Reconhecimento de Sinais Estáticos a partir de Informação RGB-D usando um Descritor Kernel

Karla Catherine Otiniano Rodríguez
Universidade Federal de Ouro Preto

Dissertação submetida ao
Departamento de Ciência da Computação (DECOM)
Instituto de Ciências Exatas e Biológicas (ICEB)
Universidade Federal de Ouro Preto
para obtenção do título de Mestre em Ciência da Computação
Minas Gerais, Brasil

R696r Rodríguez, Karla Catherine Otiniano.
Reconhecimento de sinais estáticos a partir de informação RGB-D usando um Descritor Kernel [manuscrito] / Karla Catherine Otiniano Rodríguez – 2014.

79f.: il.; color.; grafs.; tabs.

Orientador: Prof. Guillermo Cámara Chávez.

Dissertação (Mestrado) - Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Departamento de Computação. Programa de Pós-graduação em Ciência da Computação.
Área de concentração: Recuperação e Tratamento da Informação.

1. Língua de sinais – Teses. 2. Profundidade - Percepção - Teses. 3. Descritores - Teses. 4. Sistemas de recuperação da informação - Teses. I. Cámara Chávez, Guillermo. II. Universidade Federal de Ouro Preto. III. Título.

CDU:004.424.5:621.397.4

Catalogação: sisbin@sisbin.ufop.br

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas – ICEB
Programa de Pós-Graduação em Ciência da Computação

decom
departamento de computação

Ata da Defesa Pública de Dissertação de Mestrado

Aos 06 dias do mês de setembro de 2013, às 10 horas na Sala de Seminários do DECOM no Instituto de Ciências Exatas e Biológicas (ICEB), reuniram-se os membros da banca examinadora composta pelos professores: **Prof. Dr. Guillermo Cámara Chávez (presidente e orientador), Prof. Dr. David Menotti Gomes, Prof. Dr. Frederico Gadelha Guimarães e Prof. Dr. William Robson Schwartz**, aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Computação, a fim de arguirem a mestranda **Karla Catherine Otiniano Rodriguez**, com o título **"Reconhecimento de Sinais Manuais (*finger-spelling*) Através de Informações RGB-D e Descritores de Kernels"**. Aberta a sessão pelo presidente, coube à candidata, na forma regimental, expor o tema de sua dissertação, dentro do tempo regulamentar, sendo em seguida questionado pelos membros da banca examinadora, tendo dado as explicações que foram necessárias.

Recomendações da Banca:

( ) Aprovada sem recomendações

( ) Reprovada

(X) Aprovada com recomendações: conforme sugestões da banca

Banca Examinadora:

Prof. Dr. Guillermo Cámara Chávez

Prof. Dr. David Menotti Gomes

Prof. Dr. Frederico Gadelha Guimarães

Prof. Dr. William Robson Schwartz

Prof. Dr. Haroldo Gambini Santos
Coordenador do Programa de Pós-Graduação em Ciência da Computação
DECOM/ICEB/UFOP

Ouro Preto, 06 de setembro de 2013.

Reconhecimento de Sinais Estáticos a partir de Informação RGB-D usando um Descritor Kernel

Resumo

Durante os últimos anos, têm sido desenvolvidas diversas abordagens para o reconhecimento de sinais. Muitas delas baseadas somente em informação de intensidade, o que tornava o pre-processamento mais complexo. Devido ao avanço da tecnologia, têm sido desenvolvidos novos dispositivos para a obtenção de informação mais complexa, além da informação de intensidade também é fornecida informação de profundidade e localização. O sensor Kinect é um deles e foi criado no ano 2010. Com esse dispositivo, é possível obter dois tipos de informações: intensidade e profundidade. Isso significa uma vantagem quando se quer desenvolver um modelo para reconhecimento de sinais. Como já foi dito, usar somente informação de intensidade, que é o mais usual, implica ter processos mais complexos e algumas vezes imprecisos. Para solucionar isso, é possível usar informação de profundidade que, além de conter informação da mão, facilita o pre-processamento. Nesta dissertação apresentamos um modelo para o reconhecimento de sinais estáticos, usando informação de intensidade e profundidade (RGB-D) de cada sinal. As imagens de intensidade oferecem informação visual do sinal. Enquanto que as imagens de profundidade permitem obter informação da forma da mão com a qual é executada o sinal. Além disso, usando este último tipo de imagem, o processo de segmentação é facilitado. Uma avaliação entre o descritor local SIFT e o descritor kernel gradiente foi realizada na etapa de extração de características. A partir das características obtidas, foi extraída informação semântica usando a técnica BoW (*Bag-of-Words*), para então finalmente classificar os sinais usando SVM (*Support Vector Machine*). Os resultados reportados nesta dissertação se mostraram superiores a outros modelos da literatura. Foi alcançado

um incremento na acurácia de 20%, sendo o melhor resultado de 95,63% de acurácia média, isso demonstra que o modelo proposto é promissor no reconhecimento de sinais.

Agradecimentos

Agradeço em primeiro lugar a Deus pela vida e por ter me dado a oportunidade de terminar com sucesso o mestrado. Além dele, muitas pessoas me apoiaram ao longo destes dois anos e gostaria agradecê-lhes.

Quero agradecer muito a meus pais Heli e Angélica pelo apoio incondicional que eles me deram sempre na minha vida, junto a meus irmãos James e Robert isto foi possível.

Tive a sorte de ter tido mais que um orientador, um amigo, que me apoiou sempre, Prof. Guillermo Cámara Chávez, graças a ele e sua esposa Yudy pela ajuda dada desde que cheguei ao Brasil.

Não posso deixar de agradecer a alguém que foi fundamental na realização desta dissertação, Edward Cayllahua, meu grande amor, sem seu apoio e companhia tudo teria sido mais difícil.

Também, gostaria de fazer menção a duas pessoas que me acompanharam neste tempo em Ouro Preto, minhas amigas Mara e Larissa, estou muito agradecida pelo apoio e carinho dado.

Finalmente, mas não menos importante, agradecer a meus avós Bertha e Roberto. Ela sempre orando por mim e ele me acompanhando desde o céu. É para vocês que eu dedico esta conquista.

Sumário

Lista de Figuras **xi**

Lista de Tabelas **xiii**

1 Introdução **1**

1.1 Introdução 1

1.2 Motivação 5

1.3 Objetivos 7

1.3.1 Objetivo geral 7

1.3.2 Objetivos específicos 8

1.4 Contribuições 8

1.5 Organização da Dissertação 9

2 Estado da Arte **11**

2.1 Introdução 11

2.2 Métodos baseados em Imagens 12

2.3 Métodos baseados em Dispositivos 14

2.4 Métodos Híbridos 15

2.5 Considerações Finais 17

3 Fundamentos Teóricos **19**

3.1 Língua de Sinais 19

3.1.1 Sinais 19

3.2 Aquisição de dados 22

3.2.1 Luvas 22

3.2.2 Kinect 22

3.3 Extração de Características 26

3.3.1 *Scale-Invariant Feature Transform (SIFT)* 26

3.3.2 *Gradient Kernel Descriptor* 29

3.3.3 Filtros de Gabor 31

3.4 *Algoritmos de aprendizagem máquina* 31

3.4.1 Clustering 32

3.4.2 Classificação 33

3.5 *Bag-of-Visual-Words* 37

3.6 Considerações Finais 38

4 Modelo Proposto **41**

4.1 Modelo proposto para Reconhecimento de Sinais 41

4.2 Segmentação 42

4.3 *Bag-of-Visual-Words* 44

4.4 Classificação 46

4.5 Considerações Finais 46

5 Experimentos e Resultados **47**

5.1 Base de dados 47

5.2 Definição de parâmetros 49

5.3 Avaliação do modelo 50
5.4 Experimentos 50
5.4.1 Resultados usando características de intensidade 51
5.4.2 Resultados usando características de profundidade 52
5.4.3 Resultados usando características RGB-D 53
5.5 Análise de Resultados 57
5.6 Considerações Finais 58

6 Conclusões **59**
6.1 Trabalhos Futuros 60

Referências Bibliográficas **61**

Lista de Figuras

1.1 Exemplos dos cinco parâmetros dos sinais (Felipe & Monteiro 2007).. . . 3

1.2 Alfabeto ASL: Sinais estáticos e dinâmicos (William Vicars 97-2013a). . . 4

1.3 Alfabeto LIBRAS: Sinais estáticos e dinâmicos (Espaço das LIBRAS 2011). 5

1.4 Tradutores de sinais: Em uma conferência e em um Jornal na TV. 7

1.5 Criança surda aprendendo a língua de sinais. 7

3.1 Exemplo de gestos usando a forma do sinal S (Felipe & Monteiro 2007). . 20

3.2 Exemplo de orientações da mão (Felipe & Monteiro 2007). 20

3.3 Exemplo de sinais com distintas localizações (Felipe & Monteiro 2007). . 21

3.4 Exemplo de sinais estáticos e dinâmicos (Felipe & Monteiro 2007). 21

3.5 Exemplo de expressões faciais (Felipe & Monteiro 2007). 21

3.6 Luvas coloridas. (a) Palma da luva dominante. (b) Parte de trás da luva dominante. (c) Palma da luva não-dominante. (d) Parte de trás da luva não-dominante. 23

3.7 Luvas com sensores. (a)Luvas com acelerômetros (b) Detalhe dos acelerômetros montados no pulso. 23

3.8 O Kinect e seus componentes. 24

3.9 Kinect para XBOX 360. 24

3.10 Exemplo das imagens obtidas pelo Kinect. Esquerda: imagem RGB. Direita: imagem de profundidade. 25

3.11 Pontos de articulação do corpo obtidos pelo Kinect. 25

3.12 Descritor kernel gradiente. Kernel de casamento aplicado sobre regiões da imagem. 30

3.13 (a) Componente real do filtro bidimensional de Gabor no domínio espacial; (b) Resposta de frequência do filtro definido em a. 31

3.14 Hiperplano com margem máximo e margens para uma SVM treinada com amostras de duas classes. Amostras na margem são chamadas de vetores de suporte. 34

3.15 Máquinas de kernel são usados para calcular uma série de funções não-linearmente separáveis em uma função linearmente separável de maior dimensão. 35

3.16 Exemplo da execução de *Random Forests*. 36

3.17 Bag-of-Visual-Words baseado em características visuais. 38

3.18 Exemplo do modelo Bag of visual words. 39

4.1 Modelo proposto para o reconhecimento de sinais estáticos. 43

4.2 Exemplo de imagens de (a) intensidade e (b) profundidade para um mesmo sinal. 44

4.3 Exemplo de imagens de (a) intensidade (b) profundidade após a segmentação. 44

4.4 Exemplo das (a) magnitudes e (b) orientações dos gradientes obtidos pelo descritor kernel gradiente aplicadas em uma imagem de intensidade. . . . 45

4.5 Exemplo das (a) magnitudes e (b) orientações dos gradientes obtidos pelo descritor kernel gradiente aplicadas em uma imagem de profundidade. . . 45

5.1 ASL Finger Spelling Dataset: 24 sinais estáticos a partir de 5 usuários. Exemplo da variedade do conjunto de dados. Esta matriz mostra uma imagem a partir de cada usuário e de cada letra. 48

5.2 Sinal m com variações de orientação e fundo. 48

5.3 Sinais mais conflituosos na base de dados. 49

5.4 Exemplo dos pontos característicos de uma imagem de intensidade obtidos pelo descritor SIFT. 52

Lista de Tabelas

5.1 Matriz de confusão da classificação dos 24 sinais a partir de informação de intensidade usando o descritor SIFT. 51

5.2 Matriz de confusão da classificação dos 24 sinais a partir de informação de intensidade usando o descritor kernel gradiente. 53

5.3 Matriz de confusão da classificação dos 24 sinais a partir de informação de profundidade usando o descritor kernel gradiente. 54

5.4 Matriz de confusão da classificação dos 24 sinais a partir de informação RGB-D usando os descritores SIFT e kernel gradiente. 55

5.5 Matriz de confusão da classificação dos 24 sinais a partir de informação RGB-D usando o descritor kernel gradiente. 56

5.6 Acurácia e desvio padrão da classificação usando informação RGB-D com diferentes quantidades de treinamento e teste. 57

5.7 Acurácia e desvio padrão de cada um dos 5 experimentos. 58

Capítulo 1

Introdução

Neste capítulo serão apresentadas a introdução e motivação desta dissertação, além dos objetivos perseguidos e contribuições alcançadas.

1.1 Introdução

É bem conhecido que, pessoas surdas têm problemas de comunicação com as demais pessoas (pessoas ouvintes). Do mesmo modo, as demais pessoas são incapazes de estabelecer uma comunicação com eles ou conseguir entender o que eles tentam dizer. Assim, como em cada país existe um idioma definido para falar, os surdos também têm sua própria língua, conhecida como Língua de Sinais. Há no mundo muitas línguas de sinais usadas como forma de comunicação entre pessoas surdas ou com problemas auditivos. Muitas dessas línguas têm reconhecimento oficial em vários países. As mais conhecidas são ASL (*American Sign Language*, Língua de Sinais Americana), BSL (*British Sign Language*), Auslan (*Australian Sign Language*) e LIBRAS (Língua Brasileira de Sinais).

A língua de sinais é uma forma complexa de comunicação em que as mãos, braços, cabeça, expressão facial e linguagem corporal são utilizados para comunicar uma língua visual-espacial sem som. Portanto, é possível definir alguns parâmetros para um sinal, os quais são cinco: (i) ***Forma***, que indica o formato da mão, (ii) ***Orientação***, que indica para onde os dedos estão apontando, (iii) ***Localização***, que indica a posição da mão em relação ao corpo, e (iv) ***Movimento***, é a mudança no tempo dos três parâmetros anteriores, e finalmente (v) ***Expressões faciais*** que representam a entonação da palavra. Na Figura 1.1, são encontrados exemplos de cada um destes parâmetros. Contudo,

todos os sinais não contêm as características dos cinco parâmetros. É por isso que na língua de sinais são encontrados dois tipos de sinais: estáticos e dinâmicos. Os sinais estáticos são os que representam as letras do alfabeto e os números. Existem algumas exceções, onde a letra é representada com movimentos tornando-se um sinal dinâmico. Os sinais dinâmicos são baseados no movimento da mão, onde varia a forma, orientação e localização da mão no tempo. Esses sinais são os que representam as palavras que logo conformam as frases.

O alfabeto em ASL é um exemplo de um conjunto de sinais estáticos e dinâmicos, realizados com uma única mão. Na Figura 1.2 pode ser visto um exemplo de cada sinal, onde as letras J e Z têm movimento, representando sinais dinâmicos. As demais 24 letras formam parte dos sinais estáticos. Esses sinais estáticos são os que usaremos nesta dissertação. Outro exemplo pode ser visto na Figura 1.3, onde é mostrado o alfabeto da LIBRAS. Deste exemplo, é possível notar a diferença do alfabeto LIBRAS com o alfabeto ASL. LIBRAS contém no seu alfabeto mais sinais dinâmicos, os quais são: *h*, *j*, *k*, *x*, *y* e *z*. Esta informação é obtida do site de LIBRAS: http://www.libras.org.br/.

Diferentes modelos foram propostos para o reconhecimento de sinais, baseados em um tipo de sinal (estático ou dinâmico). Para sinais estáticos são encontrados os modelos de Vamplew (1996), Pizzolato, dos Santos Anjo & Pedroso (2010), De Souza, Pizzolato & dos Santos Anjo (2012), Ren, Yuan & Zhang (2011), Pugeault & Bowden (2011), entre outros. Quando trata-se de sinais dinâmicos, temos os modelos Piater, Hoyoux & Du (2010), Biswas & Basu (2011), Chang, Chen & Huang (2011), Hienz, Bauer & Karl-Friedrich (2000), etc.

Na atualidade, devido ao avanço nos dispositivos usados na aquisição da informação, é possível dividir os modelos de acordo com a forma da obtenção de informação. Assim, a principal forma de obtenção de dados são as câmeras RGB. Esse tipo de informação é a mais comum e a que contém informação visual do sinal realizado. Como já foi dito, uma língua de sinais tem informação visual-geométrica, com as câmeras RGB são obtidas as informações visuais do sinal. Para a informação geométrica, são geralmente usadas luvas de dados, com as quais é obtida informação das posições das mãos. Contudo, no ano 2010, foi desenvolvido um dispositivo chamado Kinect (©2013 Microsoft Corporation 2013), o qual, devido a sua camera e aos dois sensores infravermelhos que possui na sua estrutura, obtém tanto informação visual quanto geométrica.

Diversos métodos para o reconhecimento de sinais foram propostos. Cada um deles focado em um tipo de informação de acordo com a aquisição dos dados. Desse modo,

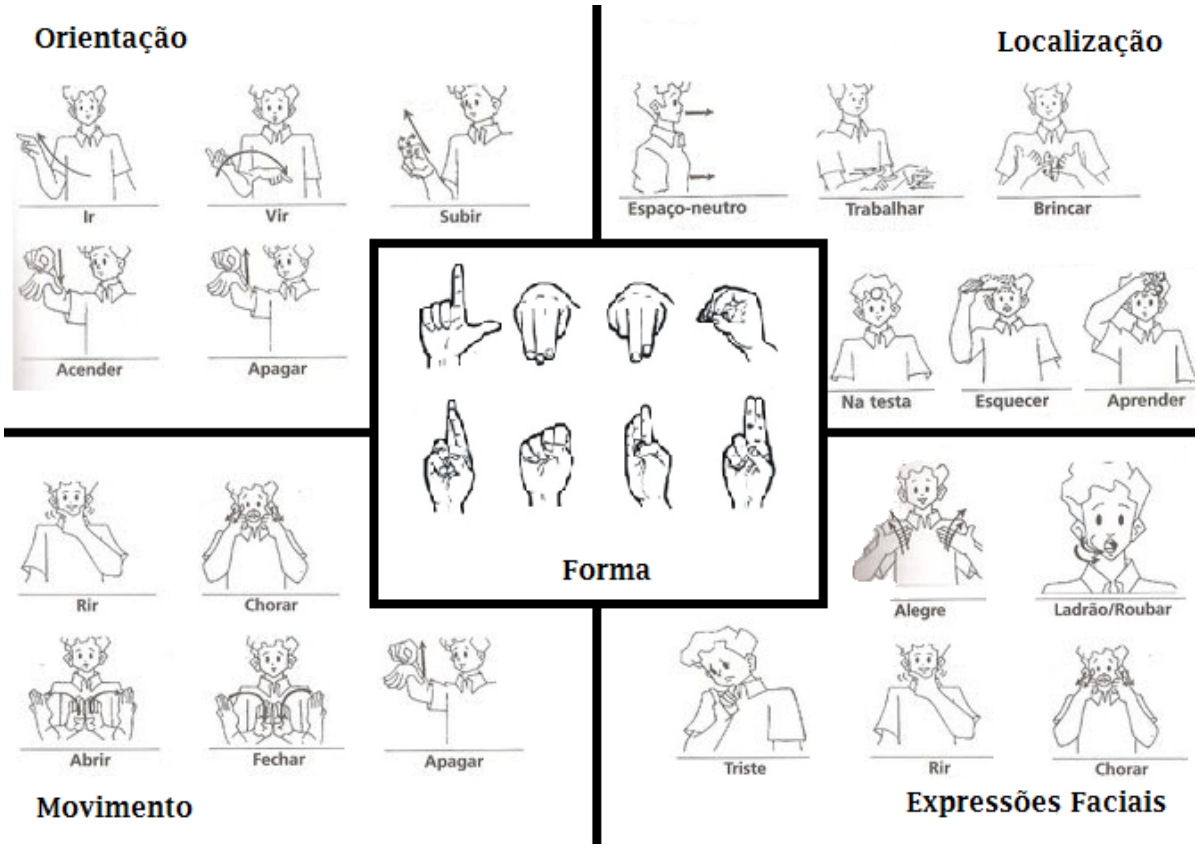

Figura 1.1: Exemplos dos cinco parâmetros dos sinais (Felipe & Monteiro 2007)..

esses métodos podem ser classificados em (a) Métodos baseados em dispositivos, os quais usam luvas de dados ou rastreadores para detectar a forma da mão e os movimentos do corpo, (b) Métodos baseados em imagens, usam câmeras simples para obter informação visual da mão. Processamento de imagens e extração de características são usados para capturar a forma e os movimentos da mão. e (c) Métodos Híbridos, os quais usam uma combinação dos dois métodos anteriores. Isto é, usar uma câmera e um outro sensor, que por exemplo pode ser um sensor infravermelho, como é o caso do Kinect.

Métodos baseados em dispositivos (sensores), tais como luvas de dados, podem fornecer medições precisas de mãos e movimento. Infelizmente, estes métodos requerem calibragem, eles também restringem o movimento natural das mãos e são muitas vezes de alto custo econômico. Métodos baseados em imagens são menos intrusivos, contudo, novos problemas surgem: localizar as mãos e segmentá-las não é uma tarefa trivial. Recentemente, câmeras de profundidade tornaram-se populares por um preço acessível. A informação de profundidade torna a tarefa de segmentar a mão do fundo muito mais

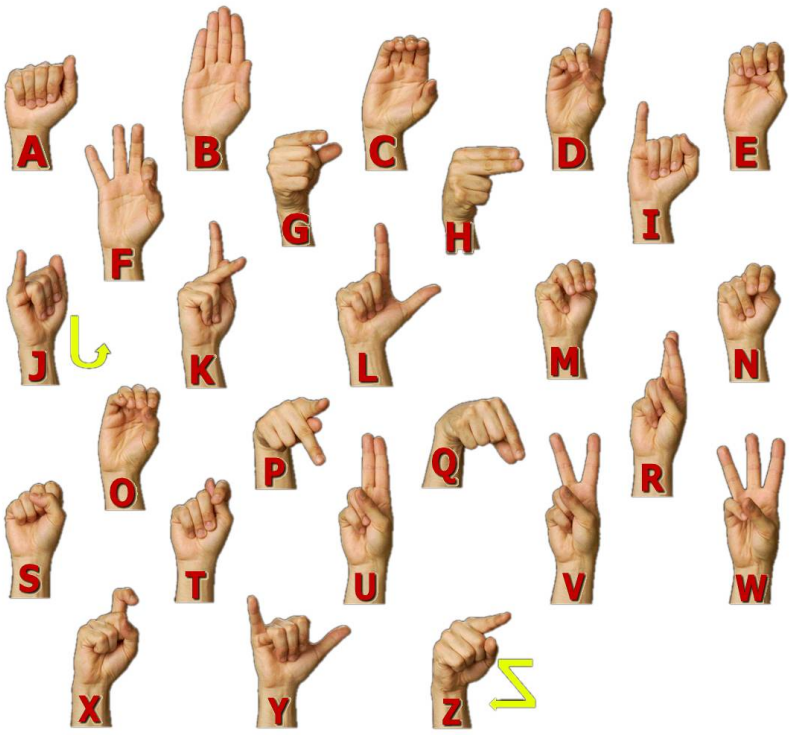

Figura 1.2: Alfabeto ASL: Sinais estáticos e dinâmicos (William Vicars 97-2013a).

fácil. Por esse motivo, a informação de profundidade pode ser usada para melhorar o processo de segmentação, como usado nos modelos (Ren, Yuan & Zhang 2011, Frati & Prattichizzo 2011, Li 2012, Mo & Neumann 2006).

As câmeras de profundidade obtiveram um grande interesse na comunidade de visão computacional devido ao seu sucesso em muitas aplicações, tais como estimativa de pose (Fanelli, Gall & Gool 2011, Shotton, Sharp, Kipman, Fitzgibbon, Finocchio, Blake, Cook & Moore 2013) rastreamento (Oikonomidis, Kyriazis & Argyros 2011), reconhecimento de objetos (Oikonomidis, Kyriazis & Argyros 2011), etc. Câmeras de profundidade também foram utilizados para reconhecimento de gestos de mão (Pugeault & Bowden 2011, Uebersax, Gall, den Bergh & Gool 2011, dos Santos Anjo, Pizzolato & Feuerstack 2012). Uebersax, Gall, den Bergh & Gool (2011) apresentam um sistema de reconhecimento de letras e palavras descritas com os dedos (*finger spelled*). Pugeault & Bowden (2011) usam o sensor Kinect da Microsoft para coletar imagens RGB e de profundidade. Eles extraem as características usando Filtros Gabor e, em seguida, uma Random Forest prediz as letras da Língua Americana de Sinais (*American Sign Language*, ASL).

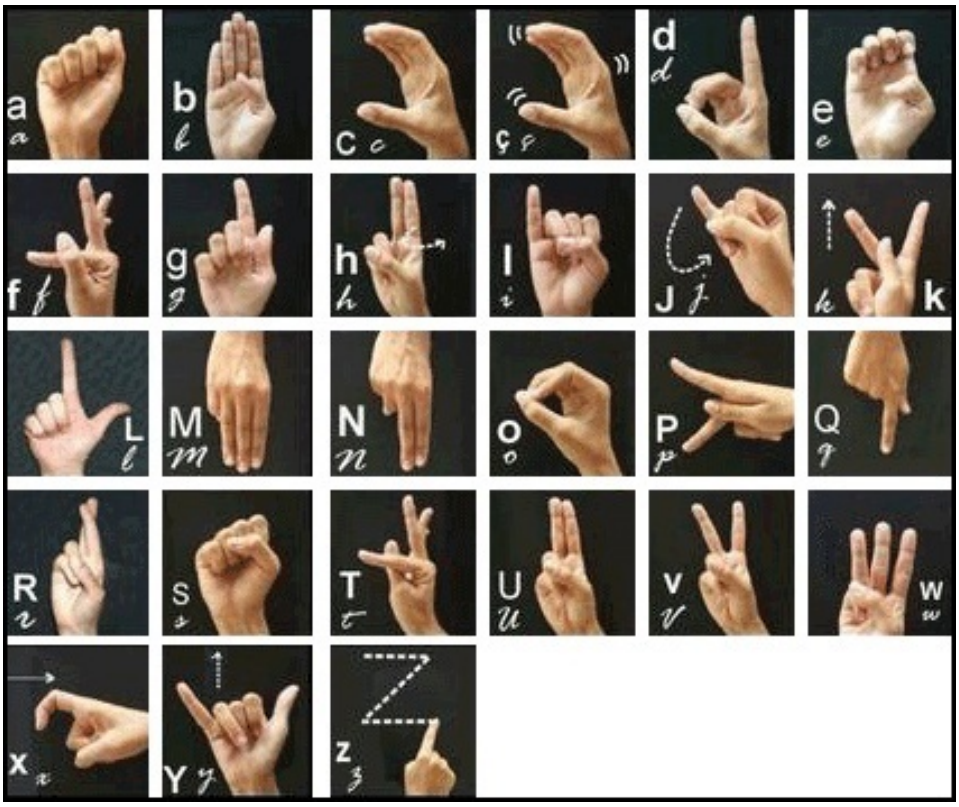

Figura 1.3: Alfabeto LIBRAS: Sinais estáticos e dinâmicos (Espaço das LIBRAS 2011).

1.2 Motivação

O reconhecimento de língua de sinais é um tema relativamente novo e em atual desenvolvimento. Reconhecer língua de sinais é uma tarefa importante para facilitar a comunicação entre pessoas falantes da língua de sinais e não falantes. Esse problema é significativo devido ao fato de que existem 360 milhões de pessoas no mundo com problemas de audição ou totalmente surdos (Lütkebohle 2013).

Nem todos os surdos tiveram sempre essas deficiências. A surdez pode ser devido a uma infecção no ouvido, doença tal como a meningite, uma lesão ou um problema hereditário. Porém, existem pessoas que nascem surdos, algumas crianças são identificadas ao nascimento ou nos primeiros meses de vida e no pior dos casos é possível chegar a 1 ou 2 anos. Nos primeiros anos de vida, a criança desenvolve a linguagem muito rapidamente. Portanto, se uma criança tem um problema de audição, é importante reconhecê-lo o quanto antes possível para que ela possa receber uma ajuda eficaz. Caso contrário, a criança não consegue aproveitar dos melhores anos para desenvolver

as habilidades necessárias para a comunicação (do nascimento aos 7 anos de idade). Quanto mais cedo a criança começar a desenvolver a língua e se comunicar, mais poderá aprender. A criança surda de pais surdos tem melhor desenvolvimento na aprendizagem da língua de sinais que as outras crianças de pais ouvintes.

Os problemas não somente estão restritos às limitações da aprendizagem da língua, eles também terão que lidar com os problemas da vida cotidiana. Por exemplo, quando eles vão ao supermercado, lojas, e o problema piora quando eles têm emergências, como ir ao hospital ou polícia por exemplo. Outro tipo de problemas, tem a ver com a limitação da obtenção da informação. Por exemplo, quando eles querem assistir à TV, eles somente conseguem obter informação visual sem conseguir capturar informação auditiva. Existem programas de televisão, como alguns jornais, onde existem pessoas que traduzem os diálogos à língua de sinais, mas lamentavelmente existem poucos programas que oferecem esse tipo de ajuda (ver Figura 1.4). Devido a isso, os surdos estão limitados a se informar, divertir ou relaxar assistindo também outro tipo de programas (esporte, novelas, entretenimento, etc.).

Esta dissertação está motivada pelos problemas descritos nos parágrafos anteriores. Com o reconhecimento de sinais é possível ajudar na comunicação entre pessoas surdas e ouvintes. Ajudar às crianças que nasceram surdas na aprendizagem da língua de sinais, ou no caso de pais ouvintes a eles também aprenderem a língua para uma melhor comunicação com seus filhos.

Outra das motivações é que o reconhecimento de sinais é um tema em atual desenvolvimento. As formas de aquisição de informação vão variando e é possível obter mais tipos de informação. Deste modo, os métodos propostos por diversos autores usam também mais informação, tentando incrementar a taxa de reconhecimento da língua de sinais. Os métodos mais básicos usam informação somente de imagens, seja em cores ou em tons de cinza, tornando o método mais complexo. Outro tipo de imagens são as imagens de profundidade, que podem ser obtidas pelo sensor Kinect ou pela câmera ToF. Este tipo de imagens fornecem informação 3D do corpo ou especificamente da mão que realiza o sinal. Usando o sensor Kinect e aplicando um software, é possível obter informação espacial a partir destas imagens. Com estes três tipos de informação: imagens de intensidade, de profundidade e informação espacial (obtida pelo software), os métodos se tornam mais robustos pela quantidade de informação usada.

Figura 1.4: Tradutores de sinais: Em uma conferência e em um Jornal na TV.

Figura 1.5: Criança surda aprendendo a língua de sinais.

1.3 Objetivos

Nesta seção, o objetivo geral e os objetivos específicos perseguidos nesta dissertação serão apresentados.

1.3.1 Objetivo geral

Propor um método para o reconhecimento de sinais estáticos a partir de informação RGB-D usando um descritor kernel.

1.3.2 Objetivos específicos

1. Analisar a robustez do descritor de características locais em comparação de um descritor kernel.

2. Avaliar a importância da segmentação de imagens.

3. Avaliar a importância da utilização de informação de intensidade e profundidade no reconhecimento de sinais.

4. Avaliar os resultados a partir dos resultados obtidos por outros modelos.

1.4 Contribuições

Nos últimos anos têm sido propostos diversos modelos para o reconhecimento de sinais. Muitos deles usando unicamente informação de intensidade, tornando difícil o pre-processamento como é o caso da segmentação. Outros modelos preferiam evitar este tipo de problemas usando dispositivos extras, como as luvas coloridas. Também eram usadas luvas de dados quando se tratava de sinais dinâmicos e era necessário obter informação das posições das mãos. Porém, este tipo de dispositivo limita a pessoa no momento de realizar o sinal.

As contribuições deste trabalho são as seguintes:

- A principal contribuição desta dissertação é apresentar um método para o reconhecimento de sinais estáticos usando informação de intensidade e de profundidade (RGB-D). Este método foi publicado no SIBGRAPI 2013 (Otiniano-Rodríguez & Cámara-Chávez 2013).

- O uso da informação de profundidade facilita o processo de segmentação, permitindo remover informações não relevantes, como as informações do fundo da imagem e partes do corpo que não estão envolvidos na realização do sinal.

- A utilização de um descritor kernel incrementa a taxa de reconhecimento,pois ele extrai informação de nível médio.

- Outra contribuição, é a busca de informação semântica a partir das características obtidas, tornando-as em características mais robustas.

1.5 Organização da Dissertação

Esta dissertação está organizada da seguinte forma. No Capítulo 2, apresentamos o estado da arte do reconhecimento de sinais. Logo, no Capítulo 3, apresentamos uma definição básica de língua de sinais, formas de aquisição de informação e técnicas importantes usadas em reconhecimento de sinais. No Capítulo 4, é explicado o modelo proposto para o reconhecimento de sinais estáticos. Em seguida, são mostrados os resultados obtidos por esta dissertação e uma avaliação dos mesmos no Capítulo 5. Finalmente, no Capítulo 6, as conclusões e trabalhos futuros são apresentados.

Capítulo 2

Estado da Arte

O reconhecimento de língua de sinais é uma área extensa dividida em duas categorias: reconhecimento aplicado em imagens estáticas que representam letras do alfabeto, também chamado de *fingerspelling*, e reconhecimento sobre imagens dinâmicas onde os sinais representam palavras ou frases. Diversos modelos têm sido desenvolvidos para reconhecer ambos tipos de sinais. Neste capítulo é apresentado uma revisão das diversas abordagens que foram propostas para reconhecimento de língua de sinais.

2.1 Introdução

A maioria dos métodos de reconhecimento de língua de sinais são aplicados sobre imagens com sinais estáticos. Geralmente essas imagens foram obtidas usando uma câmera simples em um local com fundo uniforme ou variado. É assim que existem diversos bancos de imagens para testar as novas abordagens propostas, sendo esta uma grande vantagem. Mas o problema com esse tipo de dados é a limitação da informação obtida. Por exemplo, em uma imagem onde uma pessoa realiza algum sinal, à primeira vista pareceria ser uma tarefa simples localizar a mão da pessoa mas, porém trata-se de uma tarefa complexa. Para resolver esse problema são usados dispositivos, tais como luvas ou rastreadores de posição, que facilitam o processamento das imagens e o rastreamento da mão. No caso das luvas, ajuda na extração das informações da mão facilitando a segmentação (a luva pode ter uma cor diferente do resto da cena). Usando o rastreador é possível obter a rota percorrida pela mão no caso de sinais dinâmicos. Contudo, usar esses dispositivos limita a execução do sinal, reduzindo o espaço de ação. Nos últimos anos,

estão sendo utilizados novos dispositivos, entre eles, as câmeras ToF (*Time-of-flight*), câmeras estereoscópicas e o mais popular é o sensor Kinect da Microsoft. Estes dispositivos, além de capturar informações de cor, têm a vantagem de obter informação da profundidade ou representar uma imagem 2D em uma 3D. No caso do Kinect, é possível obter uma imagem RGB e uma imagem de profundidade, facilitando assim a obtenção e processamento dos dados. A seguir apresentaremos alguns dos métodos existentes na literatura, categorizando-os de acordo com a forma de obtenção de dados.

2.2 Métodos baseados em Imagens

Uma imagem digital é uma representação bidimensional de uma cena na qual podem existir diferentes objetos. Em nosso caso, o "objeto" de interesse na cena é a mão da pessoa que executa o sinal, mas é possível encontrar diferentes tipos de objetos, como outras partes do corpo da pessoa (ombro, cabeça, braço, etc.) ou objetos no fundo da cena (mesa, cadeira, outras pessoas, etc.). É assim que cada problema em reconhecimento de sinais torna-se único pela complexidade encontrada nas imagens usadas. Os passos comumente utilizados para reconhecimento são os seguintes: segmentação, extração de características, e classificação.

A efeito de simplificar o problema de segmentação, alguns autores utilizam fundos uniformes e roupas de uma mesma cor como é o caso de (Hienz, Bauer & Karl-Friedrich 2000, Holden & Owens 2001), mas as imagens no mundo real apresentam fundos complexos. Existem diferentes abordagens que tratam o problema de segmentação em imagens com fundo complexo. A solução é conseguida através de duas técnicas. A primeira é baseada na cor da pele e a segunda é usando alguma ajuda externa (dispositivos). O modelo apresentado em (Cooper & Bowden 2007) aprende a cor da pele de um modelo gaussiano da região do rosto para conseguir detectar as mãos. Piater, Hoyoux & Du (2010) apresentam um sistema de rastreamento da mão com base na segmentação da região da cor da pele seguido pelo casamento dos padrões baseado em PCA (*Principal Component Analysis*). Porém, usar a cor da pele como atributo não é tão simples. Essa técnica pode não dar bons resultados se na cena existem variações de iluminação ou variações de cores no fundo, além da ampla variedade de cores de pele. Em (Von Agris, Knorr & Kraiss 2008) é usado um modelo que além de trabalhar com a cor da pele adiciona conhecimentos de alto nível do corpo humano para detectar e segmentar as mãos. Uma outra abordagem é apresentada em (Starner, Weaver & Pentland 1998), onde o

algoritmo explora a imagem de entrada até encontrar um pixel com a mesma cor da pele previamente aprendida a partir de um modelo. Posteriormente, operações de dilatação morfológica são executadas a partir do pixel detectado.

Após a imagem ter sido segmentada é preciso extrair características que permitam descrevê-la. Em (Bo, Ren & Fox 2010) são apresentados três descritores kernel, *Gradient match kernel*, *Color match kernel* e *Shape match kernel*, que têm a vantagem de serem aplicados diretamente em imagens com fundo complexo e obter bons resultados. Cooper & Bowden (2007) segmentam a imagem com o intuito de detectar a face e as mãos da pessoa. Para a extração de características são usados conjuntamente quatro tipos de Momentos que oferecem diferentes propriedades: (a) ***Momentos Espaciais*** obtêm informação da posição, (b) ***Momentos Centrais*** que são invariantes à posição devido a serem calculados baseando-se em um centro de gravidade, (c) ***Momentos Centrais*** Normalizados são invariantes à posição e tamanho, e (d) ***Momentos de Hu*** que oferecem invariância à rotação, escala e posição. Em (Von Agris, Knorr & Kraiss 2008) é apresentado outro modelo que extrai características das mãos e da face. As características da mão são calculadas a partir da borda. Para a extração de características da face, um modelo de aparência ativo (Active Appearance Model, AMM) é aplicado para identificar as áreas de interesse, tais como os olhos e região da boca. Em seguida uma descrição numérica da expressão facial e do contorno dos lábios são calculados.

A eleição do tipo de classificador a ser usado no reconhecimento de sinais depende do tipo de sinal. No caso de sinais estáticos é possível usar Redes Neuronais Artificiais (*Artificial Neural Network*, ANN) ou Máquinas de vetores de suporte (*Support Vector Machine*, SVM). Vamplew (1996) cria um sistema chamado SLARTI (*Sign Language Recognition*). O sistema tem uma arquitetura modular que envolve múltiplas ANN e o classificador dos K-vizinhos mais próximos para reconhecer gestos da língua de sinais australiana. As características definidas da mão foram: a forma, a orientação, a localização e o movimento. Cada ANN é usada para classificar cada uma das características. O classificador dos K-Vizinhos mais próximos foi usado como classificador final usando as saídas das ANNs como sua entrada. Em (Pizzolato, dos Santos Anjo & Pedroso 2010) é apresentado uma arquitetura de dois níveis, usando também ANNs. No primeiro nível, os sinais com posturas manuais semelhantes foram agrupados para realizar uma classificação preliminar. No segundo nível, uma ANN foi aplicada para diferenciar os sinais agrupados no nível anterior (por exemplo, os sinais das letras T e F são muito semelhantes). O vetor de características usado no segundo nível é diferente do usado no primeiro nível. Baseados neste modelo, em (De Souza, Pizzolato & dos San-

tos Anjo 2012) fizeram uma comparação usando SVM para classificar sinais estáticos, os resultados mostraram melhores resultados do SVM frente à ANN.

Para o caso de sinais dinâmicos o mais usado é *Hidden Markov Model*(HMM). HMM tem a propriedade de compensar as variações de tempo e amplitude dos sinais e tem sido comprovados para reconhecimento de fala e de caracteres. Devido a estas características, as HMMs aparecem como uma abordagem ideal para o reconhecimento de língua de sinais. Como a fala, a língua de sinais pode ser considerada um sinal de tempo não-determinístico. Portanto, HMMs são usados para dados que contenham informações temporais e são conhecidos por terem taxas elevadas de classificação. Em (Liang & Ouhyoung 1998) é usado um modelo de Markov linear de dez estados para o reconhecimento da língua de sinais de Taiwan. Os modelos baseados em HMM precisam de um treinamento intensivo para extrair as transições temporais para a classificação. Este processo requer grandes quantidades de dados para representar a variação de eventos e descrição de características. As Cadeias de Markov podem ser utilizadas para representar essas transições temporais, evitando as altas exigências dos HMM. Cooper & Bowden (2007) utilizam Cadeias de Markov para representar as transições temporais que indicam a presença de um sinal. O objetivo é calcular a cadeia que melhor descreve os dados de entrada. Ye, Yao & Jiang (2004) apresentaram para o reconhecimento de língua de sinais chinês (*Chinese Sign Language Recognition*, CSLR). O método faz uso das vantagens tanto dos HMMs e das SVMs. Como foi mencionado, os HMMs são mais adequados para sinais dinâmicos, enquanto SVM mostra desempenho superior na classificação de sinais estáticos, com boas propriedades de generalização, especialmente para amostras limitadas.

2.3 Métodos baseados em Dispositivos

Uma solução para facilitar a segmentação ou rastreamento da mão é usar dispositivos externos. As luvas coloridas ou com sensores são usadas na obtenção de dados. As luvas colorias possuem cores distintas para identificar partes diferentes da mão ou para diferenciar uma mão da outra. Uma luva com sensores consiste de um rastreador que identifica informações de localização, orientação, velocidade e direção da mão de acordo a um ponto de referência, e de sensores que permitem identificar as mãos.

Na língua de sinais, a maioria das vezes, a mão direita é quem executa o sinal. Isto é devido a que as pessoas geralmente são destras. Em (Hienz, Bauer & Karl-

Friedrich 2000, Wang, Chen, Zhang, Wang & Gao 2007) são usadas duas luvas coloridas para rastrear os pontos característicos da mão e descrevê-los mais elaboradamente. As luvas são de cores diferentes, sendo que para a mão direita foi usada uma luva com 7 cores diferentes representando os cinco dedos, a palma e a parte traseira da mão. Já, para a mão esquerda foi usada uma luva de uma única cor. Em (Hienz, Bauer & Karl-Friedrich 2000) é usado o mesmo tipo de luvas, utilizando um algoritmo limiar são obtidas áreas do fundo, corpo, pele e das luvas. Por cada uma destas áreas coloridas são calculados centros de gravidade (CG). Um vetor de características é construído com as informações obtidas das luvas para logo ser usadas no classificador HMM. Essas características contêm informações principalmente das posições das mãos relativas ao centro do corpo, distâncias entre os CGs da mão direita e esquerda, além do tamanho de cada área colorida.

A abordagem apresentada em (Liang & Ouhyoung 1998) usa uma luva com sensores como dispositivo de entrada sobre um vocabulário de 250 sinais. Para a fase de reconhecimento foi utilizado uma HMM. Uma aplicação diferente é encontrada em (Brashear, Henderson, Park, Hamilton, Lee & Starner 2006), onde é criado um sistema para ajudar as crianças surdas a praticarem suas habilidades em língua de sinais. As crianças vestem duas luvas pequenas com acelerômetros sem fio, posicionados na parte traseira dos pulsos. A abordagem usa uma adaptação do histograma de cores para uma segmentação robusta e rastreamento da mão. A informação da forma da mão é combinada com os dados do acelerômetro e são usadas para treinar os HMMs.

A luva com sensores tem um bom desempenho em reconhecimento de sinais, com alta confiabilidade e eliminação da fase de pré-processamento. Porém, a principal desvantagem é que o usuário precisa vestir a luva, prejudicando a comodidade e os movimentos naturais do usuário.

2.4 Métodos Híbridos

Como foi mencionado, uma imagem é a representação de uma cena e proporciona informação visual. Mas dependendo do cenário podem existir problemas na segmentação, seleção ou extração de dados. Uma solução foi usar dispositivos, uma luva com sensores adiciona informação de localização e uma luva colorida facilita a segmentação da imagem. Porém, esses dispositivos limitam a execução do sinal. Na atualidade existem dispositivos que permitem segmentar a imagem sem precisar utilizar dispositivos intrusi-

vos, eles obtêm informações visuais, de profundidade e/ou das posições das articulações do corpo, sem limitar os movimentos naturais do usuário. Zahedi & Manashty (2011) propõem trabalhar com câmeras de profundidade (*Time-of-Flight*, ToF), com as quais é mais fácil obter a informação dos objetos em questão, sem a necessidade de fazer uma calibração extra. A finalidade desse modelo é converter os sinais gravados com a câmera, em uma língua de sinais padrão e um arquivo XML portável, chamado SiGML (Signing Gesture Markup Language). Desta forma, é muito mais rápido transferir e converter os sinais em animações virtuais 3D.

Além das câmeras ToF, existe o sensor Kinect da Microsoft. Baseados na literatura, é possível dizer que ele é o dispositivo mais usado na atualidade. O Kinect é capaz de obter todas as informações que foram antes mencionadas. Em (Biswas & Basu 2011) as imagens de profundidade são usadas para a remoção do fundo, seguido pela geração do perfil de profundidade da pessoa que realiza o sinal. Além disso, a diferença entre os *frames* consecutivos gera o perfil de movimento da pessoa e é utilizado para o reconhecimento de gestos. Eles permitem o uso eficiente da câmera de profundidade para reconhecer com sucesso vários gestos humanos. O sistema foi treinado usando SVM.

Em (Doliotis, Athitsos, Kosmopoulos & Perantonis 2012) é descrito um método que, dada uma imagem de entrada representando uma pessoa realizando um gesto em uma cena complexa, localiza o braço em questão. Automaticamente detecta e segmenta a mão, finalmente, cria uma lista ordenada de possíveis classes de formas, orientação da posição em 3D e parâmetros de configuração da mão, para usá-los no processo da classificação. O algoritmo de segmentação da mão em uma cena complexa é baseado em dados de profundidade de uma única imagem capturada com um sensor, por exemplo, o Kinect. A possibilidade de realizar estes processos, e sem grande custo computacional, são algumas das vantagens da utilização do Kinect.

Em (Zafrulla, Brashear, Starner, Hamilton & Presti 2011) é feita uma comparação entre um sistema baseado em Kinect e o sistema CopyCat do modelo (Brashear, Zafrulla, Starner, Hamilton, Presti & Lee 2010) o qual usa luvas coloridas e acelerômetros incorporados para rastrear os movimentos das mãos. Os experimentos mostram que os resultados de verificação com o Kinect são comparáveis ao sistema CopyCat. Porém, as principais vantagens foram que o Kinect melhora a interatividade e o conforto do usuário, a robustez e a sustentabilidade do sistema, além do custo e facilidade da implantação. As aplicações para reconhecimento de sinais ou de gestos utilizando o Kinect são muito variadas. Um exemplo é o modelo apresentado em (Chang, Chen & Huang 2011), onde é avaliada uma possibilidade de reabilitar dois adultos com deficiências motoras,

utilizando um sistema baseado em Kinect em uma escola pública.

Pugeault & Bowden (2011) criaram um modelo usando informações de intensidade e profundidade obtidas usando o sensor Kinect. Na descrição de características foram aplicados Filtros de Gabor e na tarefa de classificação foi usado *Random Forests*. Obtiveram uma acurácia de 75% usando a combinação de intensidade e profundidade. Quando ambas características foram testadas de forma separada, foi obtido 73% de acurácia para imagens de intensidade e 69% para imagens de profundidade. A mesma base de dados foi usada em um outro modelo (Zhu & Wong 2012), nesse modelo primeiro são extraídas as características comuns a nível de janelas, e logo estas são combinadas por meio de descritores kernel. Esta abordagem usa apenas 40 amostras de treinamento e as restantes são usadas para teste. Eles obtiveram dois resultados, o primeiro de 77% e um segundo de 88%, onde foi acrescentado uma pirâmide de imagens. Um terceiro modelo usa também essa base de dados (Estrela, Cámara-Chávez, Campos, Schwartz & Nascimento 2013). O modelo baseia-se na estratégia *bag-of-visual-words* combinada com a técnica *Partial Least Squares* (PLS) a fim de criar modelos das letras do alfabeto manual. Ele também usa a *Binary Appearance and Shape Elements* (BASE), o qual é um descritor rápido e de baixo custo que combina informações de intensidade e forma. No modelo, são avaliados os descritores SIFT e BASE, e os classificadores SVM e PLS. Destas comparações, a melhor acurácia obtida foi usando o descritor SIFT e o classificador PLS com 71,51%.

2.5 Considerações Finais

Nesta seção foram apresentadas diversas abordagens para o reconhecimento de sinais. Nem sempre é possível realizar cada passo do reconhecimento da melhor forma. Por isso cada abordagem tem uma visão diferente. A obtenção de dados, além de ser o passo inicial, é quem faz a diferença nos processos seguintes. Como já foi mencionado, o processo de segmentação em imagens obtidas por uma câmera simples é mais complicado comparado com a segmentação de imagens que contam com informação de algum dispositivo adicional, como as luvas. Embora, esses dispositivos forneçam informações que ajudem na solução dos demais processos, eles não são os mais ideais. O Kinect além de obter mais informação, melhora a interatividade com o usuário, facilitando a realização do sinal. Também é importante determinar qual é o classificador mais adequado para nosso problema com base no desempenho que eles obtêm de acordo com o tipo de entrada.

Capítulo 3

Fundamentos Teóricos

Neste capítulo apresentaremos a teoria e as definições necessárias para entender esta dissertação. Serão revisados conceitos básicos concernentes à língua de sinais, métodos de aquisição de dados, extração de características e classificação.

3.1 Língua de Sinais

Língua de sinais é uma língua baseada em gestos, sinais e expressões faciais e/ou corporais, assim como os sons são usados na fala. Sua aquisição é basicamente visual e espacial. São línguas naturais de cada comunidade de surdos. Existem no mundo muitas línguas de sinais usadas como forma de comunicação entre pessoas surdas ou com problemas auditivos. Muitas dessas línguas têm reconhecimento oficial em vários países.

3.1.1 Sinais

Sinais estão compostos pela combinação da forma e do movimento das mãos em relação ao corpo ou a um ponto no espaço onde os sinais são realizados. Em língua de sinais são definidos os seguintes cinco parâmetros que conformam o sinal (William Vicars 97-2013b):

Forma Define a configuração das articulações da mão. Na Figura 3.1, são mostrados três exemplos de sinais que usam a mesma forma da mão.

Orientação Especifica a direção da mão e aonde os dedos estão apontando. Na Figura 3.2, são mostrados alguns exemplos deste parâmetro.

Localização Indica a posição da mão com relação ao corpo. A mão pode se encontrar em um espaço neutro (perto do corpo) ou na testa, como pode ser visto nos exemplos da Figura 3.3.

Movimento É a mudança no tempo de qualquer uma das três funções acima descritas. É a característica mais complexa. Contudo, nem todos os sinais têm movimento. Na Figura 3.4, são encontrados exemplos de sinais com e sem movimento.

Expressões faciais e/ou corporais São de importância fundamental para entender os sinais. A entonação em língua de sinais é realizada através destas expressões que são chamadas de não-manuais. Na Figura 3.5, encontramos alguns exemplos de expressões faciais.

Figura 3.1: Exemplo de gestos usando a forma do sinal *S* (Felipe & Monteiro 2007).

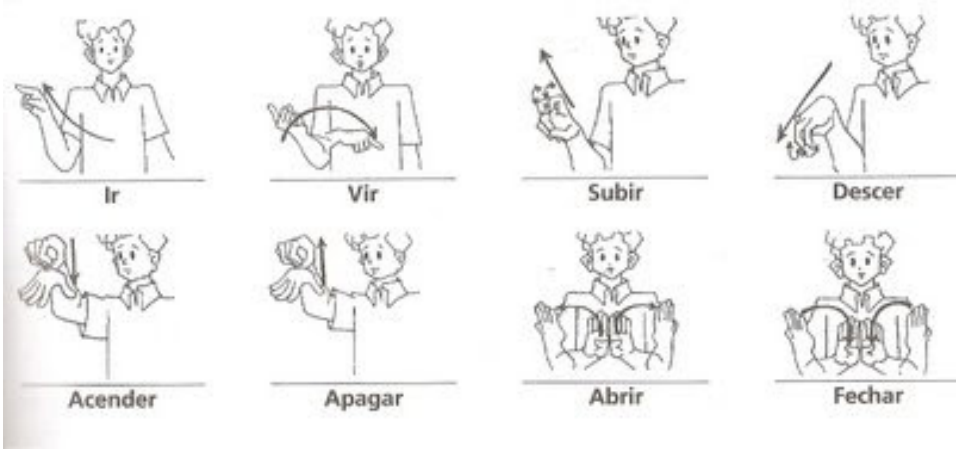

Figura 3.2: Exemplo de orientações da mão (Felipe & Monteiro 2007).

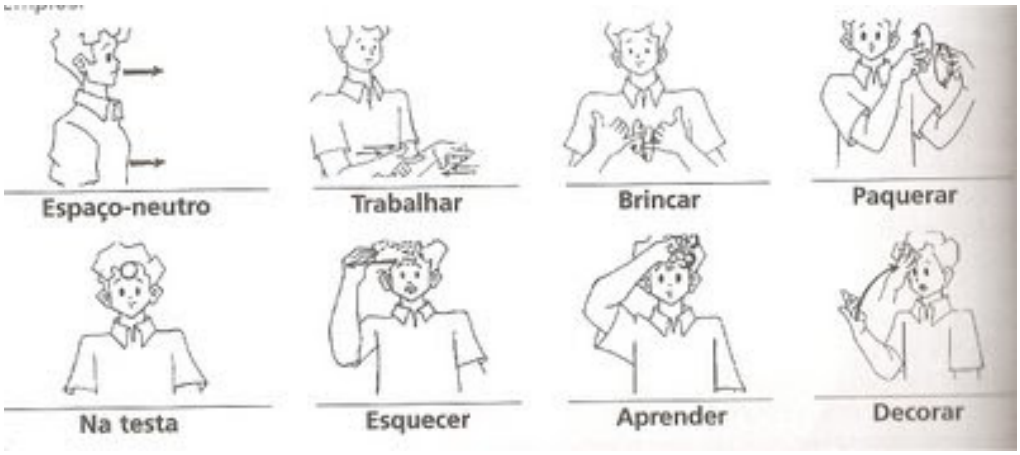

Figura 3.3: Exemplo de sinais com distintas localizações (Felipe & Monteiro 2007).

Figura 3.4: Exemplo de sinais estáticos e dinâmicos (Felipe & Monteiro 2007).

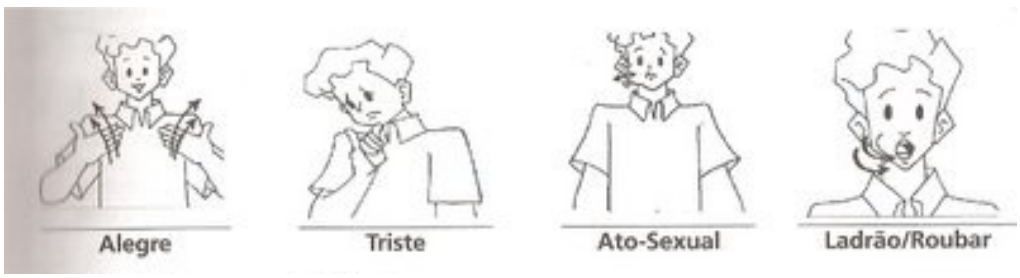

Figura 3.5: Exemplo de expressões faciais (Felipe & Monteiro 2007).

3.2 Aquisição de dados

O primeiro passo em qualquer método é a aquisição de dados. Existem diversos dispositivos que podem ser utilizados para a obtenção de informação. Alguns exemplos são as câmeras digitais, luvas, dispositivos de profundidade, entre outros.

3.2.1 Luvas

O dispositivo mais básico é uma câmera simples para obter imagens RGB. Mas é possível adicionar informação usando luvas. Estas podem ser coloridas e/ou com sensores. A Figura 3.6 mostra as luvas coloridas usadas no modelo apresentado em (Wang, Chen, Zhang, Wang & Gao 2007). Na Figura 3.7, é mostrado um exemplo onde são usadas luvas coloridas providas de sensores que foram usadas no sistema CopyCat de diferentes modelos (Brashear, Henderson, Park, Hamilton, Lee & Starner 2006, Brashear, Zafrulla, Starner, Hamilton, Presti & Lee 2010). CopyCat é um jogo baseado em língua de sinais americanas (*American Sign Language*, ASL) criada para ajudar à crianças surdas praticarem suas habilidades em ASL. Inicialmente, como já foi mencionado, a forma de aquisição era usando luvas, as quais depois foram substituídas pelo Kinect.

3.2.2 Kinect

Kinect, cujo *code name* foi *Project Natal*, foi desenvolvido pela Microsoft e PrimeSense e lançado em novembro de 2010. O Kinect possui uma câmara RGB, juntamente com um agrupamento de microfones e um sensor de profundidade. Este sensor utiliza um projetor de infravermelho para medir a profundidade de cada pixel, sendo capaz de rastrear todo o corpo do usuário independentemente das condições de iluminação. Na Figura 3.8, são mostrados os componentes do Kinect.

Inicialmente, o Kinect foi destinado para os proprietários do console *Microsoft Xbox 360*. Prometendo um videogame e experiência de entretenimento sem a necessidade de *gamepads* ou de outros dispositivos de interação como pode ser visto na Figura 3.9, mudando completamente a visão atual da interatividade com videogames.

Em junho de 2011 a Microsoft permite utilizar o sensor Kinect a partir um computador usando o SDK (Software Development Kit) *Kinect for Windows*. Empresas do mundo inteiro estão usando o sensor Kinect para desenvolver e implantar soluções

Figura 3.6: Luvas coloridas. (a) Palma da luva dominante. (b) Parte de trás da luva dominante. (c) Palma da luva não-dominante. (d) Parte de trás da luva não-dominante.

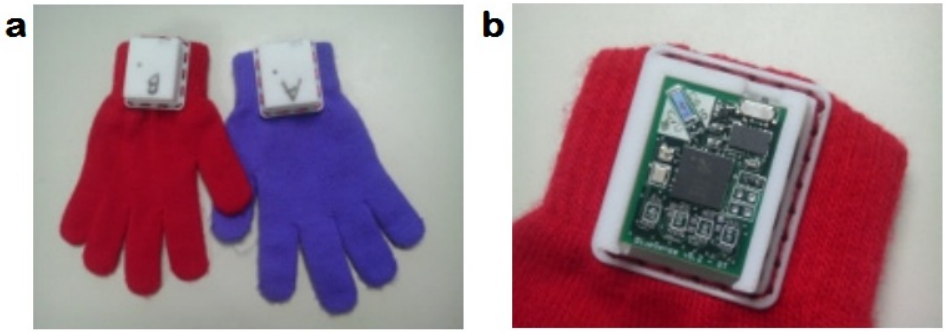

Figura 3.7: Luvas com sensores. (a)Luvas com acelerômetros (b) Detalhe dos acelerômetros montados no pulso.

que nos dão a capacidade de interagir naturalmente com os computadores simplesmente gesticulando e falando (©2013 Microsoft Corporation 2013).

Na Figura 3.10 é mostrado um exemplo da informação visual obtida usando o Kinect. Ao lado esquerdo, encontra-se a imagem em RGB e do lado direito a imagem que

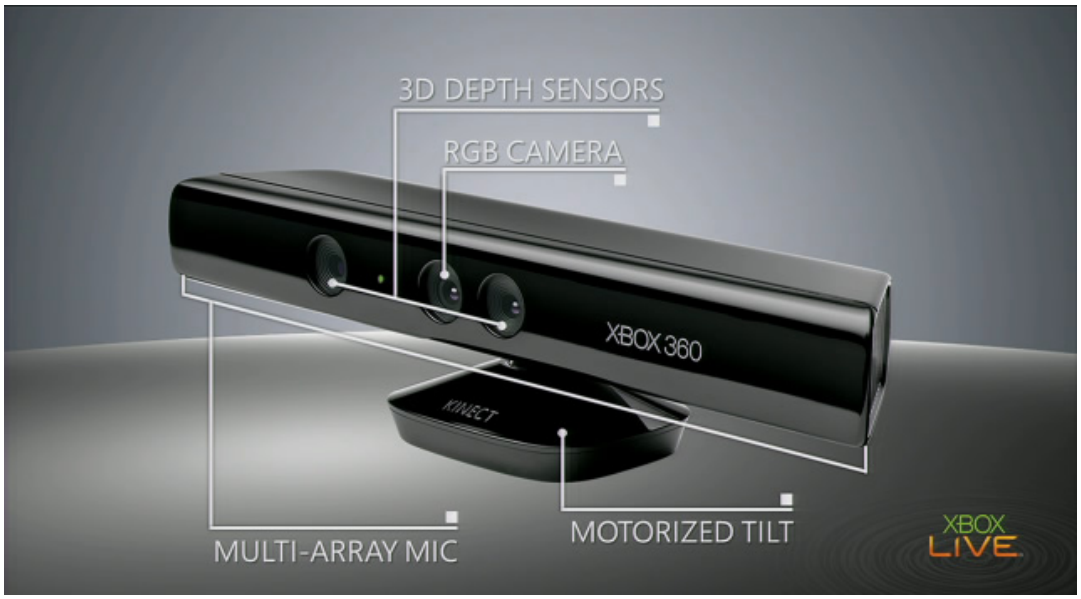

Figura 3.8: O Kinect e seus componentes.

Figura 3.9: Kinect para XBOX 360.

representa a profundidade da cena. Além desta informação, também é possível obter informação das articulações do corpo. São vinte pontos de articulação e podem ser vistos em detalhe na Figura 3.11

Figura 3.10: Exemplo das imagens obtidas pelo Kinect. Esquerda: imagem RGB. Direita: imagem de profundidade.

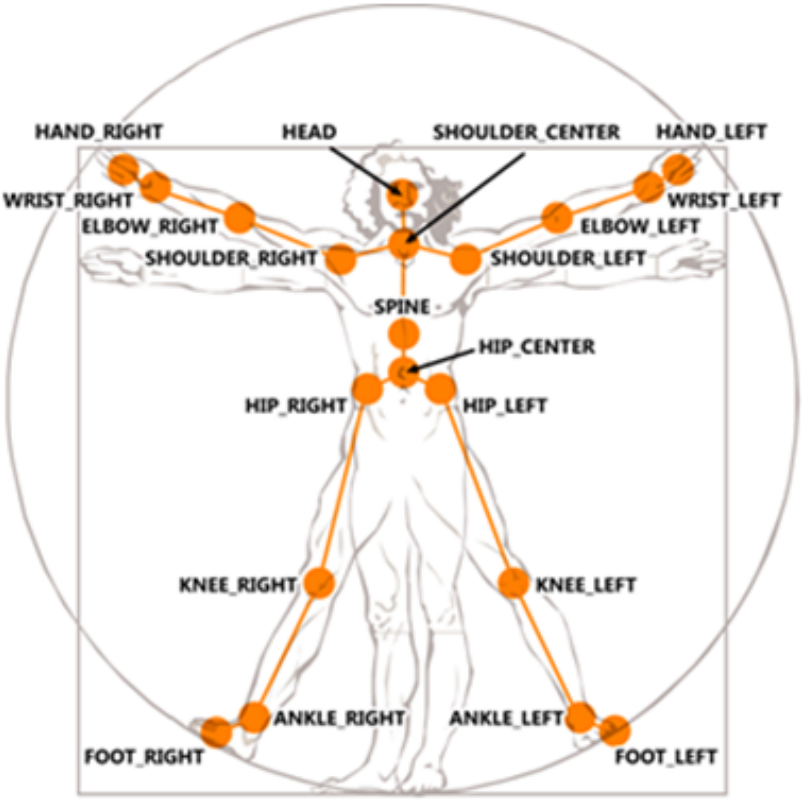

Figura 3.11: Pontos de articulação do corpo obtidos pelo Kinect.

3.3 Extração de Características

A extração de características é um passo muito importante para o processo de reconhecimento. Essas características devem representar da melhor forma possível a imagem em questão. Definiremos os descritores usados em nosso modelo.

3.3.1 Scale-Invariant Feature Transform (SIFT)

SIFT é um algoritmo utilizado em visão computacional para detectar e descrever características locais em imagens. O algoritmo foi proposto por Lowe (1999) e patenteado nos EUA pela *University of British Columbia*.

SIFT é um método para extrair características invariantes distintivas de uma imagem que podem ser usadas para realizar um casamento confiável entre as diferentes vistas de um objeto ou de uma cena. Elas estão bem localizadas em ambos domínios, espacial e da frequência, reduzindo a probabilidade de perturbação por oclusão, desordem ou ruído. As características são altamente distintivas, no sentido que, uma única característica pode ser casada corretamente com uma elevada probabilidade, fornecendo uma base para reconhecimento de objetos ou cenas. SIFT tem sido usado em muitas aplicações, tais como reconhecimento de objetos, mapeamento robótico, modelagem 3D, reconhecimento de gestos, rastreamento de vídeo e na identificação de pessoas. (Lowe 2004). A eleição do descritor SIFT é devido a ele ser invariante a escala e rotação, e parcialmente invariante a mudanças na iluminação e ponto de vista 3D. Além de obter bons resultados nas abordagens onde ele é utilizado.

SIFT minimiza o custo de extrair as características através de uma abordagem de filtragem em cascata. As operações mais custosas são aplicadas somente em locais que passam um teste inicial.

A seguir estão as etapas do cálculo utilizado para gerar o conjunto de características da imagem:

Construção de um espaço de escala Neste primeiro passo é gerado um espaço de escala. Para cria-lo, são geradas progressivamente imagens suavizadas a partir da imagem original. Cada nível de suavizamento é chamado de "escala". Logo, a imagem original é redimensionada à metade do seu tamanho. Em seguida, são geradas novamente imagens suavizadas a partir da imagem original já redimensi-

onada. Este processo é repetido tantas vezes quanto for necessário. Cada nível de imagens redimensionadas são chamadas de "*octaves*". O número de *octaves* e escalas dependem do tamanho da imagem original. No entanto, o criador do SIFT sugere que quatro "*octaves*" e cinco níveis de suavizamento são ideais para o algoritmo. A fim de suavizar a imagem é feita uma convolução de um operador gaussiano sobre cada pixel da imagem.

Aproximação do Laplaciano de Gaussianas São obtidas bordas e quinas da imagem para localizar os pontos de interesse, que serão detectados no passo seguinte. A fim de obter essa informação é possível aplicar um operador Laplaciano sobre a imagem, mas esse proceso é muito custoso computacionalmente. Então, é aplicada uma outra abordagem, usando o espaço de escalas do passo anterior. A qual consiste em calcular a diferença entre duas escalas consecutivas, o que realmente trata-se de uma diferença de Gaussianas (*Difference of Gaussians*, DoG). Isso é aplicado em todos os *octaves*. Gerando DoGs de diferentes tamanhos. Essas DoGs são aproximadamente equivalentes a aplicar o Laplaciano de Gaussianas invariante à escala (o qual é útil para detetar pontos de interesse), com a diferença de ser um processo mais simples.

Localização de pontos de interesse As imagens obtidas no passo anterior contêm pontos de interesse, os quais serão localizados aplicando dois passos. (a) No primeiro passo, são detetados os máximos e mínimos valores. O processo é simples, tem que percorrer cada pixel e verificar sua vizinhança, são usadas a imagem atual e as imagens de acima e de abaixo. Fazendo um total de 26 vizinhos. O pixel atual é marcado como "ponto de interesse" se ele é maior ou menor a todos seus 26 vizinhos. Pode-se notar que não são detetados pontos de interesse nas escalas inferiores ou superiores devido a não ter suficientes vizinhos para fazer a comparação. Uma vez feito isso, os pixels marcados são "aproximadamente" os máximos e mínimos valores devido a que esses valores quase nunca estão exatamente no pixel. Eles estão em um lugar dentro do pixel, mas não é possível acessar a esse valor. Então, a posição do sub-pixel deve ser localizado matematicamente. (b) No segundo passo, usando os dados do pixel, são gerados valores de sub-pixels. Isto é feito através da expansão de Taylor da imagem em torno do ponto de interesse aproximado. Estes valores de sub-pixel aumentam as chances de casamento e de estabilidade do algoritmo.

Filtrar os pontos de interesse Os pontos de interesse encontrados, em alguns casos, encontram-se ao longo da borda ou não têm suficiente contraste. Em qualquer

dos casos, essas características não são úteis e devem ser removidas. A abordagem usada é uma parecida à usada no detector de quinas de Harris. Para remover as características de baixo contraste, é comparada a intensidade do pixel atual na imagem DoG com um valor dado (limiar), se ele é menor, então é rejeitada. Para remover as características que estão ao longo da borda, são calculadas duas gradientes no ponto de interesse. Ambas perpendiculares entre si. A imagem em torno do ponto de interesse pode ser: (a) Uma região plana, se ambos os gradientes são pequenos. (b) Uma borda, se um gradiente é grande (perpendicular à borda), e o outro é pequeno (ao longo da borda). (c) Uma quina, se ambos os gradientes são grandes.

Quinas são ótimos candidatos para serem pontos de interesse, portanto, valores elevados de gradientes são procurados, sendo rejeitados aqueles pontos de interesse que não possuem essa característica.

Atribuição de orientação aos pontos de interesse Até este ponto foi atingida a invariância à escala. O próximo passo é atribuir uma ou mais orientações para cada ponto de interesse para atingir a invariância à rotação. A ideia é reunir direções de gradiente e magnitudes em torno de cada ponto de interesse. Logo, são calculadas as mais importantes orientações da região. Finalmente, essas orientações são atribuídas ao ponto de interesse. Qualquer cálculo posterior é feito em relação a essas orientações.

Gerar as características Finalmente, no passo final do SIFT é criada uma "impressão digital" para cada ponto de interesse com a finalidade de identificá-os. Por exemplo, se um olho é um ponto de interesse, usando esta "impressão digital", é possível distingui-lo de outros pontos de interesse, como as orelhas, nariz, dedos, etc.

Por cada ponto de interesse, é tomada uma janela de 16x16 ao seu redor. Essa janela é dividida em 16 janelas de 4x4. Para cada janela de 4x4 é gerado um histograma de 8 *bins*. Cada *bin* corresponde a orientações de gradiente, por exemplo, no primeiro *bin* corresponde a orientações encontradas no intervalo [0,44] graus, o segundo a [45,89] graus, e assim até o oitavo *bin* que completa os 360 graus. Cada valor de orientação de gradiente é contado no seu respectivo bin. Isso é feito para todos as janelas de 4x4. O resultado final é um vetor de dimensão 128. É assim que cada ponto de interesse é identificado exclusivamente por esse vetor de características.

3.3.2 Gradient Kernel Descriptor

Os histogramas de orientação, tais como SIFT ou HOG, encontram-se entre os descritores de baixo nível mais populares e com maior sucesso. Bo, Ren & Fox (2010) mostraram que os histogramas de orientação são equivalentes a um certo tipo de kernel de casamento (*match kernels*) aplicados sobre regiões da imagem (*image patches*). Isto permitiu aos autores desenhar uma familia de descritores kernel, os quais fornecem um *framework* unificado para converter atributos de pixels (gradiente, cor, *local binary pattern* (LBP), etc.) em características a nível de regiões (*patch-level*). Os *match kernels* permitem medir semelhanças entre regiões da imagem e construir descritores baseados em kernel de baixa dimensão a partir dos *match kernels* usando a análise de componentes principais (*Kernel Principal Component Analysis*, KPCA).

O descritor de características de baixo nível consiste de 3 etapas. (a) Desenhar o *match kernel* usando algum atributo dos pixels, (b) Aprender *compact basis vectors* usando KPCA, e (c) Construir os descritores kernel projetando os vetores de características de dimensão infinita para os vetores de base aprendidas.

Em (Bo, Ren & Fox 2011), é apresentado outro modelo proposto pelos mesmos autores, onde entre outros descritores é encontrado o descritor kernel gradiente. Este descritor é aplicado sobre imagens de profundidade para capturar características das bordas da imagem.

O *gradient match kernel*, consiste de três kernels. (a) O kernel de posição Gaussiana k_s mede quão perto de dois pixels estão espacialmente, (b) O kernel de orientação k_o calcula a similaridade das orientações do gradiente, e (c) O kernel linear normalizado pondera a contribuição de cada pixel usando as magnitudes do gradiente.

O *gradient match kernel*, K_{grad} (Bo, Ren & Fox 2010), é definido da seguinte forma:

$$K_{grad}(PQ) = \sum_{p \in P} \sum_{q \in Q} \tilde{m}(p)\tilde{m}(q)k_o(\tilde{\theta}(p)\tilde{\theta}(q))k_s(p,q)$$

O kernel linear normalizado $\tilde{m}(p)\tilde{m}(q)$ pondera a contribuição de cada gradiente onde $\tilde{m}(p) = m(p)/\sqrt{\sum_{p \in P} m(p)^2 + \varepsilon_g}$ é uma constante positiva pequena para garantir que o denominador é maior do que 0 e $m(p)$ é a magnitude do gradiente de profundidade no pixel p. Logo, $k_o(\tilde{\theta}(p)\tilde{\theta}(q)) = exp(-\gamma_o \|\tilde{\theta}(p) - \tilde{\theta}(q)\|^2)$ é um kernel Gaussiano sobre orientações. $\tilde{\theta}(p)$ e $\tilde{m}(p)$ são a orientação e a magnitude do gradiente da profundidade em um pixel p. Os autores de (Bo, Ren & Fox 2010) sugerem $\gamma_o = 5$. Para estimar

a diferença entre as orientações em pixels p e q, usamos os seguintes vetores gradiente normalizados em função do kernel k_o:

$$\tilde{\theta}(p) = [sin(\theta(p))cos(\theta(p))]$$
$$\tilde{\theta}(q) = [sin(\theta(q))cos(\theta(q))]$$

onde $\theta(p)$ é a orientação do gradiente de profundidade no pixel p. Kernel de posição Gaussiana $k_s(p,q) = exp(-\gamma_s\|p-q\|^2)$ com p denotando a posição 2D de um pixel em um *image patch* (normalizado a [0,1]), isto mede quão perto dois pixels estão espacialmente (Bo, Ren & Fox 2011). O valor sugerido para γ_s é 3.

Na figura 3.12 mostra que, como foi mencionado, é criada a abordagem de kernel de casamento (*match kernel*) baseados nos métodos SIFT e HOG. Onde são encontrados os três kernels de casamento descritos anteriormente. A região da imagem tomada pelo autor (Bo, Ren & Fox 2010) tem tamanho 16 $\times$ 16.

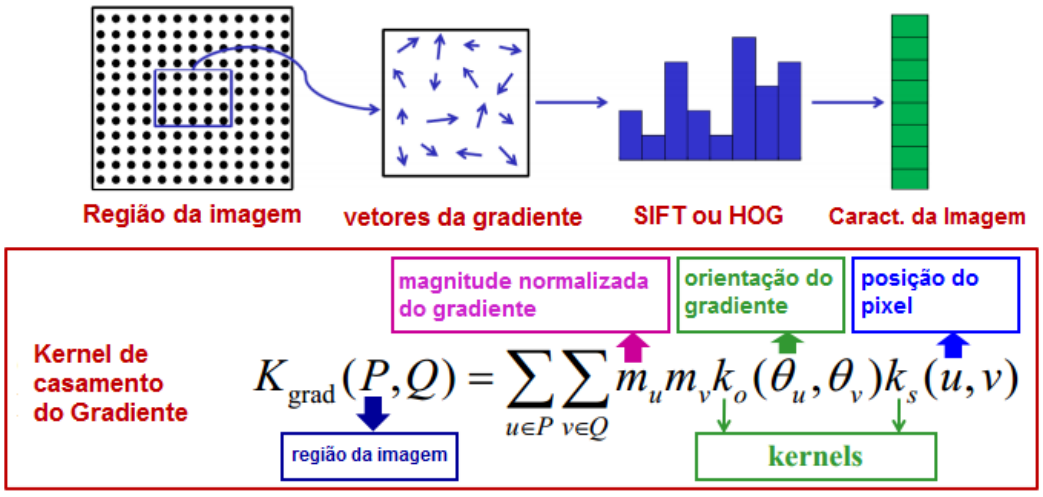

Figura 3.12: Descritor kernel gradiente. Kernel de casamento aplicado sobre regiões da imagem.

Então, kernels de casamento fornecem um princípio para medir a similaridade entre regiões de imagem, contudo a avaliação de kernels pode ser computacionalmente cara quando as regiões da imagem são grandes (Bo, Ren & Fox 2010). A fim de facilitar esse processo, o descritor correspondente ao kernel pode ser extraído a partir deste kernel de casamento projetando o vetor característico de dimensão infinita a um conjunto de vetores de base finitas.

3.3.3 Filtros de Gabor

São filtros passa-banda os quais têm sido usados em muitas aplicações, tais como segmentação de textura, análise de documentos, detecção de bordas, identificação de retina, codificação e representação de imagens.

Os filtros de Gabor são um conjunto de funções bidimensionais Gaussianas moduladas por funções sinusoidais complexas e apresentam certas propriedades extremamente úteis como: invariância à iluminação, rotação, escala e translação. A forma básica do filtro de Gabor é definida da seguinte forma (Pugeault & Bowden 2011):

$$g(x, y, \lambda, \sigma, \theta) = \exp\left(\frac{x'^2 + y'^2}{2\sigma^2}\right) \exp\left(i(\frac{2\pi x}{\lambda})\right)$$

onde $x' = x\cos\theta + y\sin\theta$ e $y' = -x\sin\theta + y\cos\theta$, λ é o comprimento de onda, θ é a orientação e σ é a variância da curvatura Gaussiana. Na Figura 3.13, é apresentado um exemplo do filtro de Gabor no domínio espacial e das frequências.

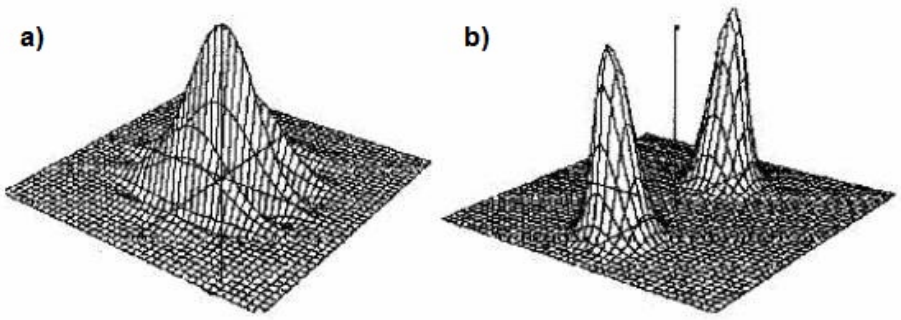

Figura 3.13: (a) Componente real do filtro bidimensional de Gabor no domínio espacial; (b) Resposta de frequência do filtro definido em a.

3.4 Algoritmos de aprendizagem máquina

Nesta seção serão mostrados alguns algoritmos de aprendizagem máquina, os quais foram encontrados no estado da arte e alguns deles são usados nesta dissertação. Temos algoritmos de agrupamento (clustering) como K-means ou LBG. Além de algoritmos de aprendizagem supervisado como SVM e Random Forest que são usados na classificação.

3.4.1 Clustering

É um procedimento de agrupamento de vetores de acordo a um critério. Esse critério pode ser baseado na distância ou na similaridade dos dados. A seguir, apresentaremos dois algoritmos: K-means e LBG.

K-means

É um algoritmo de aprendizagem máquina e mineração de dados no qual são agrupados os dados de acordo a uma relação entre eles. Esse agrupamento é realizado sem qualquer conhecimento prévio dessas relações. K-means é uma das técnicas mais simples de agrupamento e é comumente usado em biometria, medicina e áreas afins. O algoritmo consiste em dividir os dados em k grupos, onde k é um parâmetro de entrada dado pelo usuário de acordo ao seu problema. Em seguida apresentamos os passos do algoritmo:

1. Selecionar arbitrariamente k pontos como os centros iniciais de cada cluster (centróides).
2. Cada ponto na base de dados é atribuído a um cluster baseado na distância Euclidiana entre o ponto e o centróide.
3. Cada centróide é recalculado pela média dos pontos que pertencem ao cluster.
4. Passos 2 e 3 são repetidos até que os clusters convirjam. Isto é, que não há mais alterações nos clusters quando os passos 2 e 3 são executados.

Quando o algoritmo termina, teremos os dados agrupados em k clusters. Se o algoritmo é executado novamente, os resultados não serão os mesmos, devido a que os centróides são selecionados arbitrariamente. De acordo ao problema, o usuário pode definir os centróides iniciais para diminuir variações no agrupamento.

Algoritmo *Linde-Buzo-Gray*(LBG)

Linde, Buzo & Gray (1980) apresentam seu algoritmo como uma variante do algoritmo de Lloyd (também conhecido como *K-means*). Eles estenderam o algoritmo por uma abordagem cuidadosa da inicialização e, muitas vezes conseguindo um melhor desempenho em termos de minimizar a distância total dentro do cluster. O algoritmo funciona da seguinte forma:

1. Encontrar a média da base de dados total, chamaremos de centróide $z_1^{(1)}$.

2. Definir $k = 1$ e $l = 1$; onde l é o índice da iteração e k conta o número de protótipos que têm sido gerados.

3. Se $k < M$, (M é o número alvo de centróides) dividir o centróide atual adicionando um pequeno deslocamento. Da seguinte forma:

   (a) Se $M - k \geq k$, dividir todos os centróides; caso contrário; dividir só $M - k$ centróides.

   (b) Denotar o número de divisões de centróides por $\tilde{k} = min(k, M - k)$.

   (c) Exemplo: para dividir $z_1^{(1)}$ em dois centróides, fazemos $z_1^{(2)} = z_1^{(1)}$, $z_s^{(2)} = z_1^{(1)} + \epsilon$, onde ϵ é um pequeno deslocamento.

4. $k \leftarrow k + \tilde{k}$ e $l \leftarrow l + 1$. A divisão foi realizada.

5. Usar $z_1^{(l)}, z_2^{(l)}, ..., z_k^{(l)}$ como protótipos iniciais. Aplicar k-means para atualizar esses protótipos.

6. Se $k < M$, voltar para o passo 3, caso contrário, parar.

3.4.2 Classificação

Para a fase da classificação diversos algoritmos podem ser usados. Temos o SVM e o *Random Forest* que foram encontrados na literatura e serão descritos a seguir.

Support Vector Machine

Máquina de vetor de suporte (*Support Vector Machine*, SVM) é um método muito útil de classificação. O método original foi proposto por Vladimir N. Vapnik em 1963 e foi introduzida como um método de aprendizagem de máquina (*machine learning*) por Cortes & Vapnik (1995).

SVM tem sido aplicado com sucesso em muitos problemas da vida real e em diversas áreas: categorização de texto, reconhecimento de digitais, reconhecimento de objetos, etc. A SVM tem sido desenvolvida como uma ferramenta robusta para classificação e regressão em ruído e domínios complexos. SVM pode ser usado para extrair valiosa informação de conjuntos de dados e construir algoritmos de classificação rápida para grande quantidade de dados.

Através de um kernel, pontos de dados podem ser mapeados em um espaço de características de alta dimensão, onde um hiperplano separa os dados das classes com margem

máxima, como mostra a Figura 3.14. Como o algoritmo original foi proposto como um classificador linear, Boser, Guyon & Vapnik (1992) sugeriram uma maneira de criar classificadores não-lineares, aplicando o *kernel trick* para hiperplanos de margem máximo. O algoritmo resultante é formalmente semelhante, exceto que cada produto escalar é substituído por uma função de kernel não-linear. Isto permite ao algoritmo ajustar o hiperplano de margem máximo em um espaço de características transformado (ver Figura 3.15). Os kernels são usados para calcular uma série de funções não-linearmente separáveis em uma função linearmente separável de maior dimensão. É por isso, que uma importante característica do classificador SVM é que permite a classificação não-linear sem requerer explicitamente um algoritmo não-linear devido à ajuda da teoria de kernels como foi mencionado. Os kernels mais comuns são: linear, polinomial, *Radial basis function (RBF)*, distância χ^2 e triangular.

Além de escolher o tipo de kernel também é importante a normalização dos dados. Recomenda-se escalar cada atributo para o intervalo [-1,1] ou [0,1].

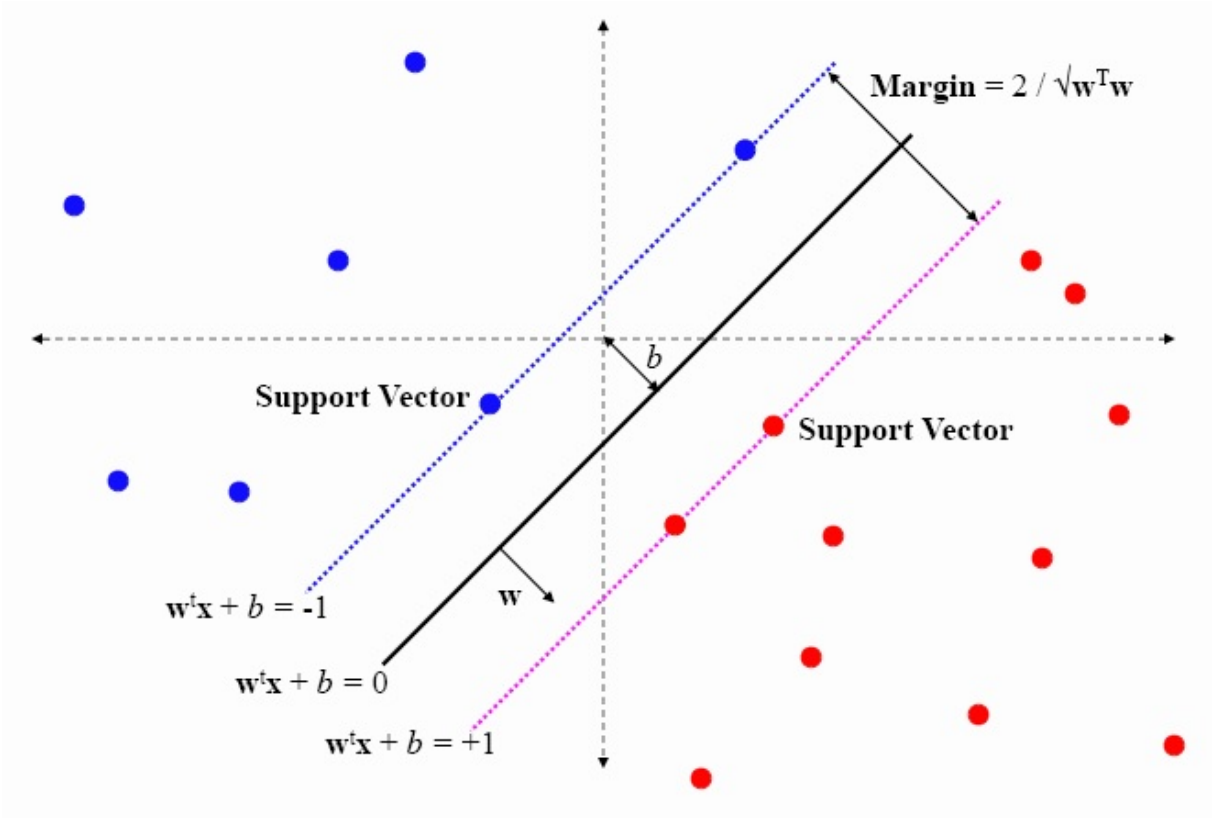

Figura 3.14: Hiperplano com margem máximo e margens para uma SVM treinada com amostras de duas classes. Amostras na margem são chamadas de vetores de suporte.

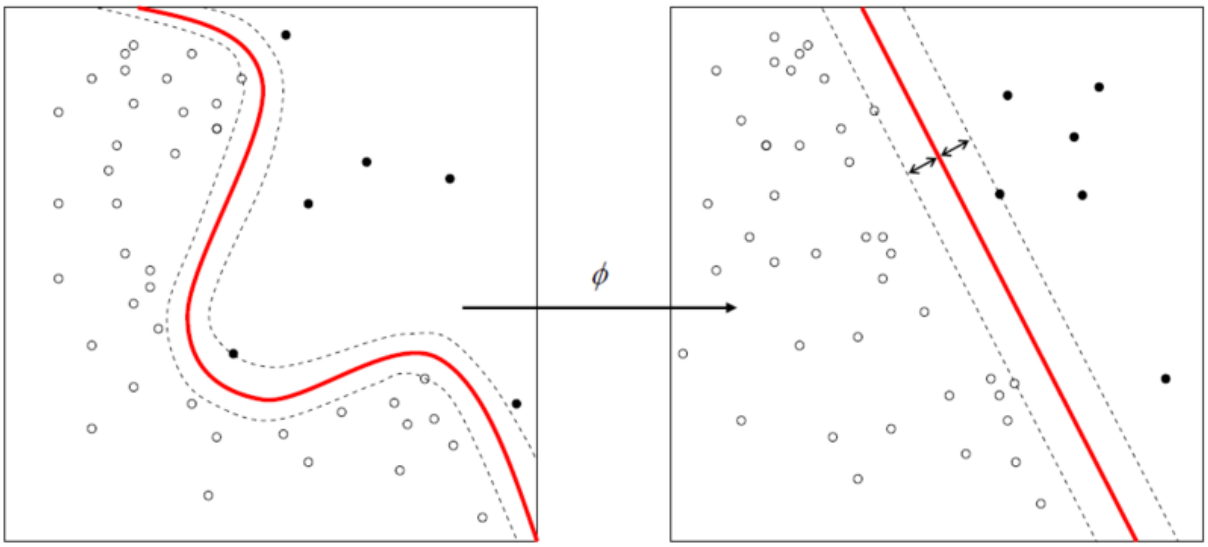

Figura 3.15: Máquinas de kernel são usados para calcular uma série de funções não-linearmente separáveis em uma função linearmente separável de maior dimensão.

Random Forests (RFs)

RFs é um método de aprendizagem em conjunto (*ensemble learning*) para classificação e regressão, que opera através da construção de várias árvores de decisão no momento do treinamento. O algoritmo foi desenvolvido por Leo Breiman (Breiman 2001) e Adele Cutler. RFs são uma combinação de árvores preditores de tal forma que cada árvore depende dos valores de um vetor de uma amostra aleatória de forma independente e com a mesma distribuição para todas as árvores da floresta.

RFs têm características de ter uma alta precisão de classificação e têm a vantagem de ser executado de forma eficiente em grandes bases de dados. Também pode lidar com centenas de variáveis de entrada sem excluir nenhuma. Além de estimar que variáveis são importantes na classificação. Ela gera uma estimativa imparcial do erro de generalização à medida que é construída a floresta. RFs têm um método eficaz para a estimativa de dados ausentes e mantêm a precisão quando uma grande parte dos dados estão ausentes.

O conjunto de treinamento para a árvore atual é definido por amostragem com reposição, quase um terço dos casos são deixados de fora da amostra. Estes dados *oob* (*out-of-bag*) são usados para obter uma estimativa imparcial do erro de classificação. É também utilizado para obter estimativas da importância de uma variável.

O funcionamento do RF pode ser resumido da seguinte forma: RFs constrói várias

árvores de classificação. Para classificar um novo objeto a partir de um vetor de entrada, tem que colocar o vetor de entrada em cada uma das árvores na floresta. Cada árvore dá uma classificação, e dizemos que a árvore "votou"para essa classe. A floresta escolhe a classificação tendo a maioria dos votos (sobre todas as árvores da floresta).

A seguir é mostrado como um sistema é treinado, para um determinado número de árvores T (ver Figura 3.16):

1. Amostrar N subconjuntos de dados de forma aleatória com reposição. O subconjunto deve ser de cerca de 66% do conjunto total. N é o número de árvores.
2. Para cada nó:
   (a) Dado um número m, m variáveis preditivas são selecionadas aleatoriamente do subconjunto total.
   (b) A variável preditiva que fornece a melhor divisão, de acordo com uma função objetivo, é usado para fazer uma divisão binária nesse nó.

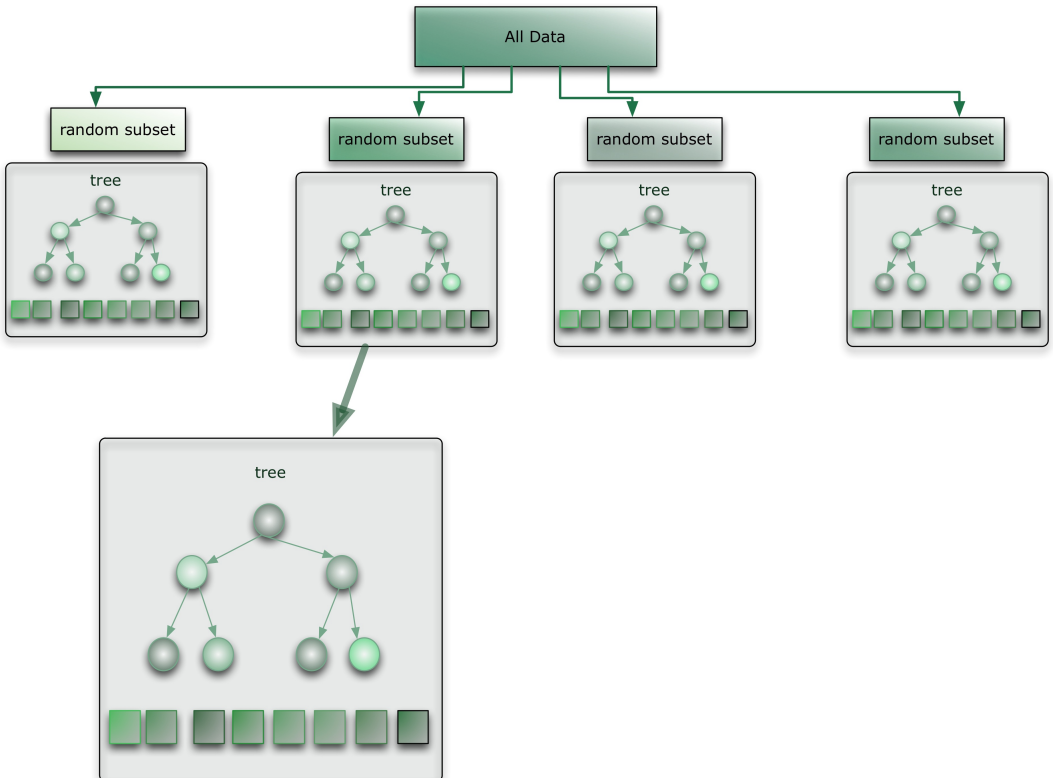

Figura 3.16: Exemplo da execução de *Random Forests*.

3.5 Bag-of-Visual-Words

O modelo *Bag-of-Visual-Words* (BoW) é comumente usado em métodos de classificação de documentos, onde a ocorrência de cada palavra é usada como um recurso para treinar um classificador.

O modelo é uma representação simplificada usada no processamento da linguagem natural e recuperação de informação (*Information retrieval*, IR). Neste modelo, um texto (como uma frase ou um documento) é representado como uma coleção não-ordenada de palavras, desconsiderando a gramática e até mesmo a ordem das palavras. O modelo *Bag-of-words* tem sido introduzido por Sivic para recuperação de vídeo (Video Retrieval) (Sivic & Zisserman 2003). Devido à sua eficiência e eficácia, tornou-se muito popular nas áreas de recuperação de imagens e categorização. O modelo original trabalha com documentos e palavras, então em visão computacional uma imagem pode ser considerada como um documento e as "palavras"(*words*) serão as entidades visuais encontradas na imagem. Neste caso, o método pode ser chamado de *Bag-of-Visual-Words*. Na Figura 3.17 é apresentado uma visão geral do modelo *Bag-of-Visual-Words*.

A abordagem de Bag-of-Visual-Words consiste de três operações: representação de características, geração do dicionário de palavras visuais (*codebook*) e geração do histograma de palavras visuais.

- Representação de características: Muitos descritores locais, como por exemplo o SIFT ou SURF, podem ser usados para a detecção e descrição das características. Onde é obtido um conjunto de vetores de características, um vetor para cada ponto de interesse encontrado na imagem.
- Geração do vocabulário: O vocabulário é gerado a partir dos vetores de características. É realizado um agrupamento (clustering) sobre todos os vetores. Os centróides de cada cluster representam as palavras visuais (*codewords*). Por tanto, cada *codeword* representa um grupo de diversas características similares. A quantidade de clusters representa o tamanho do *codebook*.
- Geração do histograma de *codewords*: Após essas operações, um histograma de palavras visuais é criado contando as ocorrências de cada *codeword*. Essas ocorrências são organizadas em um vetor. Onde cada vetor representa as características para uma imagem.

Na Figura 3.18 é mostrado um exemplo do funcionamento do modelo. A qual, mostra

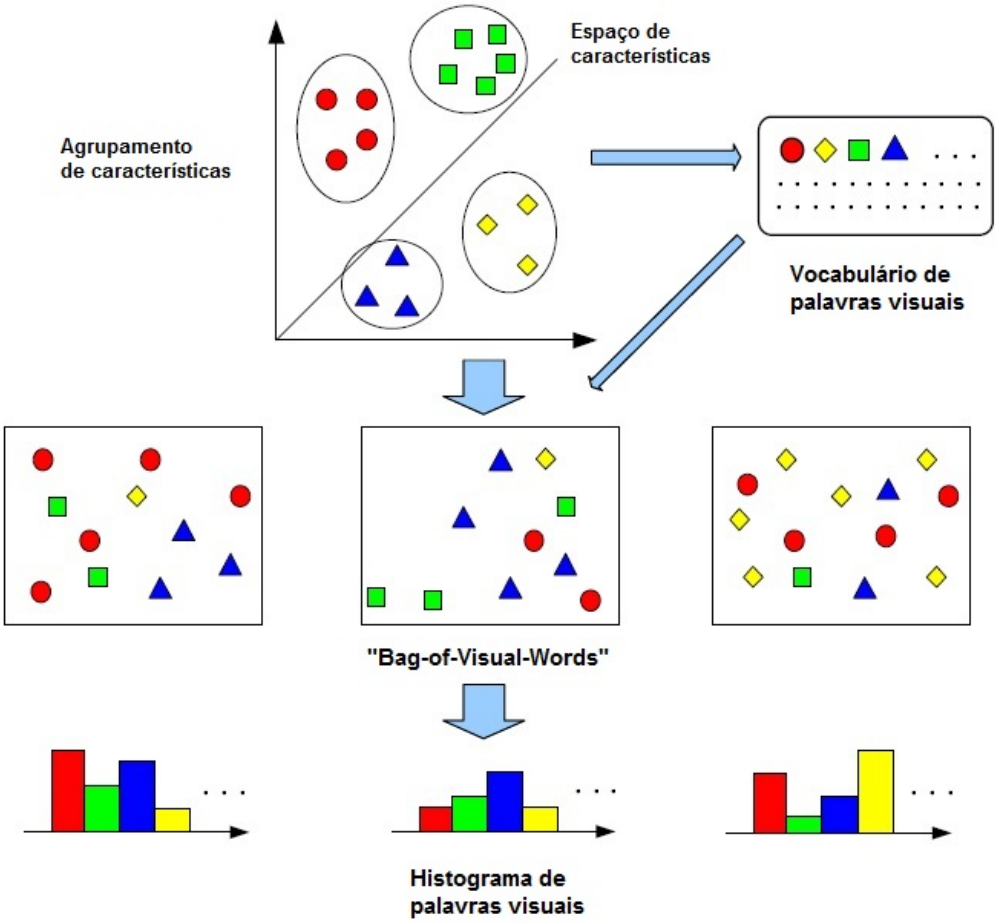

Figura 3.17: Bag-of-Visual-Words baseado em características visuais.

que na primeira parte são extraídas as características dos três objetos. Depois é realizado o agrupamento de acordo as características em comum e são obtidos os *codewords*. Finalmente cada imagem é representada por um vetor de ocorrências de palavras visuais.

3.6 Considerações Finais

Neste capítulo, foi apresentada a teoria necessária para entender esta dissertação. Foram mostrados os conceitos básicos da língua de sinais assim como a descrição do dispositivo mais usado na aquisição de dados para obter informação RGB-D, o Kinect. Principalmente, foram explicados os métodos de extração de características e mostrada uma visão geral dos métodos de *clustering* além do funcionamento do modelo *Bag-of-Words*. Finalmente, uma teoria básica dos classificadores SVM e *Random Forests* foi apresentada.

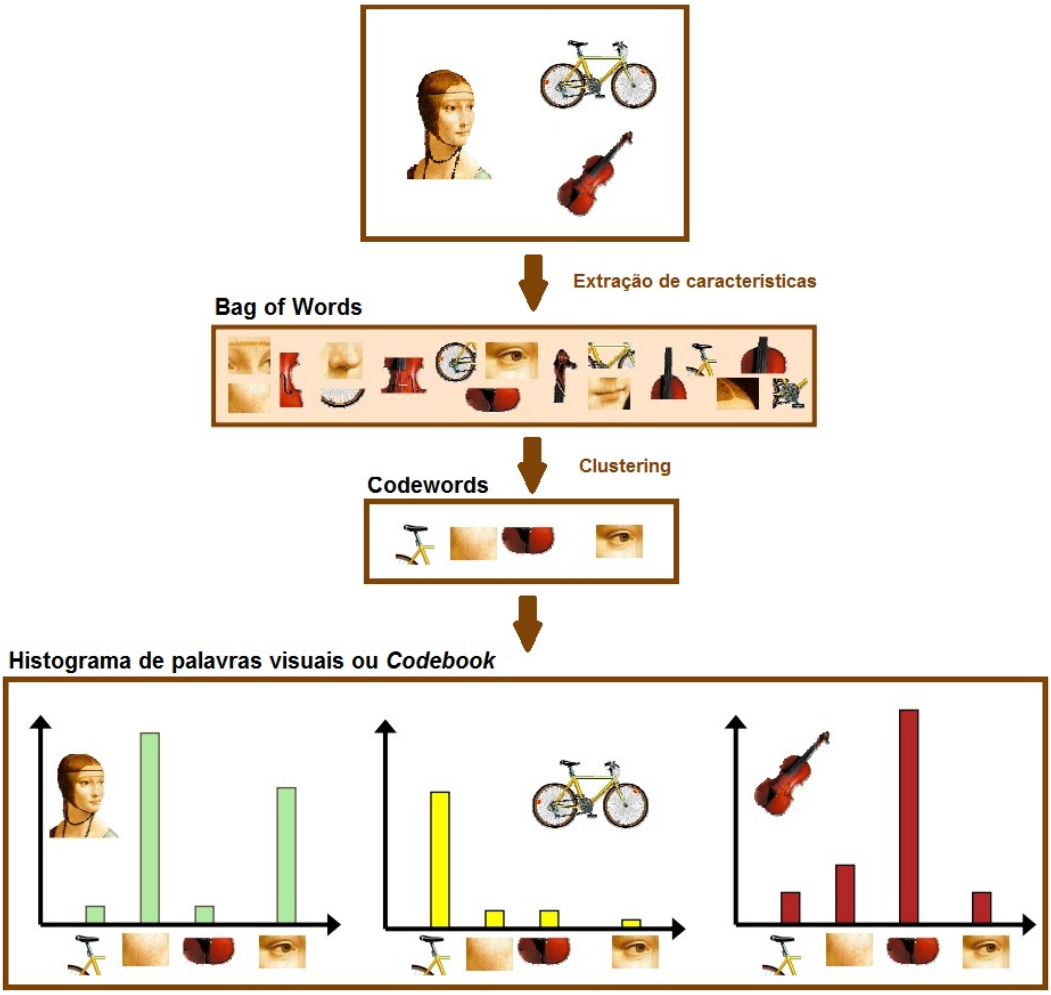

Figura 3.18: Exemplo do modelo Bag of visual words.

Capítulo 4

Modelo Proposto

Neste capítulo, é apresentado o modelo para o reconhecimento de sinais proposto nesta dissertação. Este modelo consiste de três fases: segmentação, aplicação do modelo *Bag-of-Visual-Words* e classificação. Na Seção 4.1 será descrito o modelo de forma geral e em cada seção subsequente serão detalhados cada passo do modelo.

4.1 Modelo proposto para Reconhecimento de Sinais

Esta seção descreve a metodologia desenvolvida para realizar o reconhecimento de sinais estáticos a partir de informações de intensidade e de profundidade. O modelo proposto consiste de três fases como é mostrado na Figura 4.1. Na primeira fase, a área da mão é segmentada utilizando o mapa de profundidade, logo a informação da mão é extraída usando os dados de profundidade e de cor obtidos do sensor Kinect^{TM} . Na segunda fase, a técnica Bag-of-Visual-Words é aplicada, dita técnica consiste das etapas de extração de características, geração do vocabulário visual e geração do histograma de palavras visuais. Na extração de características são usadas informações de intensidade e de profundidade. O descritor SIFT é usado para extrair as características da imagem de intensidades detectando os pontos de interesse e descrevendo-os através da sua vizinhança. O descritor kernel gradiente, aplicado nas imagens de intensidade e de profundidade, é composto por três *kernels*. O *kernel* linear normalizado pesa a contribuição de cada pixel com as magnitudes de gradiente, o *kernel* de orientação calcula a semelhança das orientações do gradiente e, finalmente, o *kernel* Gaussiano de posição mede quão perto dois pixels se encontram no espaço. O vocabulário visual é gerado a partir do

agrupamento por semelhança das características extraídas na fase anterior, o centróide de cada grupo representa uma palavra visual. Logo, o histograma de palavras visuais é obtido contando o número de ocorrências de cada palavra visual. Finalmente, os histogramas são utilizados como entrada para nosso classificador SVM. Este modelo proposto foi publicado no SIBGRAPI 2013 (Otiniano-Rodríguez & Cámara-Chávez 2013).

4.2 Segmentação

O primeiro passo consiste em obter as informações da mão, separando-a do fundo, para realizar essa tarefa é usada uma máscara. Essa máscara é obtida a partir do mapa de profundidade, usando o conceito que os menores níveis de profundidade representam as informações da mão, devido a que esta se encontra mais perto da câmera. O mapa de profundidade é dividido em N grupos. Cada grupo é etiquetado como FG (*foreground*) ou BG (*background*) a partir das seguintes condições:

1. Os centróides dos grupos são ordenados a fim de obter o menor valor de profundidade.
2. Inicialmente, o centróide com menor valor é etiquetado como FG e o resto como BG.
3. Para $i \in (1, 2...N-1)$, onde N representa o número de clusters.
   (a) Se a distância entre $cluster_i$ e $cluster_{(i+1)}$ é menor que t, então $cluster_{(i+1)}$ é etiquetado como FG, senão terminar.

Com os grupos já etiquetados, é criada a máscara onde a etiqueta FG representa valor 1 e BG o valor 0. O sensor Kinect diferencia eficientemente até cinco níveis de profundidade, portanto esse valor determina o número de clusters a ser usado na segmentação. O valor do limiar t depende do banco de dados. O valor utilizado nesta dissertação é de 100, o qual está baseado no banco de dados descrito na Seção 5.1.

Após a máscara ser obtida, é aplicada a operação morfológica de abertura, usando um elemento estruturante quadrado com 3 pixels de largura, para reduzir o ruido presente na máscara. Logo, esta é aplicada sobre as imagens de intensidade e profundidade (ver Figura 4.2, obtendo assim as informações que representam somente a mão da pessoa, como é mostrado na Figura 4.3.

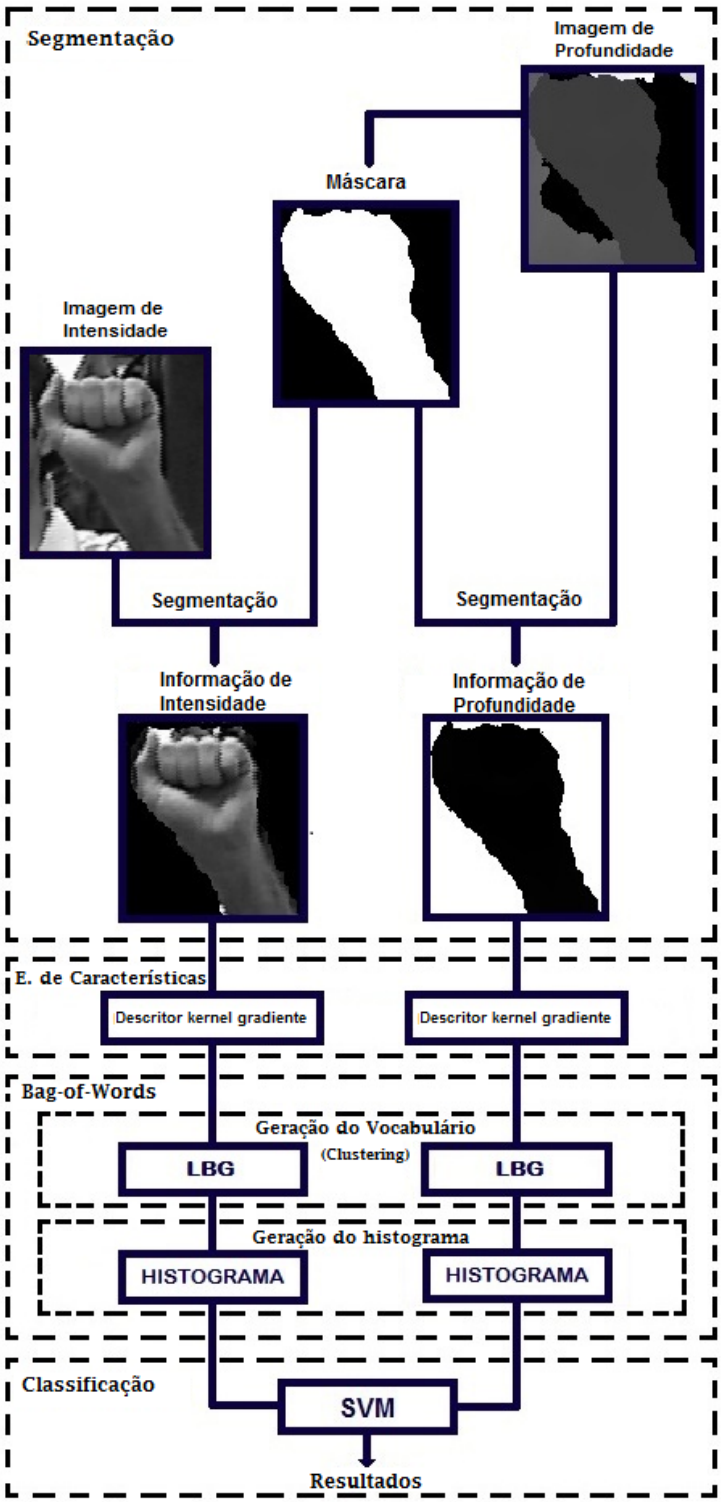

Figura 4.1: Modelo proposto para o reconhecimento de sinais estáticos.

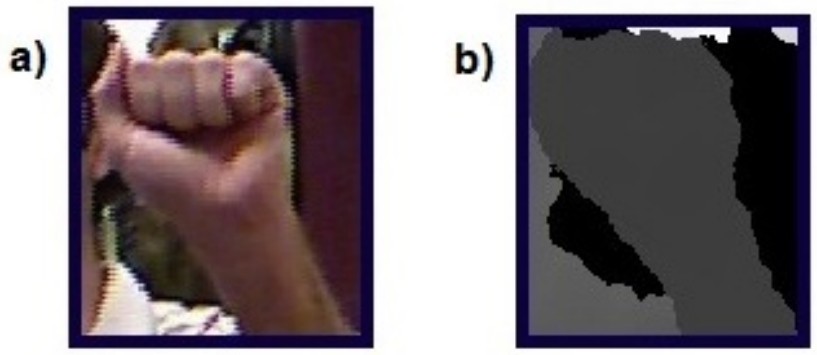

Figura 4.2: Exemplo de imagens de (a) intensidade e (b) profundidade para um mesmo sinal.

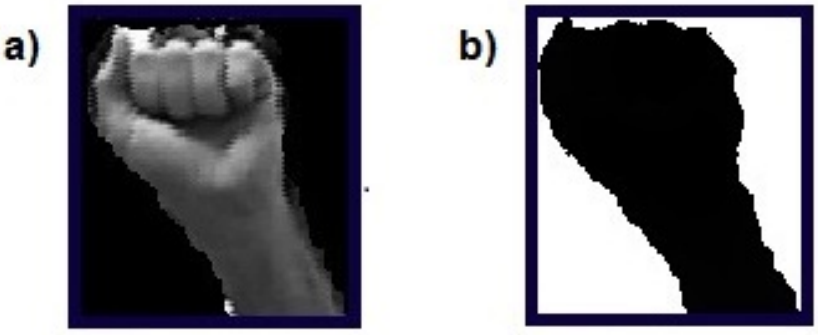

Figura 4.3: Exemplo de imagens de (a) intensidade (b) profundidade após a segmentação.

4.3 Bag-of-Visual-Words

A ideia principal da utilização deste modelo é a obtenção de informação semântica. Os passos desta técnica são: Extração de características, geração do vocabulário visual e a geração do histograma.

Para esta dissertação, as características são obtidas pelo descritor kernel gradiente. Esse descritor é aplicado tanto nas imagens de intensidade quanto nas imagens de profundidade. As características obtidas formam parte do primeiro passo do modelo BoW conforme foi descrito na Seção 3.5. Na Figuras 4.4 e 4.5, são mostrados exemplos das características (magnitudes e orientações) obtidas por este descritor para cada tipo de imagem.

Na geração do vocabulário visual, foi tomada uma amostra do 30% do total de

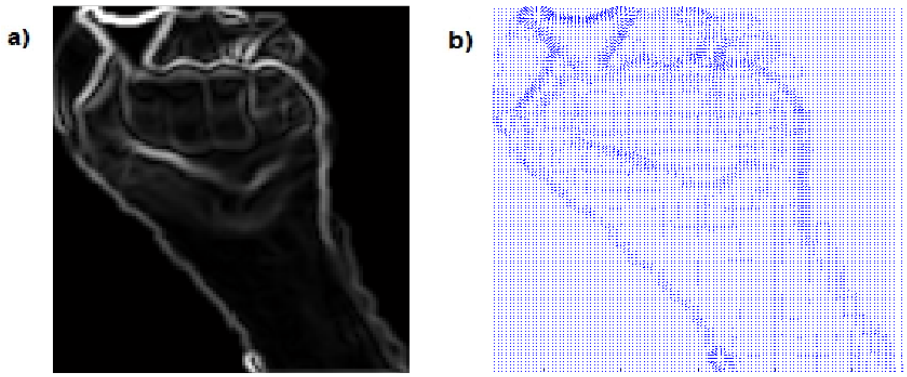

Figura 4.4: Exemplo das (a) magnitudes e (b) orientações dos gradientes obtidos pelo descritor kernel gradiente aplicadas em uma imagem de intensidade.

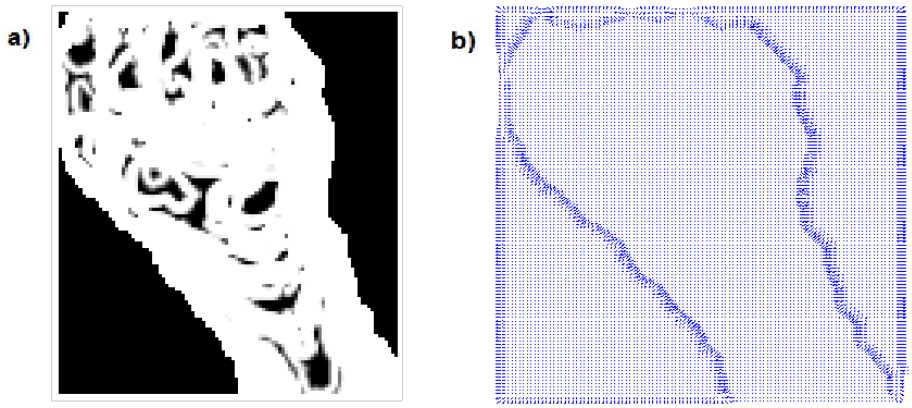

Figura 4.5: Exemplo das (a) magnitudes e (b) orientações dos gradientes obtidos pelo descritor kernel gradiente aplicadas em uma imagem de profundidade.

características. Esta geração é feita usando o algoritmo de agrupamento LBG (Linde-Buzo-Gray) (Linde, Buzo & Gray 1980), com o qual foram calculados 100 grupos. São encontradas palavras visuais ou *codewords* que são representados pelos centróides de cada grupo. Portanto, cada *codeword* representa um grupo com características similares. Este é o processo principal para a obtenção de informação de nível médio.

Logo de ter sido gerado o vocabulário, é criado um histograma contando as ocorrências de cada *codeword* na imagem. Assim, cada histograma de tamanho 100 representa as características para uma imagem.

4.4 Classificação

O método usado na classificação é SVM. Ele foi escolhido devido à alta taxa de acerto nos modelos que o usam nos problemas de reconhecimento em geral. Para o reconhecimento de sinais, SVM também foi usado em alguns modelos conforme foi mostrado no estado da arte (De Souza, Pizzolato & dos Santos Anjo 2012, Ye, Yao & Jiang 2004, Estrela, Cámara-Chávez, Campos, Schwartz & Nascimento 2013). SVM consegue uma boa generalização a partir de um pequeno conjunto de dados. SVM também tem a propriedade de fazer possível a classificação não-linear usando a teoria de kernels sem precisar de um algoritmo específico não-linear. Os kernels são usados para mapear os dados em um espaço de características de alta dimensão. O kernel RBF (*Radial basis function*) foi selecionado para esta dissertação. Ele é o kernel mais popular usado em SVM e permite trabalhar com múltiplas dimensões.

4.5 Considerações Finais

Neste capítulo foi apresentado um modelo para o reconhecimento de sinais. O qual consta de 3 fases: (a) Segmentação baseado nos mapas de profundidade, (b) Aplicação do modelo *Bag-of-Words* para extração de informação semântica, usando os descritores SIFT e kernel gradiente na etapa de extração de características e o algoritmo LBG na geração do vocabulário visual. (c) A classificação dos histogramas de palavras visuais é realizada usando SVM.

Capítulo 5

Experimentos e Resultados

Neste capítulo serão apresentados os experimentos realizados nesta dissertação. Primeiro será descrita a base de dados utilizada, logo definidos os parâmetros do modelo, e finalmente será feita uma avaliação dos resultados.

5.1 Base de dados

A base de dados usada nesta dissertação é: *ASL Finger Spelling Dataset* e encontra-se disponível no site http://personal.ee.surrey.ac.uk/Personal/N.Pugeault/index.php. Essa base contém os sinais que representam as letras do alfabeto, do A a Z, exceto os sinais J e Z. Esses sinais não são usados devido ao fato de que eles têm movimento e este modelo proposto trabalha unicamente com sinais estáticos.

A base de dados contém 500 amostras para cada um dos 24 sinais, gravados de 5 pessoas diferentes (não nativas em língua de sinais), equivalente a um total de 60,000 amostras. Cada amostra tem uma imagem RGB e uma imagem de profundidade, fazendo um total de 120,000 imagens.

As imagens da base de dados apresentam variedade de fundo e de ângulos de visualização. A Figura 5.1, mostra um exemplo de cada sinal para cada executor do sinal, onde é possível ver a variação em tamanho, fundo e orientação dos sinais. Na Figura 5.2, são mostrados exemplos das variações do sinal m, os mesmos foram realizados por diferentes pessoas.

Contudo, devido a essa variação na orientação, quando o sinal é realizado, alguns

Figura 5.1: ASL Finger Spelling Dataset: 24 sinais estáticos a partir de 5 usuários. Exemplo da variedade do conjunto de dados. Esta matriz mostra uma imagem a partir de cada usuário e de cada letra.

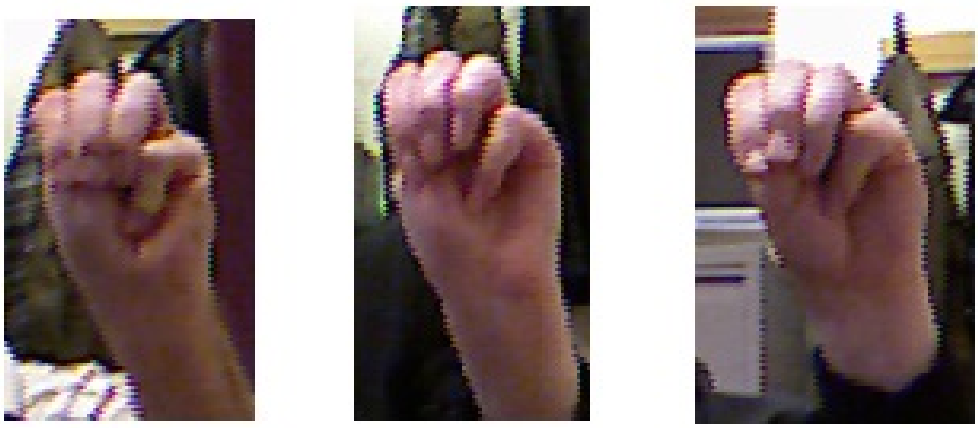

Figura 5.2: Sinal *m* com variações de orientação e fundo.

sinais tornam-se altamente semelhantes. Na Figura 5.3 são mostrados os sinais mais semelhantes: *a*, *e*, *m*, *n*, *s* e *t*. Os exemplos são tomados de uma mesma pessoa. É fácil identificar a semelhança entre esses sinais, todos são representados por um punho fechado, e só diferem pela posição do polegar, levando a níveis mais elevados de confusão. Por conseguinte, devido à grande semelhança, esses sinais são os mais difíceis de diferenciar na tarefa de classificação.

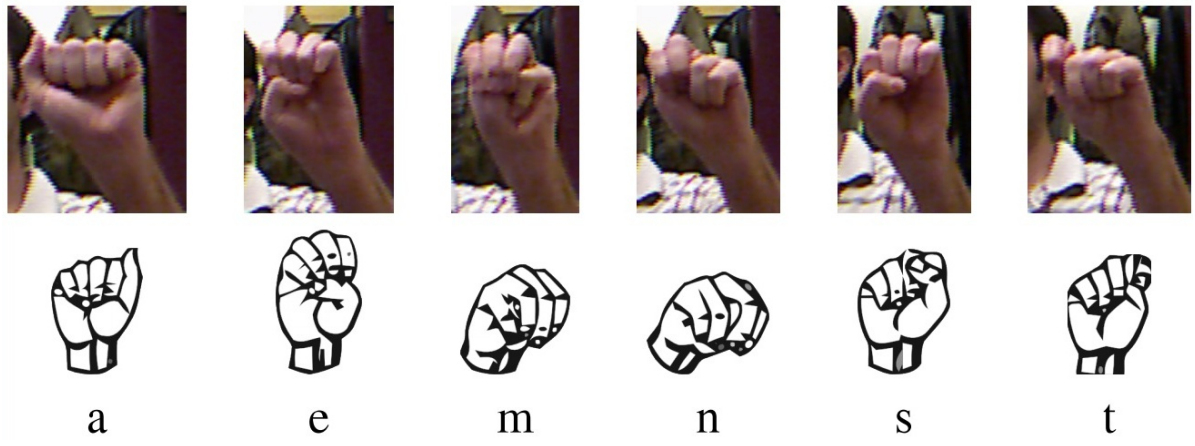

Figura 5.3: Sinais mais conflituosos na base de dados.

5.2 Definição de parâmetros

É preciso definir os valores dos parâmetros para a realização dos experimentos, os quais foram definidos experimentalmente. Os parâmetros utilizados neste trabalho são:

- Para extrair as características de baixo nível usando o descritor kernel gradiente, são usados janelas de 12×13 sobre uma malha regular densa, com espaçamento de 8 pixels (as imagens não têm tamanho uniforme).

- A fim de produzir o vocabulário de palavras visuais, o algoritmo LBG (Linde-Buzo-Gray) (Linde, Buzo & Gray 1980) foi usado para detetar cem clusters tomando uma amostra de 30% do total das características.

- Na fase de classificação, usamos o kernel RBF, cujos valores de g (gama) e c (custo) são 0,25 e 5, respectivamente. Estes valores foram descobertos na realização dos experimentos. Foi usada uma validação cruzada com 5 *folds*. A biblioteca LIBSVM

(*Library for Support Vector Machine*) (Chang & Lin 2011) foi utilizada em nossa implementação.

5.3 Avaliação do modelo

Nossos resultados serão avaliados a partir da comparação dos resultados obtidos com outros modelos usando a mesma base de dados:

Modelo 1 Pugeault & Bowden (2011), criadores da base de dados, usaram no seu modelo Filtros de Gabor para a extração de características e *Random Forests* para a tarefa de classificação. Obtiveram uma acurácia de 75%.

Modelo 2 Zhu & Wong (2012), desenvolveram um método onde primeiro são extraídas as características comuns a nível de janelas, e logo estas são combinadas por meio de descritores kernel. Esta abordagem usa apenas um pequeno número de amostras de treinamento (40 amostras por sinal) e as restantes são usadas para teste. Eles obtiveram 2 resultados, o primeiro de 77% e um segundo de 88%, onde foi acrescentado uma pirâmide de imagens.

Modelo 3 Em (Estrela, Cámara-Chávez, Campos, Schwartz & Nascimento 2013), é apresentado um modelo que utiliza informação RGB-D para o reconhecimento de sinais. Neste modelo são avaliados dois descritores (SIFT e BASE) e dois classificadores (SVM e PLS). Destas comparações, a melhor acurácia obtida foi usando o descritor SIFT e o classificador PLS com 71,51%.

5.4 Experimentos

A fim de validar nosso modelo, foram realizados três tipos de experimentos: usando as características das imagens de intensidade, usando as características das imagens de profundidade e usando ambas características de forma conjunta (RGB-D).

Os resultados serão apresentados em formato de uma matriz de confusão. A diagonal principal da matriz representa os sinais corretamente classificados. E a acurácia final é a média dos valores da diagonal.

5.4.1 Resultados usando características de intensidade

Usando este tipo de informação, foram realizados dois experimentos. O primeiro usando o descritor SIFT e segundo usando o descritor kernel gradiente.

Experimento 1: SIFT

A acurácia média obtida neste experimento é de 63%. A matriz de confusão é encontrada na Tabela 5.1. Nesta tabela, pode-se observar que os sinais *n*, *r*, *k* e *x* têm as acurácias médias mais baixas (entre 50% e 54%), e o sinal *h* tem a acurácia média mais alta (82%). Isso mostra a ampla variação dos resultados para esse experimento e significa que pode não existir suficiente informação. Como é mostrado na Figura 5.4, os pontos de interesse ficam concentrados em uma parte da mão, não descrevendo a mão completamente.

Tabela 5.1: Matriz de confusão da classificação dos 24 sinais a partir de informação de intensidade usando o descritor SIFT.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.67	0.01	0.04	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.03	0.02	0.02	0.00	0.01	0.01	0.02	0.03	0.01	0.00	0.00	0.01	0.01
b	0.02	0.74	0.02	0.01	0.01	0.02	0.00	0.00	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.01	0.03	0.02	0.01	0.01	0.01
c	0.03	0.02	0.67	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.03	0.02	0.01	0.02	0.02	0.00	0.02	0.02	0.01	0.02	0.02
d	0.01	0.02	0.02	0.55	0.03	0.03	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.03
e	0.02	0.01	0.02	0.04	0.58	0.02	0.01	0.00	0.02	0.01	0.02	0.03	0.03	0.02	0.03	0.02	0.01	0.02	0.03	0.01	0.01	0.01	0.03	0.02
f	0.00	0.01	0.01	0.02	0.02	0.72	0.00	0.01	0.03	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.05	0.01	0.02
g	0.01	0.00	0.02	0.01	0.01	0.02	0.71	0.07	0.01	0.03	0.01	0.00	0.01	0.02	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
h	0.00	0.00	0.01	0.01	0.01	0.00	0.07	0.82	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01
i	0.02	0.02	0.02	0.02	0.02	0.03	0.01	0.00	0.66	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
k	0.02	0.01	0.03	0.03	0.02	0.02	0.03	0.01	0.04	0.54	0.03	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.01	0.02	0.03	0.02	0.02	0.02
l	0.02	0.01	0.02	0.02	0.01	0.01	0.00	0.00	0.02	0.02	0.73	0.00	0.01	0.01	0.01	0.00	0.02	0.01	0.01	0.02	0.01	0.00	0.01	0.04
m	0.04	0.01	0.01	0.01	0.03	0.01	0.01	0.00	0.02	0.00	0.00	0.59	0.06	0.04	0.02	0.03	0.00	0.04	0.05	0.00	0.00	0.00	0.01	0.00
n	0.04	0.02	0.01	0.02	0.04	0.01	0.01	0.00	0.03	0.01	0.01	0.07	0.50	0.02	0.02	0.02	0.01	0.03	0.06	0.01	0.00	0.01	0.03	0.01
o	0.02	0.01	0.04	0.02	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.04	0.03	0.57	0.02	0.03	0.01	0.04	0.02	0.01	0.01	0.00	0.01	0.01
p	0.01	0.01	0.01	0.02	0.03	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.67	0.07	0.01	0.00	0.02	0.01	0.00	0.00	0.02	0.01
q	0.01	0.02	0.01	0.01	0.03	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.03	0.08	0.63	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.01
r	0.02	0.03	0.03	0.04	0.02	0.03	0.01	0.01	0.02	0.04	0.03	0.00	0.01	0.01	0.01	0.00	0.51	0.01	0.00	0.06	0.06	0.03	0.02	0.02
s	0.04	0.01	0.04	0.02	0.03	0.00	0.01	0.00	0.01	0.01	0.02	0.03	0.03	0.04	0.01	0.01	0.01	0.59	0.02	0.01	0.00	0.00	0.02	0.01
t	0.04	0.02	0.01	0.02	0.03	0.01	0.01	0.00	0.02	0.01	0.01	0.04	0.06	0.02	0.03	0.03	0.01	0.03	0.56	0.01	0.01	0.00	0.01	0.01
u	0.01	0.04	0.04	0.02	0.01	0.02	0.00	0.00	0.02	0.02	0.02	0.00	0.01	0.01	0.01	0.00	0.06	0.01	0.01	0.59	0.06	0.03	0.01	0.01
v	0.01	0.02	0.03	0.02	0.01	0.03	0.01	0.00	0.01	0.03	0.02	0.00	0.00	0.01	0.00	0.00	0.07	0.00	0.00	0.06	0.58	0.07	0.01	0.01
w	0.00	0.01	0.01	0.02	0.01	0.05	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.02	0.06	0.72	0.00	0.01
x	0.02	0.01	0.03	0.03	0.03	0.01	0.01	0.00	0.02	0.03	0.02	0.01	0.03	0.01	0.04	0.01	0.03	0.02	0.02	0.02	0.01	0.01	0.54	0.03
y	0.01	0.01	0.03	0.03	0.02	0.03	0.01	0.00	0.03	0.02	0.05	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.59

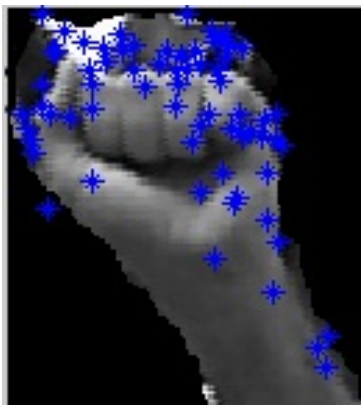

Figura 5.4: Exemplo dos pontos característicos de uma imagem de intensidade obtidos pelo descritor SIFT.

Experimento 2: Descritor kernel gradiente

Usando este descritor sobre as imagens de intensidade, é obtida uma acurácia média de 89%. Na Tabela 5.2, são encontrados os resultados para cada um dos 24 sinais. A comparação dos resultados do primeiro experimento usando SIFT, o sinais n, r, k e x incrementaram suas médias a valores entre 77% e 90%. O que significa um aumento sobre o 27% para esses sinais. De forma geral, foram obtidos melhores resultados usando este descritor, obtendo um aumento do 26%. Acredita-se que isso acontece devido a que este descritor utiliza uma malha regular descrevendo a mão completamente.

5.4.2 Resultados usando características de profundidade

Para este tipo de imagens foi utilizado o descritor de kernel gradiente. Os resultados são mostrado a seguir.

Experimento 3: Descritor kernel gradiente

Neste experimento, o descritor kernel gradiente é usado para extrair as características das imagens de profundidade. A acurácia obtida é de 86%. A Tabela 5.3, mostra a matriz de confusão da classificação usando este tipo de informação. Os sinais têm uma média sobre o 68%, sendo os sinais t, n, r e s os que obtêm menores resultados (entre 68% e 76%). Por outro lado, temos os sinais b e l com os melhores resultados, ambos com 95% de acerto. Foram obtidos esses resultados devido à informação 3D que se encontra

Tabela 5.2: Matriz de confusão da classificação dos 24 sinais a partir de informação de intensidade usando o descritor kernel gradiente.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.94	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00
b	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c	0.00	0.00	0.95	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00
e	0.01	0.00	0.01	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
f	0.00	0.01	0.01	0.00	0.00	0.96	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.09	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.94	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
k	0.00	0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.91	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00
l	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
m	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.05	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00
n	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.06	0.77	0.01	0.01	0.00	0.01	0.01	0.05	0.01	0.00	0.00	0.01	0.00
o	0.00	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.87	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.91	0.04	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.92	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
r	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.03	0.01	0.00	0.01	0.01	0.00	0.00	0.78	0.00	0.00	0.03	0.02	0.00	0.01	0.00
s	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.02	0.02	0.00	0.00	0.00	0.84	0.03	0.00	0.00	0.00	0.00	0.00
t	0.03	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.04	0.01	0.01	0.00	0.00	0.02	0.81	0.00	0.00	0.00	0.02	0.00
u	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.87	0.02	0.00	0.00	0.00
v	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.02	0.86	0.02	0.00	0.00
w	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.96	0.00	0.00
x	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.01	0.00	0.01	0.00	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.86	0.00
y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93

nos mapas de profundidade. Além disso, o processo de segmentação é realizado de forma mais exata já que a máscara é obtida a partir da imagem de profundidade.

5.4.3 Resultados usando características RGB-D

A fim de trabalhar com ambas informações (intensidade e profundidade). São realizados dois experimentos. O primeiro consiste em aplicar o descritor SIFT sobre as imagens de intensidade e o descritor kernel gradiente sobre as imagens profundidade. No segundo experimento, o descritor kernel gradiente é aplicado tanto nas imagens de intensidade quanto de profundidade. Nos dois casos, são usadas as características obtidas nos experimentos anteriores. Essas características (histogramas de palavras visuais) de intensidade e de profundidade são combinadas, obtendo assim um vector de características com informações RGB-D.

Tabela 5.3: Matriz de confusão da classificação dos 24 sinais a partir de informação de profundidade usando o descritor kernel gradiente.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.89	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.01	0.00
b	0.01	0.95	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
c	0.01	0.00	0.92	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
d	0.00	0.00	0.00	0.87	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.00	0.01	0.00
e	0.02	0.00	0.02	0.01	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00
f	0.00	0.02	0.01	0.01	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
g	0.01	0.00	0.00	0.00	0.00	0.00	0.86	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.94	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
k	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.84	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.01	0.01	0.01	0.00	0.01	0.00
l	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
m	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.07	0.01	0.00	0.00	0.01	0.05	0.04	0.00	0.00	0.00	0.00	0.00
n	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.07	0.74	0.02	0.00	0.00	0.00	0.02	0.07	0.00	0.00	0.00	0.02	0.00
o	0.01	0.00	0.02	0.01	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.01	0.78	0.01	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00
p	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.88	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
q	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	0.01	0.01	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.76	0.00	0.01	0.05	0.03	0.00	0.01	0.00
s	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.03	0.03	0.00	0.00	0.00	0.76	0.04	0.00	0.00	0.00	0.01	0.00
t	0.06	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.08	0.02	0.00	0.00	0.00	0.05	0.68	0.00	0.00	0.00	0.02	0.00
u	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.82	0.03	0.02	0.01	0.00
v	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.03	0.83	0.04	0.00	0.00
w	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.04	0.91	0.00	0.00
x	0.01	0.00	0.01	0.03	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.03	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.78	0.00
y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.94

Experimento 4: SIFT e descritor kernel gradiente

Usando esses dois descritores, um para cada tipo de informação, é obtida uma acurácia média de 91,26%. A Tabela 5.4, mostra os resultados obtidos neste experimento. Os sinais *b*, *c*, *f*, *i*, *l* e *y* têm as mais altas acurácias médias (sobre 95%). Já, os sinais *n*, *r* e *t* têm os valores mais baixos: 84% e 82%, respectivamente. O baixo valor de reconhecimento do sinal *n* (82%) é devido à grande semelhança com sinais *m* e *t*, como foi mostrado na Figura 5.3. No entanto, este mesmo valor teve um aumento de 32% e 8% sobre os resultados dos experimentos 1 e 3, respectivamente.

Tabela 5.4: Matriz de confusão da classificação dos 24 sinais a partir de informação RGB-D usando os descritores SIFT e kernel gradiente.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00
b	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
d	0.00	0.00	0.00	0.91	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
e	0.01	0.00	0.01	0.01	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00
f	0.00	0.01	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.96	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
k	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.00
l	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
m	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.05	0.01	0.00	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00
n	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.06	0.82	0.01	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.00	0.01	0.00
o	0.01	0.00	0.01	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.87	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.93	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	0.01	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.03	0.02	0.00	0.00	0.00
s	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.00	0.00	0.00	0.88	0.01	0.00	0.00	0.00	0.01	0.00
t	0.03	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.05	0.01	0.00	0.00	0.00	0.02	0.82	0.00	0.00	0.00	0.01	0.00
u	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.88	0.02	0.01	0.01	0.00
v	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.90	0.02	0.00	0.00
w	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.95	0.00	0.00
x	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.87	0.00
y	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96

Experimento 5: Descritor kernel gradiente

Para este experimento, o descritor kernel gradiente foi aplicado tanto nas imagens de intensidade quanto de profundidade. A acurácia média obtida é de 95,63%, significando

a acurácia média mais alta de todos os experimentos realizados. Na Tabela 5.5, são encontrados os resultados da classificação dos 24 sinais. A partir desses resultados, é possível observar que os sinais *n* e *t* têm as menores acurácias, 90% em ambos casos. Contudo, os 22 sinais restantes, têm uma média superior a 90%. Como é o caso dos sinais *b* e *l*, que têm as mais altas acurácias (99%). Estes resultados mostram um aumento da acurácia em comparação aos experimentos 2 e 3, devido a que ambas informações são combinadas, obtendo assim um descritor mais completo.

Tabela 5.5: Matriz de confusão da classificação dos 24 sinais a partir de informação RGB-D usando o descritor kernel gradiente.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
b	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
e	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
f	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
k	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00
l	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
m	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.02	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00
n	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.90	0.01	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.01	0.00
o	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.02	0.01	0.00	0.01	0.00
s	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.93	0.01	0.00	0.00	0.00	0.00	0.00
t	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01	0.90	0.00	0.00	0.00	0.01	0.00
u	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.93	0.01	0.00	0.00	0.00
v	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.94	0.02	0.00	0.00
w	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.98	0.00	0.00
x	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.95	0.00
y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98

A Tabela 5.6 mostra os resultados da classificação usando informação RGB-D e o descritor kernel gradiente variando as quantidades de amostras para treinamento e teste. Para obter a acurácia média este experimento foi executado 20 vezes para cada diferente porcentagem. A finalidade de este experimento é avaliar o nosso modelo com os diferentes modelos proposto na Seção 5.3. No modelo (Zhu & Wong 2012) são usadas 40 amostras de cada sinal para teste. Em nosso caso, o 10% representa 50 amostras de cada sinal por

Tabela 5.6: Acurácia e desvio padrão da classificação usando informação RGB-D com diferentes quantidades de treinamento e teste.

% Treinamento	% Teste	Acurácia	Desvio Padrão
10	90	88.85%	0.14
20	80	91.07%	0.20
30	70	91.93%	0.18
40	60	93.12%	0.13
50	50	93.67%	0.19
80	20	95.63%	0.15

pessoa, o que faz os métodos comparáveis. Os resultados mostram que nosso modelo com 10% de amostras para treinamento é o menor que os demais experimentos, no entanto o resultado não é insignificante por que tem uma acurácia média de 88.85%. Usando 50% de amostras para treinamento é obtida uma acurácia média de 93,67% superando o resultado do método proposto por Pugeault & Bowden (2011) que obteve 75% de acurácia.

5.5 Análise de Resultados

A Tabela 5.7, mostra os resultados dos cinco experimentos realizados e dos três modelos mencionados na Seção 5.3. Esta tabela inclui a acurácia média e o desvio padrão para cada experimento. Podemos ver que os experimentos usando informação RGB-D obtêm os melhores resultados, superando aos demais experimentos que usam informação de forma separada. Os resultados usando informação RGB-D também superam na acurácia aos modelos propostos por Pugeault & Bowden (2011), Zhu & Wong (2012) e Estrela, Cámara-Chávez, Campos, Schwartz & Nascimento (2013). Os quais são encontrados no estado da arte e usam a mesma base de dados no seus experimentos. Deve ser notado que o modelo proposto por Zhu & Wong (2012) tem um menor número de amostras para treinamento, 40 amostras para cada sinai, pelo qual foram realizados os experimentos da Tabela 5.6 onde foi mostrado que nossos resultados são comparáveis a esse modelo com uma acurácia média de 88,85%. Dessa mesma tabela, usando 50% de amostras para treinamento é obtida uma acurácia média de 93,67% superando o resultado do método proposto por Pugeault & Bowden (2011) que obteve 75% de acurácia.

Tabela 5.7: Acurácia e desvio padrão de cada um dos 5 experimentos.

Experimento	Método	Tipo de Inf.	Acurácia	Desvio Padrão
1	SIFT	RGB	62,70%	0,47
2	Kernel	RGB	89,56%	0,23
3	Kernel	Depth	85,18%	0,16
4	SIFT & Kernel	RGB-D	91,26%	0,18
5	Kernel & Kernel	RGB-D	95,63%	0,15
Pugeault et al.	Gabor	RGB-D	75,00%	-
Zhu et al.	SIFT & LBP	RGB-D	88,94%	-
Estrela et al.	PLS	RGB-D	71,51%	-

Tendo em conta todos os experimentos, os resultados mostram claramente as altas taxas de acerto que obteve nosso modelo (RGB-D). Além da baixa variação do desvio padrão quando a informação RGB-D é utilizada. Tornando-se em um modelo promissor para o reconhecimento de sinais estáticos. Esses resultados foram produto da combinação de informação de intensidade e de profundidade de cada sinal. Por causa disso, cada sinal foi representado com mais detalhe. As imagens de profundidade, além de conter características do sinal, elas facilitaram o processo da segmentação, permitindo concentrar-nos nas demais etapas do modelo.

5.6 Considerações Finais

Neste capítulo, a base de dados usada nos experimentos foi descrita detalhadamente. Foram realizados cinco experimentos divididos em 3 grupos: Usando informação de intensidade, de profundidade e de ambas informações (RGB-D). Logo, foram mostrados e analisados os resultados para cada um dos experimentos. Finalmente os resultados da classificação usando informações RGB-D obtiveram os melhores resultados, atingindo alguns sinais 99%.

Capítulo 6

Conclusões

Reconhecimento de língua de sinais é um tema que se encontra atualmente em desenvolvimento. Existem abordagens que variam na técnica de obtenção da informação. Desde as mais antigas que utilizam luvas coloridas ou de dados, outras, câmeras simples e os métodos mais atuais que utilizam o sensor Kinect ou câmeras ToF. A partir desses dispositivos são obtidas diversas informações. As abordagens mais comuns utilizam informação de imagens RGB com fundos uniformes para facilitar o processo de segmentação. Através de luvas de dados, é possível obter informação das posições das mãos, facilitando o processo quando se trata de sinais dinâmicos. O sensor Kinect permite obter tanto a imagem RGB quanto uma imagem de profundidade, que representa a distância entre cada parte do corpo e o Kinect.

Nesta dissertação foram usadas as imagens de intensidade e profundidade. Esta última, além de ser usada para extrair informação da mesma e definir as características, foi utilizada no processo de segmentação, facilitando esta tarefa.

Os resultados obtidos nos experimentos, mostraram que usando informação de intensidade e de profundidade conjuntamente foram obtidas as mais altas acurácias. Deste modo, aumentando a acurácia de cada um dos experimentos que usaram as informações de intensidade e profundidade de forma separada. Nossos resultados também foram comparados a outras abordagens que usaram a mesma base de dados no seus experimentos. Essa comparação demonstrou que nossos resultados também obtiveram maior acurácia média que os outros modelos, confirmando a eficácia de nosso modelo (Otiniano-Rodríguez & Cámara-Chávez 2013). Os resultados obtidos, além de ser em grande parte pela utilização de informação RGB-D, também é por causa da aplicação do descri-

tor kernel gradiente, o qual conseguiu extrair características de melhor qualidade em comparação ao descritor SIFT. Além disso, a técnica Bag-of-Words permitiu buscar informação semântica tornando as características mais robustas, permitindo assim ao classificador SVM fazer uma melhor e mais rápida diferenciação entre sinais.

6.1 Trabalhos Futuros

Em consequência deste trabalho de dissertação, consideramos os seguintes trabalhos futuros:

- Testar nosso método com outras bases de dados. (Criar uma base de dados com imagens RGB e de profundidade).
- Testar outros tipos de descritores ou outros tipos de kernels sobre as imagens de profundidade.
- Estender nosso método para sinais dinâmicos usando informação das posições da mão.

Referências Bibliográficas

©2013 Microsoft Corporation (2013). Kinect for Windows.

Biswas, K. & Basu, S. K. (2011). Gesture recognition using microsoft kinect®, *Proceedings of the 2011 5th International Conference on Automation, Robotics and Applications (ICARA)*, IEEE, pp. 100–103.

Bo, L., Ren, X. & Fox, D. (2010). Kernel descriptors for visual recognition, *Advances in Neural Information Processing Systems* **7**.

Bo, L., Ren, X. & Fox, D. (2011). Depth kernel descriptors for object recognition, *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 821–826.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, pp. 144–152.

Brashear, H., Henderson, V., Park, K.-H., Hamilton, H., Lee, S. & Starner, T. (2006). American sign language recognition in game development for deaf children, *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, ACM, pp. 79–86.

Brashear, H., Zafrulla, Z., Starner, T., Hamilton, H., Presti, P. & Lee, S. (2010). Copycat: A corpus for verifying american sign language during game play by deaf children, *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 27–32.

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**(3): 1–27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chang, Y.-J., Chen, S.-F. & Huang, J.-D. (2011). A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities, *Research in developmental disabilities* **32**(6): 2566–2570.

Cooper, H. & Bowden, R. (2007). Large lexicon detection of sign language, *Human–Computer Interaction*, Springer, pp. 88–97.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks, *Machine Learning* **20**(3): 273–297.

De Souza, C. R., Pizzolato, E. B. & dos Santos Anjo, M. (2012). Fingerspelling recognition with support vector machines and hidden conditional random fields, *Advances in Artificial Intelligence-IBERAMIA 2012*, Springer, pp. 561–570.

Doliotis, P., Athitsos, V., Kosmopoulos, D. & Perantonis, S. (2012). Hand shape and 3d pose estimation using depth data from a single cluttered frame, *Advances in Visual Computing*, Springer, pp. 148–158.

dos Santos Anjo, M., Pizzolato, E. B. & Feuerstack, S. (2012). A real-time system to recognize static gestures of brazilian sign language (Libras) alphabet using kinect, *Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems*, Brazilian Computer Society, pp. 259–268.

Espaço das LIBRAS (2011). Alfabeto Manual - Em Língua Brasileira de Sinais, `http://espacodaslibras.blogspot.com/2011/07/alfabeto-manual-em-lingua-brasileira-de.html`.

Estrela, B. N., Cámara-Chávez, G., Campos, M. F., Schwartz, W. R. & Nascimento, E. R. (2013). Sign language recognition using partial least squares and RGB-D information, *Proceedings of the IX Workshop de Visão Computacional (WVC 2013)*.

Fanelli, G., Gall, J. & Gool, L. V. (2011). Real time head pose estimation with random regression forests, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 617–624.

Frati, V. & Prattichizzo, D. (2011). Using Kinect for hand tracking and rendering in wearable haptics, *Proceedings of the IEEE World Haptics Conference (WHC)*, IEEE, pp. 317–321.

Hienz, H., Bauer, B. & Karl-Friedrich, K. (2000). HMM-based continuous sign language recognition using stochastic grammars, *Lecture Notes in Computer Science* pp. 185–196.

Holden, E.-J. & Owens, R. (2001). Visual sign language recognition, *Multi-Image Analysis*, Springer, pp. 270–287.

Li, Y. (2012). Hand gesture recognition using Kinect, *Proceedings of the 3rd IEEE International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, pp. 196–199.

Liang, R.-H. & Ouhyoung, M. (1998). A real-time continuous gesture recognition system for sign language, *Proceedings oh the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998.*, IEEE, pp. 558–567.

Linde, Y., Buzo, A. & Gray, R. (1980). An algorithm for vector quantizer design, *IEEE Transactions on Communications* **28**(1): 84–95.

Lowe, D. G. (1999). Object recognition from local scale-invariant features, *Proceedings of the seventh IEEE International Conference on Computer vision, 1999.*, Vol. 2, IEEE, pp. 1150–1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* **60**(2): 91–110.

Lütkebohle, I. (2013). World Health Organization, `http://www.who.int/mediacentre/factsheets/fs300/en/`.

Mo, Z. & Neumann, U. (2006). Real-time hand pose recognition using low-resolution depth images, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, IEEE, pp. 1499–1505.

Oikonomidis, I., Kyriazis, N. & Argyros, A. (2011). Efficient model-based 3D tracking of hand articulations using Kinect, *Proceedings of the British Machine Vision Conference*, BMVA Press, pp. 101.1–101.11.

Otiniano-Rodríguez, K. & Cámara-Chávez, G. (2013). Finger spelling recognition from RGB-D information using kernel descriptor, *Proceedings of the SIBGRAPI 2013 (XXVI Conference on Graphics, Patterns and Images)*.

Piater, J., Hoyoux, T. & Du, W. (2010). Video analysis for continuous sign language recognition, *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 22–23.

Pizzolato, E. B., dos Santos Anjo, M. & Pedroso, G. C. (2010). Automatic recognition of finger spelling for libras based on a two-layer architecture, *Proceedings of the 2010 ACM Symposium on Applied Computing*, ACM, pp. 969–973.

Pugeault, N. & Bowden, R. (2011). Spelling it out: Real-time ASL fingerspelling recognition., *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE, pp. 1114–1119.

Ren, Z., Yuan, J. & Zhang, Z. (2011). Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera, *Proceedings of the 19th ACM international conference on Multimedia*, ACM, pp. 1093–1096.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A. W., Finocchio, M., Blake, A., Cook, M. & Moore, R. (2013). Real-time human pose recognition in parts from single depth images, *Communications of the ACM* **56**(1): 116–124.

Sivic, J. & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos, *Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003.*, IEEE, pp. 1470–1477.

Starner, T., Weaver, J. & Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **20**(12): 1371–1375.

Uebersax, D., Gall, J., den Bergh, M. V. & Gool, L. J. V. (2011). Real-time sign language letter and word recognition from depth data, *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 383–390.

Vamplew, P. (1996). *Recognition of sign language using neural networks*, PhD thesis, University of Tasmania.

Von Agris, U., Knorr, M. & Kraiss, K.-F. (2008). The significance of facial features for automatic sign language recognition, *Proceedings of the 8th IEEE International*

Conference on Automatic Face & Gesture Recognition, 2008. FG'08., IEEE, pp. 1–6.

Wang, Q., Chen, X., Zhang, L.-G., Wang, C. & Gao, W. (2007). Viewpoint invariant sign language recognition, *Computer Vision and Image Understanding* **108**(1): 87–97.

William Vicars, E. (97-2013a). ASL University TM, `http://www.lifeprint.com/asl101/topics/wallpaper1.htm`.

William Vicars, E. (97-2013b). ASL University TM, `http://lifeprint.com/asl101/pages-layout/parameters.htm`.

Ye, J., Yao, H. & Jiang, F. (2004). Based on hmm and svm multilayer architecture classifier for chinese sign language recognition with large vocabulary, *Proceedings oh the Third International Conference on Image and Graphics, 2004.*, IEEE, pp. 377–380.

Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H. & Presti, P. (2011). American sign language recognition with the Kinect, *Proceedings of the 13th international conference on multimodal interfaces*, ACM, pp. 279–286.

Zahedi, M. & Manashty, A. R. (2011). Robust sign language recognition system using ToF depth cameras, *arXiv preprint arXiv:1105.0699* .

Zhu, X. & Wong, K.-Y. K. (2012). Single-frame hand gesture recognition using color and depth kernel descriptors, *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 2989–2992.