2, 5.8 e 5.9), aparentando não haver um alfabeto que melhor se destaca para o reconhecimento. Entretanto, pela comparação das Figuras 5.5 e 5.10, percebe-se uma ligeira vantagem da acurácia do reconhecimento das posturas para *Libras* em detrimento das posturas da *ASL*. Visto que os dois alfabetos possuem uma quantidade idêntica de classes (letras), não se encontraram diferenças significativas da eficiência do método em relação a cada alfabeto.

Capítulo 6

Conclusão

A língua de sinais não é só uma forma natural de comunicação entre surdos e deficientes auditivos, é também um importante mecanismo de inclusão que deve ser conhecido e estudado pela sociedade como um todo. Semelhantemente às línguas orais, a representação por sinais constitui-se em uma complexa estrutura linguística, fornecendo recursos expressivos suficientes que permitam aos seus usuários expor ideias em relação a quaisquer assuntos e situações. Em particular, o "alfabeto manual" apresenta uma série de gestos, sendo parte integrante de uma língua de sinais. Neste caso, a doletração constitui a principal aplicação prática do alfabeto manual e é utilizada entre seus praticantes, por exemplo, ao transmitir nomes pessoais ou perguntar a qual sinal da língua corresponde um determinado conceito.

Este trabalho propôs um sistema de reconhecimento automático das 26 posturas estáticas representantes das letras dos alfabetos manuais: da (i) Língua de Sinais Americana (*ASL*); e da (ii) Língua Brasileira de Sinais (*Libras*). Para atingir este objetivo, a metodologia do sistema emprega o sensor de profundidade *Kinect* na aquisição de dados; e, de posse das imagens de profundidade, aplica uma estratégia de Casamento de Modelos conjuntamente ao algoritmo de alinhamento *ICP* na etapa de reconhecimento.

Contrariando investigações similares [23], os resultados apresentados mostraram que o algoritmo *ICP* pode ser utilizado para produzir casamentos corretos entre as classes do alfabeto, mesmo quando um conjunto próximo (ambíguo) de posturas gestuais é aplicado. Com acurácias reportadas de $99{,}04\%$ (*ASL*), e de $99{,}62\%$ (*Libras*), a metodologia implementada atestou ser adequada suficiente para o reconhecimento dos alfabetos manuais. No entanto, por estar condicionado ao alinhamento pareado do algoritmo *ICP*, o paradigma de Casamento de Modelos é ainda uma limitação identificada para a aplicação do método em contextos de tempo-real (> 15 *FPS*).

Sob uma análise técnica das contribuições originais deste trabalho, duas delas apresentam maior relevância para o estado da arte:

- A proposta de um conjunto diversificado de métricas de correspondência possibilitou uma análise abrangente sobre a aplicabilidade do algoritmo *ICP* para o reconhecimento de formas *3D*. Neste ponto, identificou-se que as métricas propostas que refletem a composição estrutural dos modelos ($\boldsymbol{M}_5$, $\boldsymbol{M}_6$ e $\boldsymbol{M}_7$) apresentaram melhores resultados de acurácia no reconhecimento do que aquelas que utilizam apenas informação do cálculo de otimização do registro ($\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ e $\boldsymbol{M}_4$);
- A proposta do algoritmo de "Ajuste Aproximado por K-Baldes" (KB*-Ajuste*) contribuiu para um aumento de desempenho na eficiência computacional da metodologia implementada. Os resultados para os experimentos *online* indicaram que a técnica KB*-Ajuste*,

com $K = 1$, atinge uma frequência média de aquisição de processamento de quadros de $7{,}41$ *FPS* contra $0{,}20$ *FPS* da técnica de força bruta (*Melhor-Ajuste*). Verificou-se ainda que a eventual partição de modelos do banco de dados pela técnica KB*-Ajuste* não comprometeu significativamente a acurácia do sistema.

Por fim, como apresentado a seguir, entende-se que a metodologia e técnicas propostas possam contribuir positivamente não só com a extensão de trabalhos aplicados ao reconhecimento das línguas de sinais, mas também para a ampliação de pesquisas diversas com o estudo do casamento de padrões.

6.1 Propostas para Trabalhos Futuros

Esta seção apresenta um rol não-exaustivo de propostas identificadas para trabalhos futuros. Como forma de apresentação, estas propostas foram divididas em: (i) "extensões do trabalho", contribuições que podem ser diretamente integradas na metodologia ou aplicáveis desta; e (ii) "evoluções da pesquisa", sugestões de trabalhos que utilizem algumas das técnicas propostas.

Extensões do Trabalho

As seguintes extensões diretas deste trabalho foram identificadas:

- Da análise de reconhecimento dos alfabetos estudados, espera-se que o sistema proposto possa ser prontamente adaptável para outros alfabetos de sinais. Na verdade, dada a complexidade e proximidade das posturas corretamente identificadas, a expectativa é de que o sistema seja capaz de reconhecer um conjunto personalizado de posturas estáticas, não necessariamente vinculadas aos alfabetos manuais, exigindo-se apenas a troca da base de dados da metodologia proposta.
- A codificação do registro *ICP* proposto adicionalmente à aceleração por *hardware*, tal como na aplicação das Unidades de Processamento Gráfico de Propósito Geral (do Inglês, *General-Purpose computing on Graphics Processing Units – GPGPU*s), aparenta ser uma alternativa razoável para a utilização do sistema em contextos de tempo-real.
- Na tentativa de aprimorar a interface natural com o usuário (*NUI*), foram realizados testes preliminares (Figura 6.1) com o algoritmo de segmentação e rastreamento das mãos por imagens de intensidade, proposto em [37, 39].
- Realizou-se ainda um estudo sobre a regularidade das imagens de profundidade obtidas com o sensor *Kinect*. Este estudo é justificável pelo fato de que alguns dos modelos de posturas adquiridos apresentavam uma superfície incompleta de pontos *3D*, quando comparados com suas respectivas imagens de intensidade. Avaliou-se, então, a proposta de regularizar as imagens de profundidade por meio do preenchimento dos pontos faltantes. Uma primeira abordagem idealizada aplica o algoritmo *floodfill* (uma técnica simples de processamento de imagens) sobre a região faltante, considerando filtros de máscara de tamanho fixo da vizinhança de pontos com representação inicial nas imagens de profundidade (Figura 6.2).

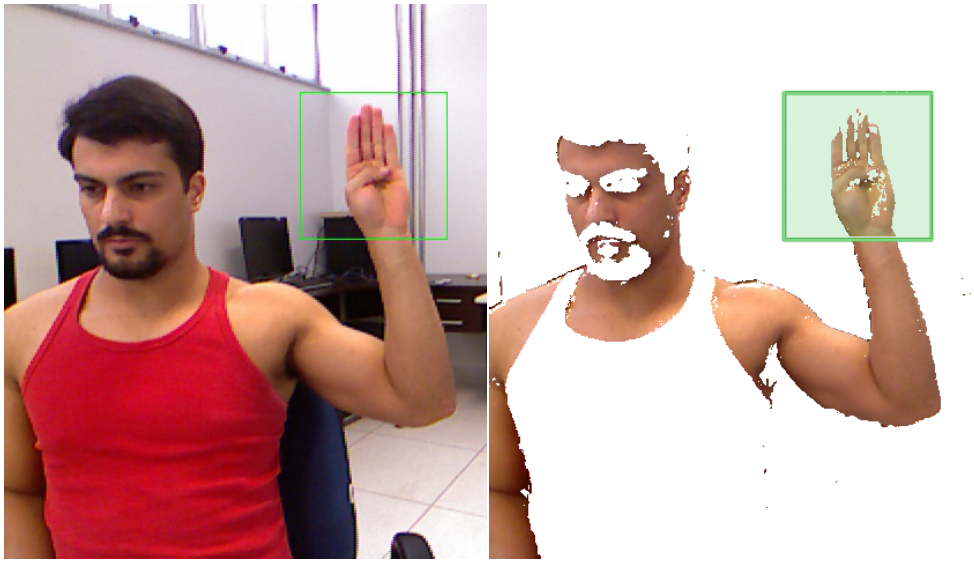


Figura 6.1: Resultados preliminares com a estratégia de segmentação e rastreamento de mãos apresentada em [37, 39].

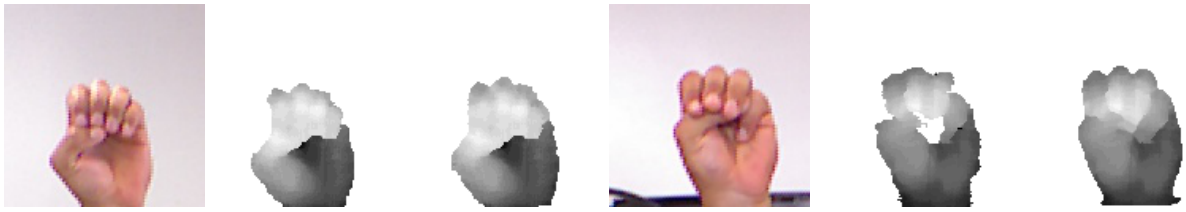


Figura 6.2: Resultados preliminares para uma proposta de regularização das imagens de profundidade obtidas com o sensor *Kinect*.

- A exemplo de outros trabalhos [15, 77], uma possível extensão direta desta pesquisa consiste na implementação de uma aplicação prática de reconhecimento para a datilologia (soletração manual de palavras). Entende-se que esta extensão apresente novos desafios, como resolver o problema de coarticulação entre as letras, porém traria ainda mais benefícios práticos para a comunidade de surdos e deficientes auditivos.

Evoluções da Pesquisa

Um conjunto de possíveis evoluções da pesquisa realizada inclui as seguintes considerações:

- Dado o relativo ganho de desempenho da técnica KB-*Ajuste*, acredita-se que esta possa ser utilizada em outras interessantes aplicações da estratégia combinada de Casamento de Modelos com o algoritmo *ICP*. De fato, uma gama maior de trabalhos com esta combinação é encontrada na área de reconhecimento biométrico [68, 80]. Logo, não é difícil pensar em aplicações do que foi apresentado nesta dissertação em pesquisas com biometria.
- Outra evolução considerada consiste em aplicar a metodologia implementada para o reconhecimento de gestos em outros contextos de interação humano-computador ou em navegação robótica [41, 89]. Neste caso, um dos pontos atrativos da pesquisa consiste na

aplicação dos aprimoramentos propostos para o registro *ICP* no rastreamento e reconhecimento simultâneo de gestos ou do ambiente de navegação.

- Finalmente, considera-se também uma pesquisa sobre o potencial da metodologia proposta e as modificações necessárias para prover o reconhecimento de sinais dinâmicos, isto é, considerando o gesto como um conjunto de posturas, possivelmente diferentes, encenadas em um certo tempo. Neste sentido, a hipótese é de que o algoritmo *ICP* possa rastrear de forma robusta o alinhamento de duas formas geométricas próximas no espaço e no tempo.

Referências

[1] D. Melcher. Visual stability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1564):468–475, 2011. (Citada nas páginas 1 e 6).

[2] R. Jain, R. Kasturi, e B. G. Schunck. *Machine vision*. McGraw-Hill, Inc., 1995. (Citada nas páginas 1, 6, 9, e 10).

[3] S. Mitra e T. Acharya. Gesture recognition: a survey. *Systems, Man and Cybernetics, Part C (Applications and Reviews), IEEE Transactions on*, 37(3):311–324, 2007. (Citada nas páginas 1, 2, 14, e 23).

[4] T. B. Moeslund e E. Granum. A survey of computer vision-based human motion capture. *Computer Vision Image Understanding*, 81(3):231–268, 2001. (Citada nas páginas 1 e 17).

[5] Microsoft Corporation. Redmond WA. Kinect for Xbox 360, 2010. (Citada nas páginas 1, 3, 10, 11, e 49).

[6] R. N. Almeida. Portuguese sign language recognition via computer vision and depth Sensor. Master's thesis, Lisbon University Institute, Department of Science and Information Technology, 2011. (Citada nas páginas 1, 3, 12, 18, 23, 24, 26, e 28).

[7] M. P. Lewis, editor. *Ethnologue: languages of the world*. SIL International, Dallas, TX, USA, 16a edition, 2009. (Citada na página 2).

[8] R. H. Damasceno e M. C. S. Domingos. *Libras: sinais de inclusão*. Unifenas, Alfenas, Arte Gráfica Atenas edition, 2010. Cartilha. (Citada nas páginas 2 e 17).

[9] Instituto Brasileiro de Geografia e Estatística. Censo demográfico, 2000. (Citada na página 2).

[10] M. E. Al-Ahdal e N. M. Tahir. Review in sign language recognition systems. Em *Computers Informatics (ISCI), 2012 IEEE Symposium on*, pages 52–57, 2012. (Citada nas páginas 2, 20, e 21).

[11] M. V. Lamar, S. Bhuiyan, e A. Iwata. Hand gesture recognition using T-CombNET: a new neural network model. *Information and Systems, IEICE Transactions on*, E83-D(11):1986–1995+, 2000. (Citada nas páginas 2, 17, e 24).

[12] R. Y. Wang e J. Popović. Real-time hand-tracking with a color glove. *ACM Trans. Graph.*, 28(3):63:1–63:8, 2009. (Citada nas páginas 2 e 17).

[13] Cyber Glove Systems. Cyber Glove III datasheet, 2010. (Citada na página 2).

[14] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, e J. M. Brady. A linguistic feature vector for the visual interpretation of sign language. Em *Computer Vision, European Conference on*. Springer-Verlag, 2004. (Citada nas páginas 2, 17, 24, e 28).

[15] A. P. A. Sousa. Interpretação da língua gestual portuguesa. Master's thesis, Universidade de Lisboa, Lisboa, 2012. (Citada nas páginas 2, 23, 24, 25, 28, 53, e 72).

[16] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati, e M. Hadap. Hand gesture recognition for Indian Sign Language. Em *Computer Communication and Informatics (ICCCI), 2012 International Conference on*, pages 1–4, Jan 2012. (Citada na página 2).

[17] J. Suarez e R. R. Murphy. Hand gesture recognition with depth images: a review. Em *Robots and Human Interactive Communications, 2012 IEEE International Workshop on*, pages 411–417, 2012. (Citada nas páginas 2, 6, 21, 23, 24, 38, e 49).

[18] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, e A. Fitzgibbon. KinectFusion: real-time dense surface mapping and tracking. Em *Mixed and Augmented Reality, Proceedings of the 2011 10th IEEE International Symposium on*, ISMAR '11, pages 127–136, Washington, DC, USA, 2011. IEEE Computer Society. (Citada nas páginas 2 e 12).

[19] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, e A. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. Em *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 559–568, New York, NY, USA, 2011. ACM. (Citada nas páginas 2 e 3).

[20] I. Oikonomidis, N. Kyriazis, e A. A. Argyros. Tracking the articulated motion of two strongly interacting hands. Em *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, Providence, Rhode Island, USA, 2012. (Citada na página 2).

[21] I. Oikonomidis, N. Kyriazis, e A. A. Argyros. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. Em *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2088 –2095, nov. 2011. (Citada na página 3).

[22] P. J. Besl e N. D. McKay. A method for registration of 3-D shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):239–256, 1992. (Citada nas páginas 3, 26, 29, 30, 38, e 39).

[23] P. Trindade, J. Lobo, e J. P. Barreto. Hand gesture recognition using color and depth images enhanced with hand angular pose data. Em *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, pages 71–76, 2012. (Citada nas páginas 3, 24, 26, 27, 28, 38, 41, 55, 59, e 70).

[24] Z. Li e R. Jarvis. Real time hand gesture recognition using a range camera. Em *Robotics and Automation, (ACRA). Australasian Conference on*, pages 529–534, 2009. (Citada nas páginas 3 e 26).

[25] S. Rusinkiewicz e M. Levoy. Efficient variants of the ICP algorithm. Em *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152, 2001. (Citada nas páginas ix, 3, 29, 30, 31, 32, 35, e 41).

[26] M Livingstone. *Vision and art: the biology of seeing*. Abrams, New York, 2008. (Citada na página 6).

[27] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1a edition, 2010. (Citada nas páginas 7, 18, e 27).

[28] L. Zhang, N. Snavely, B. Curless, e S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. Em *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*, pages 548–558. ACM, 2004. (Citada nas páginas 7 e 11).

[29] J. Sivic, C. Lawrence, e Z. R. Szeliski. Finding people in repeated shots of the same scene. Em *Proceedings of the 16th British Machine Vision Conference*, pages 909–918, 2006. (Citada na página 7).

[30] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, e A. Blake. Real-time human pose recognition in parts from single depth images. Em *Computer Vision and Pattern Recognition*, 2011. (Citada nas páginas 7, 12, 21, 23, e 28).

[31] M. Pollefeys. Tutorial on 3D modeling from images. Computer Vision (ECCV), European Conference on, 2000. (Citada nas páginas 7 e 10).

[32] E. Trucco e A. Verri. *Introductory techniques for 3-D computer vision*. Prentice Hall PTR, 1998. (Citada nas páginas 7 e 9).

[33] International Imaging Industry Association. Photography - Digital still cameras - Guidelines for reporting pixel-related specifications, 2004. Norma ANSI/I3A IT10.7000-2004. (Citada na página 8).

[34] Wikipedia. The Free Encyclopedia, 2014. (Citada nas páginas ix, 8, 15, e 16).

[35] R. C. Gonzalez e R. E. Woods. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001. (Citada na página 8).

[36] K. Jack. *Video demystified: a handbook for the digital engineer*. Newnes, Newton, MA, USA, 5a edition, 2007. (Citada na página 8).

[37] A. A. Argyros e M. I. A. Lourakis. Real-time tracking of multiple skin-colored objects with a possibly moving camera. Em *Computer Vision - ECCV 2004*, volume 3023 of *Lecture Notes in Computer Science*, pages 368–379. Springer Berlin Heidelberg, 2004. (Citada nas páginas x, 8, 23, 71, e 72).

[38] A. Johnson. *Spin-Images: a representation for 3-D surface matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 1997. (Citada nas páginas 9 e 29).

[39] I. Oikonomidis, N. Kyriazis, e A. Argyros. Efficient model-based 3D tracking of hand articulations using Kinect. Em *Proceedings of the British Machine Vision Conference*, pages 101.1–101.11. BMVA Press, 2011. (Citada nas páginas ix, x, 9, 18, 23, 71, e 72).

[40] S. Ghobadi, O. Loepprich, K. Hartmann, e O. Loffeld. Hand segmentation using 2D/3D Images. Em M. J. Cree, editor, *Image and Vision Computing New Zealand, 27th International Conference on*, pages 64–69. University of Waikato, 2007. (Citada nas páginas 9, 18, e 23).

[41] M. Van den Bergh e L. Van Gool. Combining RGB and ToF cameras for real-time 3D hand gesture interaction. Em *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 66–72, 2011. (Citada nas páginas 9, 28, e 72).

[42] B. L. Curless. *New methods for surface reconstruction from range images*. PhD thesis, Stanford University, Stanford, CA, USA, 1998. (Citada na página 10).

[43] K. A. Pulli. *Surface reconstruction and display from range and color data*. PhD thesis, University of Washington, 1997. Chairperson-Shapiro, Linda G. (Citada na página 10).

[44] K. Konolige e D. Beymer. *SRI small vision system: user's manual software version*. Videre Design, 4.4a edition, 2007. (Citada nas páginas 10 e 11).

[45] R. Zhang, P. Tsai, J. E. Cryer, e M. Shah. Shape from Shading: A Survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(8):690–706, 1999. (Citada na página 10).

[46] C. Hernández. *Stereo and silhouette fusion for 3D object modeling from uncalibrated images under circular motion*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 2004. (Citada nas páginas 10 e 11).

[47] Y. Aloimonos e A. Rosenfeld. Principles of computer vision. Em *Handbook of pattern recognition and image processing (vol. 2): computer vision*, pages 1–15. Academic Press, Inc., 1994. (Citada na página 10).

[48] P. Favaro. Shape from focus and defocus: convexity, quasiconvexity and defocus-invariant textures. Em *Computer Vision, 11th IEEE International Conference on*, pages 1–7, 2007. (Citada nas páginas 10 e 11).

[49] E. Mouaddib, J. Batlle, e J. Salvi. Recent progress in structured light in order to solve the correspondence problem in stereovision. *Robotics and Automation, 1997 IEEE International Conference on*, 1:130 –136 vol.1, 1997. (Citada na página 11).

[50] P. Cho, H. Anderson, R. Hatch, e P. Ramaswami. Real-time 3D ladar imaging. Em *Applied Imagery and Pattern Recognition Workshop, 2006. AIPR 2006. 35th IEEE*, page 5, 2006. (Citada na página 11).

[51] A. Soloviev e M. U. de Haag. Three-dimensional navigation with scanning ladars: concept and initial verification. *Aerospace and Electronic Systems, IEEE Transactions on*, 46(1):14 –31, 2010. (Citada na página 11).

[52] P. Vuylsteke, C. B. Price, e A. Oosterlinck. Image sensors for real-time 3D acquisition: part 1. Em *Traditional and non-traditional robotic sensors*, pages 187–210. Springer-Verlag New York, Inc., 1990. (Citada na página 11).

[53] S. S. Sinha e R. Jain. Range image analysis. Em *Handbook of pattern recognition and image processing (vol. 2): computer vision*, pages 185–237, Orlando, FL, USA, 1994. Academic Press, Inc. (Citada na página 12).

[54] C. Aguiar. *Segmentation and region-based registration applied to 3D raw point clouds. Application to cultural heritage*. PhD thesis, Université Montpellier II, 2009. (Citada na página 12).

[55] L. A. de Albuquerque. Alinhamento de imagens de profundidade na reconstrução 3D de objetos de forma livre. Master's thesis, Universidade de Brasília, 2006. (Citada na página 12).

[56] PrimeSense Incorporation. Prime sensor NiTE 1.3 algorithms notes, 2010. (Citada nas páginas 12, 13, e 49).

[57] Microsoft Corporation. Kinect for Windows SDK, 2012. (Citada na página 13).

[58] J. Tong, J. Zhou, L. Liu, Z. Pan, e H. Yan. Scanning 3D full human bodies using kinects. *Visualization and Computer Graphics, IEEE Transactions on*, 18(4):643–650, 2012. (Citada na página 13).

[59] OpenNI Software Development Kit. Online, 2014. (Citada nas páginas 13 e 49).

[60] E. C. Viotti. Introdução aos Estudos Linguísticos. Universidade Federal de Santa Catarina, Curso de Licenciatura em Letras-Libras, 2007. Desenvolvimento de material didático ou instrucional - Material didático. (Citada na página 14).

[61] J. B. Dias, K. P. Souza, e H. Pistori. Conjunto de treinamento para algoritmos de reconhecimento de Libras. Em *II Workshop de Visão Computacional*, São Carlos, 2006. (Citada nas páginas ix e 15).

[62] F. C. Capovilla e W. D. Raphael. *Enciclopédia da língua de sinais brasileiras: o mundo do surdo em libras*. Edusp, 2005. (Citada nas páginas ix, 14, e 15).

[63] K. Strobel. História da educação de surdos, 2009. Universidade Federal de Santa Catarina. Licenciatura em Letras-LIBRAS na modalidade a distância. (Citada na página 16).

[64] A. Baalbaki e B. Caldas. Impacto do congresso de Milão sobre a língua dos sinais. Em *Anais do XV Congresso Nacional de Linguística e Filologia*, 2011. (Citada na página 16).

[65] A. dos Santos Figueira. *Material de apoio para o aprendizado de Libras*. PHORTE Editora, 2011. (Citada na página 16).

[66] J. Esminger e J. E. G. Souza Junior. Curso básico de Libras: comunicando com as mãos, 2014. Associação dos Profissionais Tradutores – Intérpretes de Língua Brasileira de Sinais de Mato Grosso do Sul. (Citada nas páginas ix e 17).

[67] G. R. S. Murthy e R. S. Jadon. A review of vision based hand gestures recognition. *Information Technology and Knowledge Management, International Journal of*, 2:pp. 405–410, 2009. (Citada na página 18).

[68] P. Yan e K. W. Bowyer. A fast algorithm for ICP-based 3D shape biometrics. Em *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, pages 213–218, 2005. (Citada nas páginas 18, 24, 27, 28, 41, 59, e 72).

[69] T. H. Cormen, C. Stein, R. L. Rivest, e C. E. Leiserson. *Introduction to algorithms*. MIT Press and McGraw-Hill, 3a edition, 2009. (Citada nas páginas 18 e 19).

[70] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, e X. Twombly. Vision-based hand pose estimation: a review. *Computer Vision and Image Understanding*, 108(1-2):52–73, 2007. (Citada nas páginas ix, 20, 22, 23, e 24).

[71] Y. Zhu e K. Fujimura. A bayesian framework for human body pose tracking from depth image sequences. *Sensors*, 10(5):5280–5293, 2010. (Citada na página 21).

[72] S. B. Gokturk e C. Tomasi. 3D head tracking based on recognition and interpolation using a time-of-flight depth sensor. Em *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–211 – II–217 Vol.2, 2004. (Citada na página 21).

[73] N. Grammalidis, G. Goussis, G. Troufakos, e M. G. Strintzis. 3-D human body tracking from depth images using analysis by synthesis. Em *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pages 185 –188 vol.2, 2001. (Citada na página 21).

[74] C. Plagemann, V. Ganapathi, D. Koller, e S. Thrun. Real-time identification and localization of body parts from depth images. Em *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3108 –3113, 2010. (Citada na página 21).

[75] K. Fujimura e Xia Liu. Sign recognition using depth image streams. Em *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 381–386, 2006. (Citada nas páginas 22, 23, e 28).

[76] C. Keskin, F. Kirac, Y. E. Kara, e L. Akarun. Real time hand pose estimation using depth sensors. Em *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1228–1234, 2011. (Citada nas páginas 23, 24, 25, e 28).

[77] N. Pugeault e R. Bowden. Spelling it out: real-time ASL fingerspelling recognition. Em *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1114–1119, 2011. (Citada nas páginas ix, 23, 24, 25, 28, 56, e 72).

[78] Z. Ren, J. Yuan, e Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. Em *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 1093–1096, New York, NY, USA, 2011. ACM. (Citada nas páginas 23 e 26).

[79] D. Uebersax, J. Gall, M. Van den Bergh, e L. Van Gool. Real-time sign language letter and word recognition from depth data. Em *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 383–390, 2011. (Citada nas páginas 23, 24, e 28).

[80] B. B. Amor, M. Ardabilian, e C. Liming. New experiments on ICP-based 3D face recognition and authentication. Em *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1195–1199, 2006. (Citada nas páginas 24, 26, 28, 41, 59, e 72).

[81] M. V. Lamar, M. S. Bhuiyan, e A. Iwata. Hand gesture recognition using morphological principal component analysis and an improved CombNET-II. Em *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, volume 4, pages 57–62 vol.4, 1999. (Citada nas páginas 24 e 28).

[82] R. C. B. Madeo. Máquinas de vetores suporte e a análise de gestos : incorporando aspectos temporais. Master's thesis, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2013. (Citada na página 24).

[83] Xia Liu e K. Fujimura. Hand gesture recognition using depth data. Em *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 529–534, 2004. (Citada nas páginas 24, 26, e 28).

[84] S. Liwicki e M. Everingham. Automatic recognition of fingerspelled words in British Sign Language. *Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on*, 0:50–57, 2009. (Citada nas páginas 24 e 28).

[85] C. R. Souza, E. B. Pizzolato, e M. Santos Anjo. Fingerspelling recognition with support vector machines and hidden conditional random fields. Em *Advances in Artificial Intelligence – IBERAMIA 2012*, volume 7637 of *Lecture Notes in Computer Science*, pages 561–570. Springer Berlin Heidelberg, 2012. (Citada nas páginas 24, 25, e 28).

[86] K. R. Konda, A. Königs, H. Schulz, e D. Schulz. Real time interaction with mobile robots using hand gestures. Em *Human-Robot Interaction, Proceedings of the seventh annual ACM/IEEE international conference on*, HRI '12, pages 177–178, New York, NY, USA, 2012. ACM. (Citada na página 24).

[87] M. Van den Bergh, D. Carton, R. de Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, e M. Buss. Real-time 3D hand gesture interaction with a robot for understanding directions from humans. Em *Robots and Human Interactive Communications, 2011 IEEE*, pages 357–362, 2011. (Citada na página 24).

[88] The Univertsity of Waikato. WEKA. Online, 2014. (Citada na página 25).

[89] J. Gutmann, M. Fukuchi, e M. Fujita. 3D perception and environment map generation for humanoid robot navigation. *The International journal of robotics research*, 27(10):1117–1134, 2008. (Citada nas páginas 27 e 72).

[90] J.P. Silva Júnior. Visualização e alinhamento de vistas parciais com geometria rígida aplicados na reconstrução 3D de objetos de forma livre, 2010. Monografia (Bacharel em Ciência da Computação), Universidade de Brasília, Brasília, Brasil. (Citada nas páginas 28 e 40).

[91] A. Ardeshir Goshtasby. *2-D and 3-D image registration: for medical, remote sensing, and industrial applications*. Wiley-Interscience, 2005. (Citada na página 28).

[92] Y. Chen e G. Medioni. Object modelling by registration of multiple range images. Em *Robotics and Automation, IEEE International Conference on*, pages 2724–2729, 1991. (Citada nas páginas 29, 35, e 42).

[93] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal on Computer Vision*, 13(2):119–152, 1994. (Citada na página 29).

[94] M. Greenspan e G. Godin. A nearest neighbor method for efficient ICP. *3D Digital Imaging and Modeling, International Conference on*, 0:161, 2001. (Citada nas páginas 29 e 33).

[95] T. Zinsser, H. Schmidt, e J. Niermann. A refined ICP algorithm for robust 3-D correspondences estimation. Em *Image Processing, Proceedings of the International Conference on*, pages 695–698, 2003. (Citada na página 29).

[96] T. Masuda. Generation of geometric model by registration and integration of multiple range images. *3D Digital Imaging and Modeling, International Conference on*, 0:254, 2001. (Citada na página 29).

[97] C. K. Chow, H. Tat Tsui, e T. Lee. Surface registration using a dynamic genetic algorithm. *Pattern Recognition*, 37(1):105 – 117, 2004. (Citada na página 29).

[98] J. P. Silva Júnior, D. L. Borges, e F. de Barros Vidal. A dynamic approach for approximate pairwise alignment based on 4-points congruence sets of 3d points. Em *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 889–892, 2011. (Citada na página 29).

[99] Z. Zhang, Sim Heng Ong, e K. Foong. Improved spin images for 3D surface matching using signed angles. Em *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 537–540, 2012. (Citada na página 29).

[100] C. Chen, Y. Hung, e J. Cheng. RANSAC-based DARCES: a new approach to fast automatic registration of partially overlapping range images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(11):1229 –1234, Novembro de 1999. (Citada na página 29).

[101] J. Salvi, C. Matabosch, D. Fofi, e J. Forest. A review of recent range image registration methods with accuracy evaluation. *Image Vision Computing*, 25(5):578–596, 2007. (Citada nas páginas 30 e 35).

[102] G. Godin, M. Rioux, e R. Baribeau. Three-dimensional registration using range and intensity information. *Videometrics III*, 2350(1):279–290, 1994. (Citada na página 31).

[103] J. L. Bentley. Multidimensional divide-and-conquer. *Communications of the ACM*, 23(4):214–229, 1980. (Citada na página 33).

[104] T. Masuda, K. Sakaue, e N. Yokoya. Registration and integration of multiple range images for 3-d model construction. *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, 1:879 –883 vol.1, 1996. (Citada na página 34).

[105] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987. (Citada nas páginas 34 e 40).

[106] K. S. Arun, T. S. Huang, e S. D. Blostein. Least-squares fitting of two 3-D point sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 9(5):698–700, 1987. (Citada na página 34).

[107] S. Malassiotis, N. Aifanti, e M. G. Strintzis. A gesture recognition system using 3D data. Em *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pages 190–193, 2002. (Citada na página 47).