Sílvia Grasiella Moreira Almeida

Extração de Características em Reconhecimento de Parâmetros Fonológicos da Língua Brasileira de Sinais utilizando Sensores RGB-D

Belo Horizonte-MG

Setembro/2014

Sílvia Grasiella Moreira Almeida

Extração de Características em Reconhecimento de Parâmetros Fonológicos da Língua Brasileira de Sinais utilizando Sensores RGB-D

Tese submetida à Escola de Engenharia da Universidade Federal de Minas Gerais como requisito para obtenção do título de Doutor no Programa de Pós-Graduação em Engenharia Elétrica.

Orientador Frederico Gadelha Guimarães

Coorientador: Jaime Arturo Ramírez

Universidade Federal de Minas Gerais – UFMG

Programa de Pós-Graduação em Engenharia Elétrica – PPGEE

Belo Horizonte-MG

Setembro/2014

Sílvia Grasiella Moreira Almeida

Extração de Características em Reconhecimento de Parâmetros Fonológicos da Língua Brasileira de Sinais utilizando Sensores RGB-D/ Sílvia Grasiella Moreira Almeida. – Belo Horizonte-MG, Setembro/2014-

144 p. : il. (algumas color.) ; 30 cm.

Orientador Frederico Gadelha Guimarães

Tese – Universidade Federal de Minas Gerais – UFMG

Programa de Pós-Graduação em Engenharia Elétrica – PPGEE, Setembro/2014.

1. Extração de características. 2. Reconhecimento de Língua Brasileira de Sinais. 3. Sensores RGB-D. I. Frederico Gadelha Guimarães. II. Universidade Federal de Minas Gerais. III. Escola de Engenharia. IV. Extração de Características em Reconhecimento de Parâmetros Fonológicos da Língua Brasileira de Sinais e utilizando Sensores RGB-D.

CDU

Dedicatória

Este trabalho é dedicado a todas as pessoas que acreditam que a educação é capaz de transformar uma sociedade.

Agradecimentos

Agradeço primeiramente ao tempo que, de forma tão sábia, permite que o amadurecimento científico e profissional possa ser alcançado.

Em seguida agradeço a todos que estiveram comigo ao longo deste tempo. Ao meu grande amigo e orientador Frederico Gadelha Guimarães, que já perdeu a conta de quantos “obrigada” eu já lhe dirigi ao longo da nossa parceria! Somem-se a todos eles mais um “Obrigada!”. Agradeço à paciência e orientações valiosíssimas do meu co-orientador Jaime Arturo Ramírez, uma das poucas pessoas capazes de me ajudar a resolver a minha confusão mental.

Agradeço a todos os meus professores, pessoas que tanto admiro e respeito. Não é à toa que tento me tornar uma professora e espelho-me em cada um que já passou pela minha vida.

Um especial agradecimento aos meus colegas de trabalho da Automação Industrial em Ouro Preto, a minha segunda casa. Os tempos da educação não são fáceis e admiro a força e paciência que vocês demonstram ao longo de toda uma vida dedicada aos ensinamentos. E vão além, transformando a vida daqueles que tem a felicidade de conviver com vocês.

Agradeço a meus alunos de BIC-Jr, rapazes ainda e que sempre estiveram comigo nas gravações dos sinais, nas apresentações das Semanas de Tecnologia e em todas as demandas em que foram solicitados. Obrigada Cássio, Gean e Matheus! E a todos os outros que participaram dos projetos e ideias que envolvem a Libras na escola.

Agradeço ao Instituto Federal Minas Gerais que, diante da minha urgência mental em pesquisar sobre este assunto, permitiu-me fazê-lo e me apoiou em tudo que foi possível ao longo destes 4 poucos e ao mesmo tempo longos anos.

Agadeço aos professores David Lowther e Frank Ferrie, professors da McGill University, pelo acolhimento, disciplina e grandes ideias compartilhadas comigo em minha breve passagem pelo rigoroso inverno canadense.

Agradeço ao Ricardo Prado e ao Alan de Freitas, indispensáveis na condução deste trabalho.

Um agradecimento todo especial a duas das pessoas que me fizeram trilhar este assunto tão instigante: a amiga Andréia Chagas Rocha Toffolo, minha professora

e tutora em Libras e a minha irmã, Priscilla, escritora e conselheira em caminhos a seguir dentro do projeto. Ambas são parte desta pesquisa, sempre me apoiando, incitando e com as quais conto na continuidade desta.

Enfim, amigos e família! Não há espaço aqui para dizer-lhes o quanto cada um esteve presente nestes momentos. Citá-los nominalmente não será ato capaz de representar o quanto sou grata por ter cada um de vocês em vários ou rápidos momentos da minha vida. Agradeço a todos pela tolerância, pela compreensão, pelo apoio.

E aqueles que leram os agradecimentos até aqui, eu agradeço! Sinal que são mesmo meus amigos! Obrigada!

Resumo

A extração de características em vídeo para reconhecimento de línguas de sinais é um problema que apresenta uma grande quantidade de desafios. O presente trabalho tem como proposta apresentar uma metodologia para extração de características em Língua Brasileira de Sinais que trata parte destes desafios. Nesta metodologia exploram-se parâmetros da própria estrutura fonológica da Língua Brasileira de Sinais. Um sensor RGB-D é utilizado para gravar sinais pré-estabelecidos e fornece como informação imagens de intensidade de cor em formato RGB, imagens de profundidade e imagens que marcam as posições do corpo humano. Destas informações obtidas por meio do sensor RGB-D são extraídas sete características baseadas em visão. Cada uma delas está relacionada a um, dois ou três parâmetros fonológicos estruturais da Língua Brasileira de Sinais. Investiga-se esta relação entre as características extraídas e os parâmetros estruturais baseados em forma, movimento e posição das mãos. Uma Máquina de Vetor de Suporte (SVM) é utilizada para classificar tais parâmetros estruturais linguísticos. Finalmente, calculam-se as distâncias entre os parâmetros classificados pela SVM e os desejados. A classificação dos sinais selecionados é realizada, assim, a partir dos valores destas distâncias. Os experimentos implementados mostram que os atributos destes parâmetros podem ser reconhecidos com sucesso em termos de características obtidas a partir das imagens RGB-D, com desempenhos individuais acima de 80% em média, o que pode ser considerado regular quando comparado a outros sistemas. Pode-se concluir que a metodologia proposta para a extração das características a partir da decomposição dos sinais em sua estrutura fonológica é um método promissor que pode auxiliar sistemas especialistas projetados para reconhecimento de sinais.

Palavras-chaves: Extração de características. Reconhecimento de Língua Brasileira de Sinais. Sensores RGB-D.

Abstract

The feature extraction in Sign Language Recognition (SLR) is a challenging problem in Computer Vision. In this work, a methodology for feature extraction in Brazilian Sign Language (BSL, or Libras in Portuguese) that addresses some of these challenges is proposed. In this methodology the phonological structure of the language, relying on RGB-D sensor for obtaining intensity, position and depth data is explored. From the RGB-D images we obtain seven vision-based features. Each feature is related to one, two or three structural elements in BSL. This relation between extracted features and structural elements based on shape, movement and position of the hands is investigated. A Support Vector Machines (SVM) is employed to classify elements based on these features. Finally, distances between classified and desired elements are calculated. From these distances, the signs classification is performed. The experiments show that the attributes of these elements can be successfully recognized in terms of the features obtained from the RGB-D images, with accuracy results individually above 80% on average. It can be concluded that the proposed feature extraction methodology and the decomposition of the signs into their phonological structure is a promising method to help expert systems designed for SLR.

Key-words: Feature extraction. Brazilian Sign Language Recognition. RGB-D sensors.

Lista de Ilustrações

Figura 1 Línguas de sinais cujas características linguísticas de partículas negativas irregulares e interrogativas são descritas quantitativamente em (DRYER; HASPELMATH, 2014). Figura originalmente disponível no sítio http://wals.info/languoid/ e editada em Julho de 2014. 23

Figura 2 Parâmetros fonológicos da Língua Brasileira de Sinais (baseado em (QUADROS; KARNOPP, 2004)). ENM: expressões não-manuais. PA: ponto de articulação. M: movimento. CM: configuração de mão. Or: orientação da palma da mão. 42

Figura 3 As quatro áreas principais de articulação dos sinais, conforme Battison (1974 apud QUADROS; KARNOPP, 2004, p. 49), no espaço de enunciação dos sinais. 43

Figura 4 Atributos e valores para o parâmetro Ponto de articulação. Cada um dos 3 atributos (cabeça, ombro e tronco) pode assumir individualmente os valores D (direita), C (centro) e E (esquerda). 44

Figura 5 As 46 configurações de mão (CMs) da Língua Brasileira de Sinais de acordo com Ferreira-Brito (1995). 44

Figura 6 Os 13 agrupamentos para as 134 configurações de mão propostas em (**??**). As configurações de mão são agrupadas por semelhança em sua forma. 45

Figura 7 Orientações da palma da mão, conforme Quadros e Karnopp (2004). 49

Figura 8 Os 34 sinais gravados na base de sinais e utilizados nesta tese. (1)“Pessoa”, (2)“Espalhar”, (3)“Copiar”, (4)“Agarrar”, (5)“Recolher”, (6)“Sumir”, (7)“Olhar”, (8)“Justo”, (9)“Verdade”, (10)“Peso”, (11)“Justiça”, (12)“Quem”, (13)“Nada”, (14)“Acreditar”, (15)“Esquecer”, (16)“Amar”, (17)“Angustiado”, (18)“Comemorar”, (19)“Rancor”, (20)“Assembleia”, (21)“Comparar”, (22)“Gritar”, (23)“Falar”, (24)“Absorver”, (25)“Engordar”, (26)“Brigar”, (27)“Esperto”, (28)“Brilhar”, (29)“Empregado”, (30)“Substituir”, (31)“Prisão”, (32)“Televisão”, (33)“Ontem”, (34)“Futuro”. As figuras de cada sinal foram editadas a partir de Capovilla, Raphael e Maurício (2012a), Capovilla, Raphael e Maurício (2012b). 53

Figura 9 Valores dos atributos dos parâmetros fonológicos resumidos para a mão direita dos 34 sinais selecionados. (1) *Ponto de articulação*: (i) 1. À direita da cabeça (D). 2. Ao centro da cabeça (C). (ii) 3. À direita do ombro (D). (iii) 4. À direita do tronco (D). 5. Ao centro do tronco (C). / (2) *Configuração da mão*: (i) 1. G1. 2. G2. 3. G4. 4. G6. 5. G7. 6. G8. 7. G10. (ii) 8. Alinhamento no eixo x (x). 9. Alinhamento no eixo y (y) . 10. Alinhamento no eixo z (z). (iii) 11. Mudança na configuração da mão ou alinhamento: sim (S). 12. Mudança na configuração da mão ou alinhamento: não (N). / (3) *Movimento*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para a direita (D). 4. Para a esquerda (E). 5. Para dentro (I). 6. Para fora (F). (ii) 7. Frequência: simples (S). 8. Frequência: repetida (R). / (4) *Orientação da palma da mão*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para dentro (D). 4. Para fora (F). 5. Para o lado (L). (iii) 6. Variação: sim (S). 7. Variação: não (N). 54

Figura 10 Valores dos atributos dos parâmetros fonológicos resumidos para a mão esquerda dos 34 sinais selecionados. (1) *Ponto de articulação*: (i) 1. À esquerda da cabeça (E). (ii) 2. À esquerda do ombro (E). (iii) 3. À esquerda do tronco (E). 4. Ao centro do tronco (C). / (2) *Configuração da mão*: (i) 1. G1. 2. G4. 3. G6. 4. G8. 5. G10. (ii) 6. Alinhamento no eixo x (x). 7. Alinhamento no eixo y (y). 8. Alinhamento no eixo z (z). (iii) 9. Mudança na configuração da mão ou alinhamento: sim (S). 10. Mudança na configuração da mão ou alinhamento: não (N). / (3) *Movimento*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para a esquerda (E). 4. Para a direita (D). 5. Para dentro (I). 6. Para fora (F). (ii) 7. Frequência: simples (S). 8. Frequência: repetida (R). / (4) *Orientação da palma da mão*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para dentro (D). 4. Para fora (F). 5. Para o lado (L). (ii) 6. Variação: sim (S). 7. Variação: não (N). 55

Figura 11 Pontos do corpo humano acessados pelos *softwares* que operam o Kinect. 59

Figura 12 Formatos dos vídeos gravados para o sinal “Comemorar” utilizando o *software nuiCaptureAnalyze* operando o sensor Kinect: (a) Intensidade, (b) Profundidade e (c) Posição. 65

Figura 13 Primeira parte de duas do vídeo para o sinal “Justiça”: quadros 1 a 60 para a sua primeira gravação. Este sinal apresenta a maior quantidade de quadros na média entre as cinco gravações realizadas e, consequentemente, o maior tempo de execução. Os $m = 5$ quadros selecionados por meio da sumarização estão em destaque com moldura vermelha. 73
Figura 14 Segunda parte de duas do vídeo para o sinal “Justiça”: quadros 61 a 94 para a sua primeira gravação. Este sinal apresenta a maior quantidade de quadros na média entre as cinco gravações realizadas e, consequentemente, o maior tempo de execução. Os $m = 5$ quadros selecionados por meio da sumarização estão em destaque com moldura vermelha. 74
Figura 15 Quadros do sinal “Esquecer”, para a sua primeira gravação. Este sinal apresenta a menor quantidade de quadros na média entre as cinco gravações realizadas e, consequentemente, o menor tempo de execução. Os $m = 5$ quadros selecionados por meio da sumarização estão em destaque com moldura vermelha. 75
Figura 16 Detecção da região de interesse - mãos direita e esquerda para o sinal “Comemorar”. Vídeos gravados utilizando o *software nuiCaptureAnalyze* operando o sensor Kinect. Em (a), vê-se um quadro colorido no formato RGB do vídeo gravado; em (b) um quadro de profundidade do vídeo de profundidade; em (c) um quadro do esqueleto do vídeo do esqueleto. Em (d) estão os quadros selecionados pela sumarização de vídeos por meio da solução do Problema da Diversidade Máxima. Já em (e) e (f) tem-se as regiões de interesse em RGB, detectada para as mãos direita e esquerda, respectivamente. Em (g) e (h) estas regiões de interesse estão detectadas em Preto & Branco por meio do algoritmo de detecção de cor da pele. 78
Figura 17 Sinais utilizados como exemplo para a extração das características. São exibidos os cinco quadros selecionados sequencialmente e todos para a mão direita. Os cálculos das características também são realizados para a mão esquerda, embora não apareçam nos exemplos. 82
Figura 18 (a) Exemplo das posições (x, y) extraídas da imagem do esqueleto. A região da mão direita está gravada na cor azul; a região da mão esquerda está gravada na cor alaranjada. O ombro está gravado na cor verde. 83
Figura 19 Estrutura computacional do arquivo em formato Matlab (MathWorks (2012) gravado pelo *NuiCaptureAnalyze*. 84

Figura 20 Sistema de coordenadas de referência para a gravação da distância tridimensional (x, y, z) utilizada como característica. 85
Figura 21 Vetores de movimento para cada uma das cinco imagens da mão direita dos três sinais exemplo: “Espalhar”, “Agarrar” e “Futuro”. Os vetores estão destacados em vermelho e posicionados onde há variação de brilho entre os quadros. 86
Figura 22 Vetores de velocidade calculados pelo fluxo óptico para cada um dos cinco quadros dos três sinais exemplo: “Espalhar”, “Agarrar” e “Futuro”. 87
Figura 23 Áreas das mãos definidas pelo retângulo em cor verde em cada um dos cinco quadros para a mão direita dos sinais exemplo “Espalhar”, “Agarrar” e “Futuro”. 88
Figura 24 Valores das áreas das mãos definida pelo retângulo em *pixels*^{2} em cada um dos cinco quadros para a mão direita para os sinais “Espalhar”, “Agarrar” e “Futuro”. 89
Figura 25 Característica extraída “Média das posições das quinas” entre os sinais “Espalhar”, “Agarrar” e “Futuro”. 90
Figura 26 Valores da característica “Média das posições das quinas” entre os sinais “Espalhar”, “Agarrar” e “Futuro”. 91
Figura 27 Definição das variáveis ρ e θ na detecção de linhas por meio da transformada de *Hough*. 91
Figura 28 Característica “Linhas detectadas” para cada um dos quadros dos sinais exemplos “Espalhar”, “Agarrar” e “Futuro”. As cinco maiores linhas estão descritas em verde em cada quadro, com o ponto inicial em vermelho e o final em amarelo. A linha detectada é aquela de maior tamanho e seus valores de ρ e θ são as características extraídas. Elas estão destacadas em vermelho. 92
Figura 29 Valores dos pares ordenados (ρ, θ) das cinco maiores linhas detectadas em cada um dos cinco quadros dos sinais “Espalhar”, “Agarrar” e “Futuro.” 93
Figura 30 Diferenças entre os quadros 1-2, 2-3, 3-4 e 4-5 em sequência para os sinais exemplo “Espalhar”, “Agarrar” e “Futuro”. 95
Figura 31 Quantidades de pontos em comum para cada uma das combinações 1-2, 2-3, 3-4 e 4-5 para os sinais “Espalhar”, “Agarrar” e “Futuro”. . 96

Figura 32 Esquemático do sistema implementado. (a) Características: sete características extraídas. (b) Vetor: veja Tabela 19 para a estrutura, $n = 5$ quadros selecionados e M = número de características para um dado elemento em Libras, veja a Tabela 17 para referência. (c) Classificação: k_{pa} saídas para o Ponto de articulação ($k_{pa} = 5$ para a mão direita e $k_{pa} = 4$ para a mão esquerda); k_{cm} saídas para a Configuração de mão ($k_{cm} = 12$ para a mão direita e $k_{cm} = 10$ para a mão esquerda); k_{mov} saídas para o Movimento ($k_{mov} = 8$ para as mãos direita e esquerda); k_{or} saídas para a orientação da palma da mão ($k_{or} = 7$ para ambas as mãos, direita e esquerda). 101
Figura 33 Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 19. 104
Figura 34 Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 19. 105
Figura 35 Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 20. 105
Figura 36 Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 20. 106
Figura 37 Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21. 107

Figura 38 Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21. 107
Figura 39 Comparação entre os vetores de características na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov), (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21. 108
Figura 40 Comparação entre os vetores de características na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov), (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21. 109

Lista de Tabelas

Tabela 1 Países que utilizam Língua de Sinais própria, de acordo com Lewis et al. (2009). 22

Tabela 2 População residente, com Tipo de deficiência auditiva por grupo de idade - Brasil 2010, IBGE. 34

Tabela 3 População residente, com Tipo de deficiência auditiva por alfabetização - Brasil 2010, IBGE. 35

Tabela 4 Locações definidas por Ferreira-Brito (1995 apud QUADROS; KARNOPP, 2004). 43

Tabela 5 Categorias do parâmetro Movimento na Língua Brasileira de Sinais segundo Ferreira-Brito (1990). 47

Tabela 6 Atributos e valores para o parâmetro Movimento. (1) Para cima: sinal “Comemorar”. (2) Para baixo: sinal “Esperto”. (3) Para a direita: sinal “Pessoa”. (4) Para a esquerda: sinal “Engordar”. (5) Para dentro: sinal “Ontem”. (6) Para fora: sinal ”Futuro“. 48

Tabela 7 Valores possíveis para o parâmetro Orientação da palma da mão. . 49

Tabela 8 Atributos e valores para o parâmetro Orientação da palma da mão. (1) Para cima: sinal “Verdade”. (2) Para baixo: sinal “Espalhar”. (3) Para dentro: sinal “Sumir”. (4) Para fora: sinal “Olhar”. (5) Para o lado: sinal “Amar”. 50

Tabela 9 Expressões não-manuais da Língua Brasileira de Sinais, definidas por Ferreira-Brito (1995). 51

Tabela 10 Parâmetros da estrutura fonológica da Libras originalmente proposta por Quadros e Karnopp (2004) e resumida nesta tese. Cada parâmetro consiste de um conjunto de atributos com os seus valores possíveis discriminados na coluna mais à direita. 51

Tabela 11 (a) Sensor Kinect produzido entre os anos de 2010 e 2014 e comercializado como acessório de movimento do videogame XBOX, da Microsoft. (b) Sensor Kinect produzido a partir de 2013 e atualmente comercializado como acessório de movimento do videogame XBOX, da Microsoft. (c) Sensor PrimeSense Carmine desenvolvido pela empresa de mesmo nome e vendido para a Apple Incorporation em novembro de 2013. (d) Sensor ASUS Xtion produzido pela empresa ASUS. 58

(a) Kinect lançado em 2010. 58
(b) Kinect lançado em 2013. 58
(c) PrimeSense Carmine 58
(d) ASUS Xtion 58
Tabela 12 Comparação entre o sensor Kinect, da Microsoft e os sensores PrimeSense Carmine / ASUS Xtion. 60
Tabela 13 Resumo das soluções apresentadas 63
Tabela 14 Parâmetros fonológicos da mão direita dos sinais com maior e menor tempo de duração na média, respectivamente, sinais “Justiça” e “Esquecer”. 71
(a) G4 - Forma 2 71
(b) G1 - Forma 3 71
Tabela 15 Parâmetros fonológicos da mão esquerda dos sinais com maior e menor tempo de duração na média, respectivamente, sinais “Justiça” e “Esquecer”. 71
(a) G4 - Forma 2 71
Tabela 16 Quantidade média de quadros e tempo médio de execução em segundos para cada um dos 34 sinais da base de sinais utilizada. Cada sinal foi gravado cinco vezes a uma taxa de 30 quadros por segundo e por um único sinalizador. 72
Tabela 17 As características extraídas da região de interesse são exibidas na coluna à esquerda. Os parâmetros fonológicos da Libras associados a estas características são exibidos na coluna da direita. 80
Tabela 18 Parâmetros fonológicos dos sinais utilizados como exemplo. Todos os parâmetros são para a mão direita. 81
(a) G1 - Forma 1 81
(b) G1 - Forma 4 81
(c) G4 - Forma 8 81
Tabela 19 Estrutura do vetor de características para *M* características e *n* quadros. Neste estudo são analisadas $M = 1, 2, 3$ características, dependendo do parâmetro e $n = 5$ quadros. 96
Tabela 20 Estrutura do vetor de características para *M* características e *n* quadros com a variável tempo multiplicada a cada característica. 97
Tabela 21 Estrutura do vetor de características para *M* características e *n* quadros com a variável tempo adicionada ao final. 97
Tabela 22 Saídas das Máquinas de Vetor de Suporte para cada um dos quatro parâmetros fonológicos da Língua Brasileira de Sinais para a mão direta. 101

Tabela 23 Saídas das Máquinas de Vetor de Suporte para cada um dos quatro parâmetros fonológicos da Língua Brasileira de Sinais para a mão esquerda. 102
Tabela 24 Características relacionadas a cada um dos quatro parâmetros fonológicos da Língua Brasileira de Sinais que compõem o vetor de entrada de cada bloco de máquinas de vetor de suporte. 102
Tabela 25 Eficácia média para o reconhecimento de cada um dos 34 sinais na base de sinais. Cada vetor de saída da Máquina de Vetor de Suporte é comparado às saídas desejadas apresentadas nas Figuras 9 e 10. A distância de *Hamming* é calculada entre estes vetores e os resultados são apresentados para os casos em que o acerto ocorre entre os **k = 1, 3, 5** vizinhos mais próximos. 111

Lista de algoritmos

1 Cálculo das posições (x_d, y_d) e (x_e, y_e) 77

Sumário

1 Introdução **21**

1.1 Problema de pesquisa 21
1.2 Revisão bibliográfica 25
1.2.1 As línguas de sinais 26
1.2.2 Reconhecimento automático de movimentos do corpo 28
1.2.2.1 Sistemas baseados em luvas 29
1.2.2.2 Sistemas baseados em vídeo 30
1.3 Objetivos 32
1.3.1 Objetivos Específicos 32
1.4 Motivação 33
1.5 Contribuições realizadas 35
1.6 Organização do trabalho 36

2 Estrutura da Língua Brasileira de Sinais **37**

2.1 Introdução 37
2.2 Estrutura linguística da Língua Brasileira de Sinais 39
2.3 Estrutura fonológica da Língua Brasileira de Sinais e sua adaptação para o problema de Visão Computacional 40
2.3.1 A fonologia em Língua Brasileira de Sinais 40
2.3.2 Ponto de articulação (PA) 42
2.3.3 Configuração de mão (CM) 43
2.3.4 Movimento (M) 46
2.3.5 Orientação da palma da mão (Or) 47
2.3.6 Expressões não-manuais (ENM) 50
2.3.7 Resumo dos atributos e valores para os quatro parâmetros adaptados 50
2.4 Sinais selecionados 51
2.5 Considerações finais 52

3 Criação da base de sinais **56**

3.1 Introdução 56
3.2 Ferramentas computacionais 57
3.2.1 Sensores RGB-D 57
3.2.1.1 O sensor Kinect 58

3.2.1.2 O sensor PrimeSense Carmine 59
3.2.1.3 Sensor Kinect *versus* Xtion/PrimeSense Carmine . . . 60
3.2.2 *Frameworks* 60
3.2.2.1 OpenNI 60
3.2.2.2 Microsoft Kinect for Windows SDK 61
3.2.3 Bibliotecas de comandos em Visão Computacional 61
3.2.3.1 OpenCV 61
3.2.3.2 Matlab 62
3.2.3.3 OpenKinect 62
3.2.4 Soluções utilizadas 62
3.2.5 Resumo dos recursos 63
3.3 A base dos sinais 63
3.4 Considerações finais 65

4 Sumarização de vídeos **67**
4.1 Introdução 67
4.2 Sumarização de vídeo 68
4.3 Sumarização dos sinais 70
4.4 Detecção da região de interesse 76
4.5 Considerações finais 78

5 Extração de características **80**
5.1 Introdução 80
5.2 Distância bidimensional 82
5.3 Distância tridimensional 84
5.4 Velocidade 85
5.5 Área das mãos 87
5.6 Média das posições das quinas das mãos 89
5.7 Linhas detectadas 90
5.8 Quantidade de pontos comuns entre quadros 93
5.9 Construção do vetor de características 94
5.10 Considerações finais 97

6 Experimentos e resultados **99**
6.1 Introdução 99
6.2 Experimentos 100
6.3 Classificação dos parâmetros fonológicos dos sinais 103
6.3.1 Comparação entre os *kernels* Linear e RBF 103
6.3.2 Comparação entre os três vetores de características propostos 108
6.4 Reconhecimento dos sinais 109

7 Conclusões **112**
7.1 Introdução 112
7.2 Conclusões 112
7.3 Trabalhos futuros 114

Referências Bibliográficas **116**

ANEXO A Anexo 1 **124**
A.1 Periódico - *Experts Systems With Applications* - 2014 124
A.2 Proceedings - Workshop de Visão Computacional - 2013 138

1 Introdução

Uma das marcas que diferencia os seres humanos de outros animais é a utilização natural da língua como meio de comunicação e expressão cultural, de acordo com Sacks (1990). Comunidades, sejam elas grandes ou pequenas, criam os mais diversos códigos para permitir que seus integrantes comuniquem-se entre si. Definições de estruturas linguísticas sintáticas, morfológicas, gramaticais e fonológicas, resultados do amadurecimento destes códigos, acabam por estabelecer uma estrutura completa e complexa para as línguas. É a utilização dos sentidos inerentes aos seres humanos que leva à criação de línguas de diferentes estruturas e modalidades. As línguas orais-escritas, tal como a Língua Portuguesa, beneficiam-se do nosso sistema visual e auditivo. Para aqueles que são privados total ou parcialmente da visão, a utilização do Braile [1] possibilita-lhes o estabelecimento da comunicação. Já a língua de sinais [2], devido à sua característica visual e espacial, é a forma mais natural de comunicação para as pessoas privadas de forma parcial ou total da audição. É esta característica espacial das línguas de sinais que chama a atenção quando projetamos sistemas computacionais baseados em visão e é objeto de pesquisa neste trabalho, com vistas à criação de um sistema computacional capaz de reconhecer seus padrões.

1.1 Problema de pesquisa

Ao redor do mundo há uma grande variedade de línguas de sinais, assim como ocorre com as línguas orais-escritas. Em (LEWIS et al., 2009), os autores realizam uma classificação genética das línguas em 136 diferentes famílias, que chama de grupos genéticos de nível superior. Seis destas famílias destacam-se como as maiores famílias de línguas do mundo. Juntas, elas respondem por cerca de 2/3 de todas as línguas e atingem 5/6 da população mundial.

As outras 130 famílias incluem categorias especiais de línguas definidas como línguas construídas, línguas de sinais, línguas isoladas, línguas misturadas, línguas de contato e línguas ainda não classificadas. Em todos estes casos as línguas são utilizadas atualmente e, portanto, denominadas “vivas”. Na Tabela 1 são exibidos os 105 países citados em (LEWIS et al., 2009) nos quais as línguas de sinais são estabelecidas como línguas oficiais. Nestes países as línguas de sinais são faladas, ao todo,

[1] O Braile não é uma língua em si. Ele é um sistema de leitura baseado em código no qual línguas orais são escritas e lidas.

[2] Um sinal em Língua de sinais é o equivalente a uma palavra em línguas orais-escritas.

Países que utilizam Língua de Sinais		
1. Afeganistão	36. França	71. Nicarágua
2. África do Sul	37. Gana	72. Nigéria
3. Albânia	38. Grécia	73. Noruega
4. Alemanha	39. Guatemala	74. Nova Zelândia
5. Algéria	40. Guiné	75. Panamá
6. Arábia Saudita	41. Holanda	76. Paquistão
7. Argentina	42. Honduras	77. Paraguai
8. Armênia	43. Hong Kong	78. Peru
9. Austrália	44. Hungria	79. Polônia
10. Áustria	45. Ilhas Maurício	80. Porto Rico
11. Bélgica	46. Índia	81. Portugal
12. Bolívia	47. Indonésia	82. Quênia
13. Brasil	48. Irã	83. Reino Unido
14. Bulgária	49. Irlanda	84. República Dominicana
15. Canadá	50. Islândia	85. República Tcheca
16. Chade	51. Israel	86. Romênia
17. Chile	52. Itália	87. Rússia
18. China	53. Jamaica	88. Serra Leoa
19. Cingapura	54. Japão	89. Sérvia
20. Colômbia	55. Jordânia	90. Sri Lanka
21. Coréia do Sul	56. Laos	91. Suécia
22. Costa Rica	57. Letônia	92. Suíca
23. Croácia	58. Líbia	93. Tailândia
24. Cuba	59. Lituânia	94. Taiwan
25. Dinamarca	60. Madagascar	95. Tanzânia
26. Egito	61. Malásia	96. Trinidad e Tobago
27. El Salvador	62. Mali	97. Tunísia
28. Equador	63. Malta	98. Turquia
29. Eslováquia	64. Marrocos	99. Ucrânia
30. Espanha	65. México	100. Uganda
31.Estados Unidos	66. Moçambique	101. Uruguai
32. Estônia	67. Moldávia	102. Venezuela
33. Etiópia	68. Mongólia	103. Vietnã
34. Filipinas	69. Namíbia	104. Zâmbia
35. Finlândia	70. Nepal	105. Zimbabue

Tabela 1: Países que utilizam Língua de Sinais própria, de acordo com Lewis et al. (2009).

por cerca de mais de 6.700.00 pessoas, o que representa aproximadamente 0,11% da população mundial.

Outra abordagem sobre as línguas de sinais existentes pode ser encontrada de forma *online* no sítio *Atlas Mundial Online de Estruturas Linguísticas* [3]. Nele, a distribui-

[3] Atlas Mundial Online de Estruturas Linguísticas - http://wals.info/languoid/genus/

ção das línguas de sinais ocorre geograficamente de acordo com duas características linguísticas pré-estabelecidas: a quantidade de partículas negativas irregulares e de partículas interrogativas nas línguas.

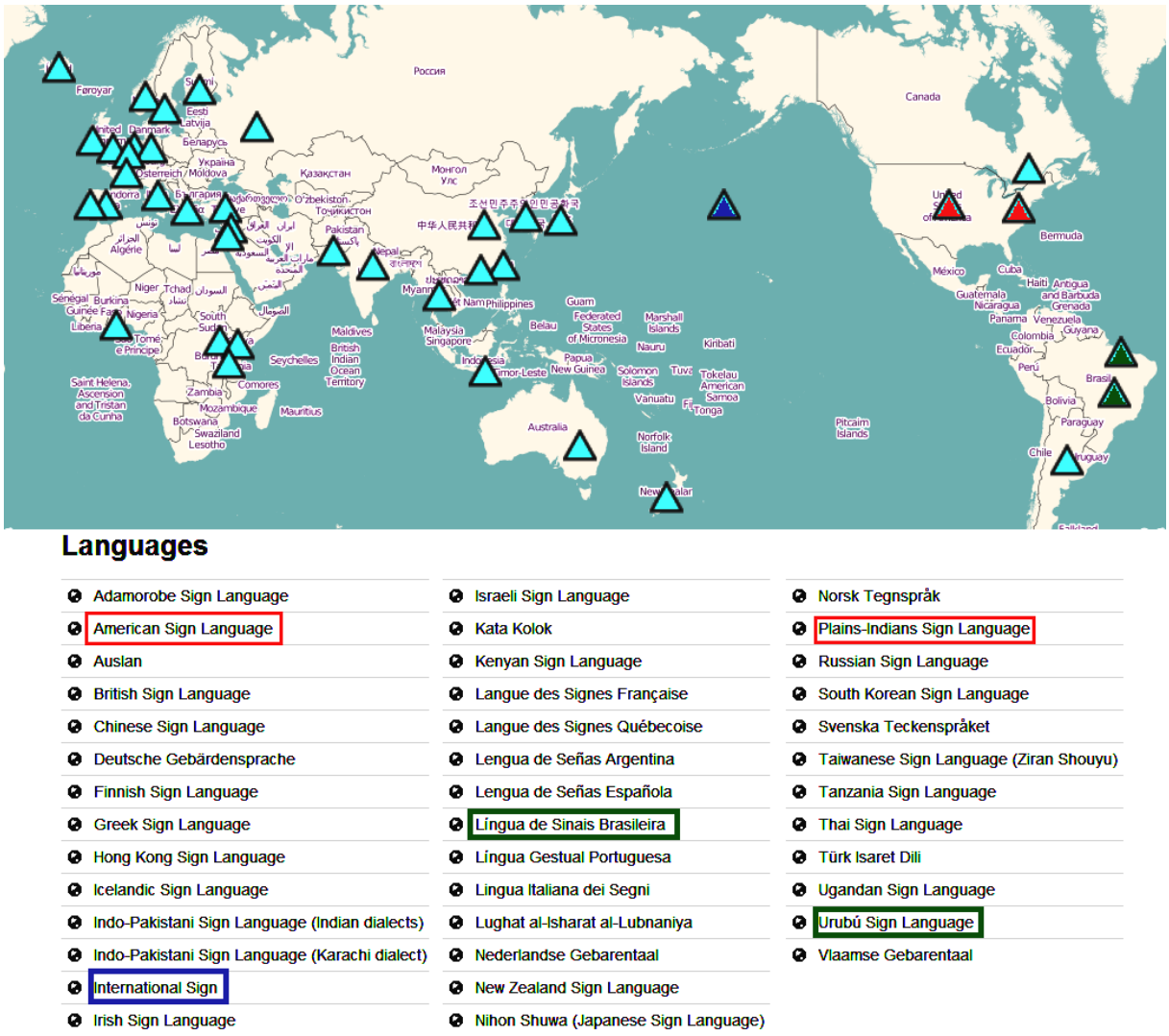


Figura 1: Línguas de sinais cujas características linguísticas de partículas negativas irregulares e interrogativas são descritas quantitativamente em (DRYER; HASPELMATH, 2014). Figura originalmente disponível no sítio http://wals.info/languoid/ e editada em Julho de 2014.

A referência para pesquisas destas características linguísticas está disponível para algumas das línguas. Nota-se nesta abordagem que há países com mais de uma língua de sinais, ainda que oficialmente haja somente uma. É o caso dos Estados Unidos, onde há duas línguas de sinais, a *American Sign Language (ASL)*, língua de sinais oficial do país e a *Plains-Indians Sign Language*, língua de sinais utilizada por nativos norte-americanos das grandes planícies dos Estados Unidos e Canadá. Ambas estão destacadas em triângulos de cor vermelha no mapa da Figura 1. O caso do Brasil está em destaque em triângulos de cor verde na mesma figura. A *Língua Brasileira de Sinais* é a língua de sinais oficial no Brasil, mas há também a *Urubú Sign*

Language (conhecida por *Kaapor Sign Language*) existente nos estados do Maranhão e do Pará. É possível ainda verificar na Figura 1 uma língua de sinais internacional, a *International Sign*, destacada em um triângulo na cor azul. Ela é uma tentativa de se criar uma língua única para os surdos. Este é um dos muitos mitos que envolvem as línguas de sinais. A ideia geral que se tem é que é possível criar uma língua única, com sinais comuns ao mundo todo para todas as pessoas surdas comunicarem-se. Mas as línguas de sinais sofrem forte influência regional como qualquer outra língua e, embora sejam similares em algumas questões estruturais, são bastante distintas nos sinais que representam as suas palavras.

As línguas de sinais são compostas por um conjunto limitado de fonemas, tais quais as línguas orais-escritas. E é a combinação destes fonemas que estabelece as palavras de cada língua. Os fonemas são bastante similares mesmo para línguas de sinais de países distintos. Os sinais, entretanto, diferem bastante de um país para outro e mesmo entre regiões do mesmo país.

A Língua Brasileira de Sinais, também chamada Libras, foi reconhecida como meio legal de comunicação e expressão da comunidade surda por meio da Lei de n^o 10.436, de 24 de abril de 2002. Sua existência, no entanto, data do final do século XVIII, com a criação do Imperial Instituto dos Surdos-Mudos no Brasil pelo professor francês Ernest Huet, o qual tornou-se surdo ao longo da vida. Devido ao fato de Ernest Huet ser francês, a Libras acabou por sofrer grande influência da Língua Francesa de Sinais. Assim, muitos de seus sinais derivam da Língua Francesa de Sinais ou são comuns a ambas.

A estrutura fonológica de uma língua de sinais, de forma geral, possui cinco parâmetros. A Língua Brasileira de Sinais não é exceção e, portanto, também é constituída fonologicamente por estes parâmetros, que são: (i) o ponto de articulação, (ii) a configuração da mão, (iii) o movimento da mão, (iv) a orientação da palma da mão e (v) as expressões faciais e corporais.

As características espaciais e temporais destes parâmetros da estrutura fonológica em língua de sinais podem ser exploradas no desenvolvimento de sistemas automáticos especialistas na área de Reconhecimento de Línguas de Sinais (SLR). Há grandes desafios para implementar um sistema computacional que seja capaz de compreender a complexidade e as sutilezas das línguas de sinais. Estes desafios estão presentes em todas as etapas necessárias à sua criação: na etapa de aquisição da informação de entrada; no processamento desta informação; na extração e posterior seleção de características; nas bases de treinamento necessárias para validação do reconhecimento de sinais; na classificação dos mesmos. Para lidar com tais desafios tanto técnicas de processamento de imagens quanto de reconhecimento de padrões e de visão computacional são combinadas em uma busca por um sistema capaz de

reconhecer sinais com um desempenho confiável.

Sobre a implementação destes sistemas/interfaces computacionais, cujo objetivo é a interação entre homem e máquina em reconhecimento de sinais, há sistemas que são baseados em luvas instrumentalizadas ou luvas total ou parcialmente coloridas. Mais recentemente, os sistemas baseados na captura de imagens por meio de sensores que gravam intensidade e profundidade, os chamados sensores RGB-D (Red, Green, Blue and Depth), têm sido bastante pesquisados. Nestes últimos não há a necessidade do uso de quaisquer equipamentos pela pessoa que realiza o sinal, chamado nesta tese de sinalizador.

No presente trabalho uma nova abordagem especificamente do problema de extração de características no reconhecimento de parâmetros fonológicos de sinais dinâmicos em Libras é apresentada. Tal abordagem relaciona as características extraídas dos sinais aos elementos fonológicos da Libras. Os sinais selecionados para validação desta abordagem baseiam-se nos trabalhos de Capovilla, Raphael e Maurício (2012a), Capovilla, Raphael e Maurício (2012b), que agrupam os mais de 10 mil sinais documentados na língua em 34 morfemasmorfemas [4]. Assim, um sinal de cada agrupamento de morfemas foi escolhido e gravado por um sensor RGB-D. As informações obtidas nesta gravação auxiliam as técnicas de processamento das imagens, das quais são extraídas as características utilizadas na implementação do reconhecimento de parâmetros fonológicos da Língua Brasileira de Sinais.

1.2 Revisão bibliográfica

As línguas de sinais são línguas completas, possuindo regras gramaticais, ortográficas, morfológicas e fonológicas complexas tal como acontece com as línguas de modalidade oral-auditiva. Devido à sua característica espacial, elas despertam grande interesse das mais diversas áreas de pesquisa, que abordam desde a estrutura e construção da língua até o desenvolvimento de sistemas computacionais capazes de reconhecê-las de forma mais automática e precisa, passando pelos estudos cognitivos da aquisição da língua de sinais e o desenvolvimento humano que se faz a partir do uso desta.

A abordagem realizada nesta tese trata da extração de características visuais partindo da estrutura fonológica da Língua Brasileira de Sinais, com vistas ao desenvolvimento futuro de um sistema que seja capaz de reconhecer sinais automaticamente.

Duas linhas de pesquisas tem suas abordagens citadas a seguir, diante da tamanha diversidade de pesquisas na área de Reconhecimento de Línguas de Sinais.

[4] Um morfema é a menor unidade sublexical que codifica significado.

Primeiramente são citadas abordagens acerca de pesquisas sobre a documentação das línguas de sinais. Logo em seguida são apresentados sistemas computacionais que reconhecem gestos e movimentos em diversas aplicações. Aprofunda-se aqui, então, a aplicação específica de reconhecimento dos sinais estáticos ou dinâmicos em várias línguas de sinais.

1.2.1 As línguas de sinais

O estudo da formação, padronização e documentação de regras que constituem a língua de sinais é de interesse especial da área de linguística. A criação e/ou construção de dicionários em língua de sinais é um tema bastante amplo e que envolve uma série de questões de ordem não só linguística, mas também social. No caso das línguas orais-escritas, por exemplo, a forma para os seus dicionários é conhecida e os vocábulos são organizados em ordem alfabética, baseando-se em um alfabeto que possui códigos bem definidos. Isto é possível porque a escrita existe na documentação das línguas orais-escritas e é por meio dela que se descrevem os conceitos de cada língua desta modalidade. E para o caso das línguas de sinais, cuja modalidade é visual-espacial? Como organizar os vocábulos (sim, ela tem vocábulos, que são os sinais) desta língua em um dicionário? Se a escrita é utilizada, então supõe-se que os usuários da respectiva língua de sinais daquele país conheçam alguma língua oral-escrita, sejam surdos ou ouvintes. Entretanto, a documentação das línguas de sinais não se faz por meio da escrita que relaciona palavra-objeto, mas sim por meio de uma descrição da forma visual de cada sinal. Esta descrição precisa, necessariamente, ser o mais visual possível.

No dicionário trilíngue composto por 2 volumes e apresentado em (CAPOVILLA; RAPHAEL; MAURíCIO, 2012a) e em (CAPOVILLA; RAPHAEL; MAURíCIO, 2012b) mais de 10.000 verbetes são catalogados em ordem alfabética e ilustrações são utilizadas para que cada verbete possa ser reproduzido em Língua Brasileira de Sinais, Língua Portuguesa e Língua Inglesa. Ao longo da documentação dos sinais deste dicionário, outros trabalhos foram realizados. Um dos resultados discutidos nas pesquisas em (CAPOVILLA et al., 2003) é a criação do *software* denominado *BuscaSigno* [5], onde o usuário pode procurar pelo sinal que deseja por meio da seleção de parâmetros que o constitui. O interessante é que o usuário pode selecionar o sinal apenas sabendo como ele é realizado no espaço. Este *software*, entretanto, não encontra-se disponível ainda para utilização geral, distintamente do dicionário trilíngue citado anteriormente. Um outro resultado apresentado em (CAPOVILLA et al., 2003) é um sistema criado para que surdos com quadriplegia [6] possam interagir com outras

[5] Ver Duduchi e Capovilla (2006).

[6] Uma pessoa quadriplégica é “uma pessoa com paralisia de ambos os braços e pernas causada por alguma lesão ou doença na coluna”, conforme descrito em (MERRIAM-WEBSTER, 2014).

pessoas por meio da categorização de sinais previamente selecionados que permitam uma comunicação mais direta destes com ouvintes e/ou outros surdos utilizando frases comuns e pré-estabelecidas.

Em (SCHEMER, 2003), o autor descreve o processo de padronização da gramática e do léxico da Língua Holandesa de Sinais. Para tal, dois grandes projetos são organizados e questões como a regionalização dos sinais são discutidas. Anteriormente ao reconhecimento recente das línguas de sinais como línguas completas, o ensino destas ocorria somente em escolas especiais para surdos. Assim, cada escola acabava criando seus próprios sinais, o que dificultou e ainda dificulta bastante a compilação de um dicionário que contenha todos os sinais e cubra as regionalidades existentes. Assim como em outras línguas, a regionalização dos verbetes também está presente em línguas de sinais e é um dos componentes que torna as línguas dinâmicas.

Uma abordagem distinta para a compilação de um dicionário e padronização dos sinais na língua é apresentada em (JOHNSTON, 2003). Nela, os sinais são agrupados pela forma da mão. Estas formas consideradas como referência são os números de 0 a 9 conjuntamente com mais 3 possíveis formatos. O autor chama a atenção para o fato de que, na realidade, o que é feito é uma documentação dos sinais e não padronização, uma vez que as diferenças não só regionais mas também entre grupos distintos classificados por idade, região, religião e sexo é bastante evidente. E nenhum grupo aceita com facilidade tentativas de padronização, pois para cada grupo aceitar um padrão é como abrir mão da própria identidade em prol de outra desconhecida.

Dois trabalhos sobre a compilação de dicionários são apresentados em (CLEVE, 2003) e em (WILCOX, 2003), ambos tendo como assunto a Língua Americana de Sinais (ASL - American Sign Language). No primeiro caso, o dicionário apresentado é criado na Universidade Gallaudet [7], que é uma universidade privada dedicada à educação de pessoas total ou parcialmente surdas, localizada em Washington DC, EUA. No segundo caso, o dicionário foi desenvolvido por meio de um projeto patrocinado pelo National Institutes of Health e executado em um Centro de Inovação de Pequenos Negócios (SBIR - Small Business Innovative Research). Uma versão proprietária deste dicionário foi desenvolvida para a plataforma Macintosh em linguagem C++ e distribuída em CD-ROM.

Sobre a Língua Brasileira de Sinais, Quadros (2012) apresenta o planejamento e as políticas públicas relacionadas ao seu reconhecimento e ensino no Brasil. Já Felipe (2006) realiza um estudo da formação da palavra, que é o sinal visual realizado pelos usuários da língua. Diversas questões específicas da linguística são analisadas. Chaveiro et al. (2013) e Capovilla (2008) trazem avaliações sobre a qualidade de vida

[7] http://www.gallaudet.edu/. Acessado em 06 de junho de 2014.

da população surda e do desenvolvimento de competências linguísticas em surdos, respectivamente. Por fim, Verdu et al. (2012) verificam, em seu trabalho, a capacidade de discriminar partes dos sinais em adolescentes com surdez bilateral.

Um texto bastante completo sobre as questões da linguagem e da cultura surda pode ser encontrado em Sacks (1990), em seu livro *Vendo Vozes: uma viagem ao mundo dos surdos*, tradução de Laura Teixeira Motta. Nele, o autor aborda questões diversas que nos provocam questionamentos sobre a natureza das línguas, tais como: "O que é necessário [...] para nos tornarmos seres humanos completos? ", "O que denominamos nossa humanidade dependerá parcialmente da linguagem?", "A linguagem desenvolve-se de um modo espontâneo e natural ou requer contato contato com outros seres humanos?" Ao descrever seus próprios pensamentos e de outros autores a respeito da linguagem, o autor faz uma viagem ao mundo dos surdos por meio de histórias e questionamentos acerca da aprendizagem da língua natural destes, que é a língua de sinais e da observação da aquisição de todas as habilidades cognitivas pelos surdos a partir da sua aprendizagem.

A documentação tanto de línguas orais quanto de línguas visuais apresentam dificuldades em comum. O importante é ter clareza que línguas, independente da modalidade a que pertençam, são objetos vivos, dinâmicos e, portanto, mutáveis. Afinal, elas retratam a cultura e expressão de pensamento da comunidade a qual pertencem.

1.2.2 Reconhecimento automático de movimentos do corpo

A variedade de padrões que compõem os sinais linguísticos e as sutilezas entre estes trazem à tona o desafio de se fazer com que os sistemas computacionais sejam capazes de compreender estes padrões e estas sutilezas, tão facilmente perceptíveis aos seres humanos.

Sistemas que reconhecem movimentos e gestos corporais são úteis em várias áreas de conhecimento. Quatro trabalhos citados a seguir apresentam análises sobre aplicações e técnicas gerais utilizadas em reconhecimento automático de movimentos e gestos. Em (WATSON, 1993) a comparação é realizada entre as técnicas de reconhecimento de gestos aplicadas à manipulação de sistemas em três dimensões e reconhecimento de sinais linguísticos. Loeding et al. (2004) realizam um extenso levantamento no qual analisam trabalhos de reconhecimento de gestos em etapas que compõem a grande maioria dos sistemas computacionais. Tais etapas são constituídas por: (i) análise dos dados de entrada; (ii) o modelo de reconhecimento de gestos proposto; (iii) as bases de dados utilizadas para validação da aproximação proposta; (iv) a performance da aproximação proposta e, (v) atributos não manuais dos sinais. Ao todo, Loeding et al. (2004) analisam 23 trabalhos tendo como referência estes 5 itens. Já em (ONG; RANGANATH, 2005), a análise de sistemas automáticos atual e

futura é realizada, com foco não apenas no vocabulário que compõe as línguas de sinais, mas em todos os aspectos que também fazem parte da língua, tais como os parâmetros não manuais dos sinais e processos gramaticais de construção de frases. O quarto artigo, de autoria de Parton (2006), analisa a multidisciplinaridade necessária para a criação de sistemas de reconhecimento de sinais na área de Inteligência Artificial. Estes quatro trabalhos resumem aspectos gerais de reconhecimento de movimentos, sem necessariamente especificar uma única aplicação.

Em se tratando de reconhecimento específico de sinais linguísticos, o que é o caso para a aplicação de línguas de sinais, duas possíveis linhas de pesquisa coexistem em visão computacional. Quando estes sinais são capturados por meio de sistemas intrusivos, tais como luvas ou roupas especiais com sensores eletrônicos acoplados nestas luvas ou roupas, os sistemas são ditos sistemas baseados em luvas. Já quando a leitura do sinal é realizada por meio da captura de imagens por câmeras 2D ou 3D, estamos lidando com os sistemas baseados em vídeos.

Obviamente, os desafios para o reconhecimento por meio de um ou outro sistema são bastantes distintos. Em sistemas baseados em luvas, as informações de entrada são capturadas por sensores eletrônicos, que fornecem dados espaciais e temporais a partir de acelerômetros instalados nos equipamentos e que o usuário deve, necessariamente, vestir (ou calçar). Já para sistemas baseados em vídeos, tem-se que a principal característica é que os dados de entrada são imagens capturadas por câmeras 2D ou 3D. Isto evita que o usuário precise se vestir (ou se calçar) com sensores eletrônicos. Ainda assim, há trabalhos onde se opta pelo uso de uma luva total ou parcialmente colorida, sem sensores acoplados a ela. Neste caso, procura-se facilitar a detecção da região de interesse por meio do reconhecimento das luvas ou regiões coloridas nas imagens. Estas luvas vão funcionar como marcadores. Trabalhos mais recentes, entretanto, vêm apresentando a gravação das imagens sem utilização de nenhuma forma de marcador para detectar a região de interesse. Esta função é transferida para o processamento, no qual algoritmos de detecção da cor da pele detectam tanto as mãos quanto o rosto. A seguir são apresentados trabalhos baseados tanto em luvas quanto em vídeos. Para o caso dos sistemas baseados em vídeos, estes são gravados com o sinalizador de frente para a câmera, o que chamamos aqui de sinalização em 1^a pessoa.

1.2.2.1 Sistemas baseados em luvas

Sistemas baseados em luvas instrumentalizadas podem apresentar sensores com ou sem fios acoplados a si. Neste caso, as informações fornecidas como entrada para os sistemas de reconhecimento são obtidas a partir destes sensores que, em geral, são acelerômetros que coletam informações de velocidade, ângulo e po-

sição dos movimentos realizados. Dipietro, Sabatini e Dario (2008) apresentam um extenso trabalho com diversos sistemas que utilizam luvas e roupas especiais para as mais diversas aplicações, incluindo o reconhecimento em língua de sinais. Comparações entre técnicas de redes neurais e cadeia oculta de Markov são apresentadas em (PARVINI et al., 2009), com as informações de entrada capturadas pela luva comercial *CyberGlove*[8]. Tanto nos trabalhos apresentados em (KIM; JANG; BIEN, 1996) quanto em (YIN et al., 2009) e em (ZHANG et al., 2011) sensores são acoplados às luvas para gerar as informações sobre os movimentos realizados pelas mãos. Os sinais detectados são sempre dinâmicos para o caso destes sistemas baseados em luvas instrumentalizadas.

1.2.2.2 Sistemas baseados em vídeo

A entrada de informações em sistemas para reconhecimento baseados em vídeos são imagens gravadas por câmeras. Neste caso não há nenhum tipo de uso de luva com sensores eletrônicos. O que ocorre, principalmente em trabalhos menos atuais, é que marcadores baseados em cores podem ser utilizados para facilitar o trabalho de detecção das mãos. Portanto, luvas de tecido podem ser total ou parcialmente coloridas e, neste caso, são utilizadas como marcadores em vídeos para facilitar o trabalho de detecção da região de interesse. Trabalhos onde luvas coloridas devem ser calçadas são apresentados em (STARNER; PENTLAND, 1995), em (WANG et al., 2007), em (ARAN et al., 2009), em (HAN; AWAD; SUTHERLAND, 2009), em (ONG et al., 2012), em (COOPER et al., 2012) e em (MARAQA et al., 2012). Não há um padrão para o tipo ou a cor utilizada nas luvas. Observa-se que o importante é que se utilizem cores contrastantes com a cor da pele, tais como cores primárias. Os algoritmos computacionais terão mais facilidade em separar as regiões com luvas do restante do corpo se, por exemplo, um canal de cores for mais evidente que os outros.

Desafios como a remoção do fundo da imagem, a detecção da região de interesse (que são as mãos e rosto que podem ter os perfis de tamanho e cor tão diversos quanto a diversidade existente nas características físicas dos seres humanos), a luminosidade presente no local de gravação, a resolução dos vídeos gravados, são comuns aos sistemas baseados em vídeo. O custo computacional de se processamento de um vídeo à procura de regiões de interesse também é alto. Este item tem se tornado menos crítico, entretanto, com o crescente aumento da capacidade de processamento dos computadores pessoais. Assim, as pesquisas mais atuais apresentam trabalhos onde a imagem deve ser analisada sem nenhum tipo de marcador.

No caso do reconhecimento de línguas de sinais com imagens gravadas em 1^a pessoa, duas linhas de pesquisa podem ser encontradas: aquelas que detectam sinais

[8] http://www.cyberglovesystems.com/. Acessado em 06 de junho de 2014.

estáticos e, em grande parte dos casos, estes sinais são as letras do alfabeto local e aquelas que procuram o reconhecimento de sinais dinâmicos. Trabalhos em diversas línguas de sinais que tratam basicamente de extrair características de imagens das letras do alfabeto podem ser encontrados em (KULKARNI; LOKHANDE, 2010), em (LAHAMY; LICHTI, 2012) e em (OTINIANO, 2013) para a Língua Americana de Sinais. Já Incertis, Garcia-Bermejo e Casanova (2006) possuem como objeto de reconhecimento as letras do alfabeto da Língua Espanhola de Sinais. Em Língua Polonesa de Sinais tem-se pesquisas com sinais estáticos apresentadas por Flasiński e Myśliński (2010). Também encontram-se trabalhos nesta linha de detecção de sinais estáticos na Língua Persa de Sinais em (KARAMI; ZANJ; SARKALEH, 2011), Língua Australiana de Sinais em (SOLE; TSOEU, 2011) e Língua Indiana de Sinais em (GEETHA; MANJUSHA, 2012), dentre tantos outros trabalhos produzidos para outras línguas.

O problema de qual base de sinais utilizar para treinamento de sistemas automáticos de reconhecimento de sinais aparece em diversos trabalhos. Aqueles que utilizam pequenas bases de sinais são encontrados em (HUANG; HUANG, 1998), em (CUI; WENG, 2000), em (ZAHEDI; KEYSERS; NEY, 2005) e em (ARAN; AKARUN, 2010). Já nos trabalhos propostos em (ZIEREN; KRAISS, 2005), em (ZAKI; SHAHEEN, 2011), em (AGRIS; BLOMER; KRAISS, 2008), (KARMOKAR; ALAM; SIDDIQUEE, 2012) e em (CARIDAKIS et al., 2012), as bases de sinais são maiores. A grande maioria destas bases, sejam elas pequenas ou grandes, são para os respectivos países de origem dos autores das pesquisas. Assim, uma grande base diversa em sinais e nas formas como eles são gravados pode ser encontrada, embora não esteja concatenada.

Sensores RGB-D são utilizados em diversas gravações que compõem bases de gestos e movimentos corporais. No presente trabalho o sensor utilizado é o *Kinect*, desenvolvido pela Microsoft (2013). Ele encontra-se presente no mercado desde novembro de 2010, quando foi lançado como acessório do *videogame* XBOX para jogos em que se deseja detectar movimentos de até quatro jogadores ao mesmo tempo. Sua função de acessório de *videogame*, entretanto, ultrapassou sua vocação inicial e encontram-se as mais diversas aplicações que o utilizam na área acadêmica.

Parte destas aplicações envolvem a utilização do Kinect em conjunto com soluções baseadas em luvas, tais como nos trabalhos apresentados por Frati e Prattichizzo (2011) e Trindade, Lobo e Barreto (2012). Aplicações do Kinect como câmera de entrada em sistemas de reconhecimento de línguas de sinais são encontradas em (ZAFRULLA et al., 2011), (ONG et al., 2012), (LI, 2012), (PHADTARE; KUSHALNAGAR; CAHILL, 2012) e em (OSZUST; WYSOCKI, 2013). O grupo de pesquisa da própria fabricante do Kinect, o Microsoft Research, Microsoft (2013), mantém um sítio atualizado e disponibiliza soluções nas áreas de robótica, reconhecimento de movimentos,

jogos educacionais, entre outros, que utilizam o Kinect como sensor de profundidade.

Cabe abrir um parênteses neste momento para citar dois trabalhos que gravam a informação de movimento de forma distinta das apresentadas com o movimento realizado em 1^a pessoa. Os autores em (STARNER; WEAVER; PENTLAND, 1998) e em (BRASHEAR et al., 2003) colocam a câmera que captura as imagens montada na cabeça e, portanto, consideram a imagem gravada em 2^a pessoa. Isto significa que eles gravam os movimentos das mãos praticamente a partir dos olhos de quem tem a câmera montada na cabeça e esta mesma pessoa é o sinalizador. Embora os resultados de reconhecimento tenham acurácia acima de 95% nestes casos, a sinalização em 1^a pessoa é mais coerente para a análise que se faz neste trabalho, pois não necessita que nenhum equipamento seja acoplado ao sinalizador.

1.3 Objetivos

O objetivo geral desta tese é extrair características dos sinais a partir de vídeos RGB-D que sejam relacionadas à estrutura fonológica da Língua Brasileira de Sinais e que possam ser utilizadas em um sistema computacional de reconhecimento automático dos parâmetros fonológicos desta língua.

1.3.1 Objetivos Específicos

Para que o objetivo geral possa ser alcançado, os seguintes objetivos específicos foram estabelecidos:

- Realizar o estudo da estrutura da Língua Brasileira de Sinais, em especial a estrutura fonológica.
- A partir do conhecimento da fonologia da Língua Brasileira de Sinais, adaptar a sua complexidade, procurando simplificá-la de tal forma que um sistema baseado em vídeo seja capaz de reconhecer os parâmetros presentes nesta estrutura.
- Realizar o estudo das soluções de *software* e *hardware* existentes em Visão Computacional que permitem a gravação de sinais em 3D, uma vez que estes são realizados no espaço tridimensional.
- Selecionar sinais que sejam representativos dos cerca de 10 mil sinais existentes na Língua Brasileira de Sinais.
- Construir uma base com os sinais selecionados, gravados a partir da escolha de uma solução de *hardware* e *software* existente.

- Utilizar um processo de sumarização de vídeos para diminuir a redundância de informações existente entre os quadros dos vídeos.
- Extrair características dos vídeos sumarizados.
- Validar as características extraídas por meio de um sistema de classificação automático de parâmetros fonológicos.
- Implementar um sistema de classificação de sinais para validar a classificação de parâmetros fonológicos realizada.

1.4 Motivação

De acordo com o mais recente censo do IBGE (Instituto Brasileiro de Geografia e Estatística) realizado em 2010, a população brasileira possui 190.755.799 habitantes. Destes, 9.717.318 possuem deficiência auditiva [9] em algum grau. Portanto, cerca de 5,09% do total da população brasileira possui algum tipo de perda auditiva. Na Tabela 2 são apresentadas as quantidades de surdos existentes no Brasil, por faixa etária e divididos pelo grau de deficiência auditiva declarada: (i) não consegue escutar de modo algum; (ii) possui grande dificuldade em escutar; (iii) possui alguma dificuldade em escutar. Nota-se que a maior parte dos surdos atualmente é adulta e possui mais de 20 anos de idade. O acesso à língua de sinais, principalmente por parte das pessoas surdas, é uma das questões que se pode levantar aqui. Como ele se dá e de que forma?

O ensino da Língua Brasileira de Sinais ainda é bastante restrito a algumas instituições de surdos, em sua maioria instituições particulares, embora a Libras seja a 2^a língua oficial no Brasil. Atualmente há leis que prevêem uma carga horária mínima de aprendizagem da Libras em cursos de licenciatura e em órgãos particulares que prestam atendimento ao público em geral, tais como bancos. Estas leis pretendem capacitar mais falantes da língua de forma a melhorar o acesso à comunicação pelos surdos. Entretanto, é difícil encontrar, tanto em órgãos privados e mesmo em órgãos públicos, pessoas que saibam conversar de fato em Libras. Mesmo os surdos, muitas vezes, possuem dificuldades de toda ordem para aprendê-la. A falta de divulgação e não aceitação da língua vem de conceitos errôneos que existem sobre ela. Muitos tratam-na apenas como uma sucessão de mímicas. Outros entendem que é possível aprender apenas o alfabeto e utilizar datilologia [10] para conversar com um surdo. Estas

[9] o termo deficiência auditiva é utilizado na documentação do IBGE.

[10] Embora nos dicionários da Língua Portuguesa a datilologia seja definida como "a técnica de comunicação por meio de sinais feito com os dedos ou com as mãos, usada por surdos", a sua definição em Libras é levemente diferente. Em Libras, a datilologia é um sistema de representação das letras dos alfabetos das línguas orais-escritas por meio das mãos.

Grupos de idade	Não consegue de modo algum	Grande dificuldade	Alguma dificuldade
0 a 4 anos	13.593	10.996	54.453
5 a 9 anos	16.494	31.976	184.925
10 a 14 anos	22.379	45.914	235.471
15 a 17 anos	14.373	27.442	133.384
18 e 19 anos	10.463	17.122	86.439
20 a 24 anos	30.591	48.795	255.109
25 a 29 anos	31.146	53.492	288.966
30 a 34 anos	30.538	63.894	325.833
35 a 39 anos	26.753	70.325	362.784
40 a 44 anos	23.843	85.537	444.978
45 a 49 anos	18.724	97.630	529.426
50 a 54 anos	17.408	119.958	625.726
55 a 59 anos	15.520	130.589	668.086
60 a 64 anos	13.267	141.022	686.776
65 a 69 anos	11.925	147.136	678.305
70 a 74 anos	10.571	164.179	669.689
75 a 79	10.000	169.752	561.265
80 anos ou mais	26.618	373.207	782.529
Total	**344.206**	**1.798.967**	**7.574.145**

Tabela 2: População residente, com Tipo de deficiência auditiva por grupo de idade - Brasil 2010, IBGE.

ideias são apenas algumas daquelas que não condizem com a realidade da Libras, uma língua completa que representa a cultura de uma comunidade, permitindo aos surdos a comunicação natural e necessária ao seu desenvolvimento em meio a sociedade. A língua de sinais é sim a língua natural para comunicação entre surdos e entre surdos e ouvintes. E, como qualquer língua no mundo, seja ela oral, visual, espacial, escrita, demanda interesse e tempo de aprendizagem para que seja dominada.

A Tabela 3 traz informações importantes sobre a alfabetização dos surdos no Brasil. De acordo com o IBGE, “considerou-se como alfabetizada a pessoa capaz de ler e escrever um bilhete simples no idioma que conhecesse. Foi considerada analfabeta a pessoa que aprendeu a ler e escrever, mas que esqueceu devido a ter passado por um processo de alfabetização que não se consolidou e que apenas assinava o próprio nome.” Não se garante aqui que o idioma utilizado seja a Libras. Na verdade, fica claro que este idioma não se trata da Libras quando se diz “ler e escrever”, pois a leitura e escrita como em línguas orais-escritas não faz parte da estrutura da língua de sinais. Os índices percentuais de surdos alfabetizados parecem altos, mas resta questionar se esta alfabetização é realizada em sua língua natural.

Pensando nesta infinidade de situações, trabalhar na criação de sistemas automáticos que auxiliem a aprendizagem da língua de sinais por ouvintes e surdos,

Grupos de idade	Total	Alfabetizada	Percentual (%)
5 a 9 anos	233.395	146.406	62,73
10 a 14 anos	303.763	272.156	89,59
15 a 19 anos	289.223	267.385	92,45
20 a 24 anos	334.495	305.588	91,36
25 a 29 anos	373.604	335.392	89,77
30 a 34 anos	420.265	365.058	86,86
35 a 39 anos	459.862	391.697	85,18
40 a 44 anos	554.358	459.783	82,94
45 a 49 anos	645.779	532.142	82,40
50 anos ou mais	6.023.529	4.205.528	46,61
Total	9.638.276	7.281.134	75,54

Tabela 3: População residente, com Tipo de deficiência auditiva por alfabetização - Brasil 2010, IBGE.

nativos ou estrangeiros, torna-se uma tarefa de interesse geral, em um tempo em que a tecnologia encontra-se cada vez mais disponível e presente no cotidiano das pessoas.

O uso de um sensor RGB-D criado para jogos de videogame como câmera que captura os sinais vai de encontro a esta disponibilidade tecnológica. Ele é um sensor de baixo custo capaz de realizar gravações de intensidade de cor e profundidade por meio de seus sensores de cor CCD (Charge-Coupled Device - Dispositivo de Carga Acoplada) e sensores CMOS (Complementary Metaloxide Semiconductor) para a detecção de infravermelhos. Além do mais, ele foi desenvolvido para detectar movimentos em jogos de videogame e, portanto, é um equipamento de fácil utilização. Atualmente, a sua conexão a computadores pessoais com o objetivo de detecção de movimentos para as mais diversas aplicações tem sido bastante difundida.

1.5 Contribuições realizadas

As contribuições esperadas com o desenvolvimento de trabalho são:

- Estabelecer uma relação entre a estrutura fonológica e as características computacionais extraídas dos sinais que facilite a criação de sistemas de reconhecimento automático de sinais da Língua Brasileira de Sinais.
- Estabelecer uma diretriz para a construção de bases de sinais com diversos sinalizadores.
- Propor a sumarização dos vídeos por meio de um método de fácil implementação e alto desempenho computacional para vídeos de pequenos tamanhos.

1.6 Organização do trabalho

O restante deste trabalho é organizado como segue. O Capítulo 2 descreve a estrutura da Língua Brasileira de Sinais de forma resumida e as adaptações realizadas sobre esta estrutura para que um sistema baseado em visão possa ser implementado. Atenção especial é dada à composição fonológica da língua, que é o objeto de estudo deste trabalho.

No Capítulo 3, a criação da base de sinais utilizada neste trabalho tem suas especificações detalhadas. As ferramentas de *hardware* e *software* existentes são também apresentadas e, ao final, as escolhas realizadas para a criação da base de sinais são discutidas.

Já o Capítulo 4 apresenta outro aspecto importante para a implementação da extração de características neste trabalho: a sumarização dos vídeos dos sinais. Uma nova técnica, recém publicada e que aborda um problema clássico de otimização foi utilizada e é descrita neste capítulo.

O Capítulo 5 trata de forma detalhada das técnicas computacionais utilizadas para cada uma das características extraídas dos sinais. Três destes sinais são utilizados como exemplo para ilustrar cada uma destas características e a construção do vetor de características completo é apresentada ao final do capítulo. Este vetor é a entrada para um sistema de classificação de parâmetros fonológicos descrito no capítulo seguinte.

Assim, é no Capítulo 6 que o sistema de classificação de parâmetros fonológicos da Libras e os experimentos realizados para validação da metodologia proposta são descritos. Os resultados obtidos por este experimentos são discutidos ao final.

Por fim, o Capítulo 7 apresenta as conclusões deste trabalho e propostas futuras para a sua continuidade, que não são poucas.

2 Estrutura da Língua Brasileira de Sinais

Um breve resumo histórico da Língua Brasileira de Sinais é descrito neste capítulo. Esta descrição atua como pano de fundo para a descrição seguinte de parte da estrutura linguística da língua e a compreensão de sua complexidade. Neste sentido, o principal foco é contextualizar a estrutura fonológica da Libras dentro de uma estrutura maior. Realizada esta contextualização, parte-se para a análise e descrição das adaptações necessárias e realizadas neste trabalho especificamente com vistas ao seu reconhecimento por sistemas de visão computacional. Entretanto, deve ficar claro que não é o objetivo deste capítulo ou mesmo deste trabalho adentrar pelos meandros linguísticos e/ou questões filosóficas e sociais que envolvem as ideias sobre esta interessante língua.

2.1 Introdução

A história da Língua Brasileira de Sinais envolve uma série de ações ocorridas em períodos da história nos quais o Brasil, ainda Império e não República, buscava consolidar em sua população os ideais de nação por meio da educação formal para alunos sem deficiência e também para os deficientes. Naquele Brasil do século XIX a cegueira e a surdez eram as únicas deficências [1] reconhecidas pelo Estado como passíveis de uma abordagem com vistas às suas superações. Desta forma, o imperador Dom Pedro II criou dois institutos com o objetivo de permitir a instrução de crianças que possuíam tais deficiências, ambos na então capital do Império, a cidade do Rio de Janeiro.

Em 1854, por meio do decreto Imperial n^o 1.428, de 12 de setembro, foi criado o Imperial Instituto dos Meninos Cegos, tendo como modelo o Instituto de Meninos Cegos de Paris. Atualmente, em homenagem ao seu terceiro diretor, seu nome é Instituto Benjamim Constant (IBC)[2] e é um centro de referência com relação a quaisquer questões que envolvam a deficiência visual.

Já em 1856, por iniciativa do francês Ernest Huet, professor que tornou-se surdo ao longo da vida e era ex-diretor do Instituto de Surdos-Mudos de Bourges,

[1] O termo deficiência era utilizado naquele período para descrever tanto as deficências visuais, quanto as auditivas, quanto as mentais.

[2] http://www.ibc.gov.br/?itemid=89. Acessado em Julho de 2014.

na França, foi criado o Imperial Instituto dos Surdos-Mudos no Brasil. Na Lei nº 939 de 26 de setembro de 1857 constava que o Império subsidiaria o Instituto, inicialmente mantido por entidades particulares. Após algumas alterações em sua nomenclatura, hoje ele é denominado Instituto Nacional de Educação de Surdos (INES) [3]. A mudança mais significativa de suas denominações foi esta substituição da palavra *Mudo* pela palavra *Educação*, refletindo estudos de modernização da década de 50, no século XX. De fato, ainda é bastante comum que as pessoas surdas sejam tratadas também como mudas, embora esta não seja a realidade. A grande maioria das pessoas surdas é capaz de falar e não possui nenhuma deficiência física com relação à fala.

A Língua Brasileira de Sinais sofreu forte influência francesa, uma vez que o fundador do INES, Ernest Huet, era francês. Ela espalhou-se por todo o Brasil por meio dos alunos de diversos estados que o INES acolhia e que retornavam a estes concluídos seus estudos.

De acordo com o sítio oficial do INES, "(...) Outra ação importante para a difusão dessa língua em território brasileiro deu-se no ano de 1875, ocasião na qual o ex-aluno do Instituto, Flausino José da Gama, desenha o livro ***Iconographia dos Signaes dos Surdos-Mudos*** com cópias distribuídas para várias localidades do Brasil. A intenção principal era a de divulgar o meio pelos quais os surdos se comunicavam.(...)".

O reconhecimento da Língua Brasileira de Sinais como língua oficial no Brasil deu-se, entretanto, somente no ano de 2002, por meio da Lei nº 10.436, de 24 de abril de 2002, que "dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências."

Não é objetivo deste trabalho destrinchar tal Lei. Entretanto, vale ressaltar aqui seus artigos 1º e 4º:

- Art. 1º - É reconhecida como meio legal de comunicação e expressão a Língua Brasileira de Sinais - Libras e outros recursos de expressão a ela associados.
  Parágrafo único. Entende-se como Língua Brasileira de Sinais - Libras a forma de comunicação e expressão, em que o sistema linguístico de natureza visual-motora, com estrutura gramatical própria, constituem um sistema linguístico de transmissão de ideias e fatos, oriundos de comunidades de pessoas surdas do Brasil.

- Art. 4º - O sistema educacional federal e os sistemas educacionais estaduais, municipais e do Distrito Federal devem garantir a inclusão nos cursos de formação de Educação Especial, de Fonoaudiologia e de Magistério, em seus níveis

[3] http://www2.ines.gov.br/. Acessado em Julho de 2014.

médio e superior, do ensino da Língua Brasileira de Sinais - Libras, como parte integrante dos Parâmetros Curriculares Nacionais - PCNs, conforme legislação vigente.

Parágrafo único. A Língua Brasileira de Sinais - Libras não poderá substituir a modalidade escrita da Língua Portuguesa.

No Art. 1^o fica clara a colocação da Libras não como uma linguagem, mas como uma língua com todas as características e estruturas necessárias para ser tratada como tal. No Art. 4^o , o governo impõe por meio da lei a divulgação da Libras para que seja possível sua aprendizagem por toda ou grande parte da população brasileira, seja ela ouvinte ou surda. Desta forma, procura incluir as pessoas surdas em todos os meios, garantindo a estas a aprendizagem e pensamento por meio de sua língua natural.

2.2 Estrutura linguística da Língua Brasileira de Sinais

A linguística é o estudo científico das línguas naturais e humanas. Estam podem ser entendidas de duas formas: ou nascem arbitrariamente ou nascem com o homem. Diferentes linhas de estudo e pensamento linguístico surgem a partir destas duas abordagens, chamadas naturalista e estruturalista, respectivamente. Neste trabalho, procura-se não adentrar em questões de cunho linguístico, mas ater-se aos conceitos das áreas da linguística que definem a estrutura da Libras neste contexto. Assim, as seguintes áreas da linguística, cujas definições são válidas também para a Libras, possuem destaque:

- A fonética, que estuda os sons como entidades físico-articulatórias isoladas. Tem por objetivo estabelecer um conjunto de traços, ou propriedades, que possam descrever todos os sons utilizados na linguagem humana. A unidade da fonética é o som da fala ou fone.

- A fonologia, que estuda os sons do ponto de vista funcional como elementos que integram um sistema linguístico determinado. A ela cabe estudar as diferenças fônicas intencionais, distintivas, que se vinculam a diferenças de significação e, além disso, estabelecer como se relacionam entre si os elementos de diferenciação e as condições em que se combinam uns com os outros para formar morfemas, palavras e frases. A unidade da fonologia é o fonema.

  O trabalho descrito nesta tese explora a estrutura fonológica da Libras como geradora de informações que permitam a um sistema de visão computacional reconhecer os parâmetros que compõem a Libras. Portanto, a descrição sobre a fonologia da Libras acontece com mais detalhes nas seções a seguir.

- A morfologia, que estuda a estrutura interna das palavras, ou seja, da combinação entre os elementos que formam as palavras. As diversas formas que as palavras apresentam quanto à categoria de número, gênero, tempo e pessoa são tratadas nesta área.
- A sintaxe, que estuda a estrutura da frase, ou seja, da combinação das unidades significativas da frase. A sintaxe trata das funções, das formas e das partes do discurso.
- A semântica, que estuda o significado da palavra e da sentença. Ela trata da natureza, da função e do uso dos significados determinados ou pressupostos.
- A pragmática, que estuda a linguagem em uso (contexto) e os princípios de comunicação.
- As interfaces da linguística, que apresentam as interações da linguística com outras ciências e domínios.

2.3 Estrutura fonológica da Língua Brasileira de Sinais e sua adaptação para o problema de Visão Computacional

Os elementos que constituem a estrutura fonológica da Libras são chamados parâmetros. A seguir eles são descritos de dois pontos de vista distintos. Primeiramente, do ponto de vista da linguística, no qual utiliza-se como referência o trabalho apresentado em (QUADROS; KARNOPP, 2004). O ponto de vista apresentado em seguida retrata as adaptações realizadas para esta tese em cada um destes parâmetros, com vistas à sua utilização no sistema de visão computacional que é proposto neste trabalho.

2.3.1 A fonologia em Língua Brasileira de Sinais

A definição do termo fonologia, conforme descrito na seção anterior, pode levar-nos a crer que ela possa ser aplicada apenas a línguas orais-escritas, uma vez que a unidade da fonética, fono, está necessariamente relacionada a sons. Pode-se crer, portanto, que não cabe este termo no estudo das línguas de sinais, que são visual-espaciais. Entretanto, a língua de sinais é uma língua completa em todos os aspectos fonéticos, fonológicos, morfológicos, sintáticos, semânticos e pragmáticos de uma língua, independente da modalidade. Assim, todos estes estudos aplicam-se às línguas de sinais, variando-se as abordagens devido às diferenças na modalidade de cada língua.

Segue a definição de fonologia no contexto de línguas de sinais presente no trabalho de Quadros e Karnopp (2004):

> "(...) A fonologia das línguas de sinais é o ramo da linguística que objetiva identificar a estrutura e organização dos constituintes fonológicos, propondo modelos descritivos e explanatórios. A primeira tarefa da fonologia para línguas de sinais é determinar quais são as unidades mínimas que formam os sinais. A segunda tarefa é estabelecer quais são os padrões possíveis de combinação entre essas unidades e as variações possíveis no ambiente fonológico. (...)"

Ainda sobre a utilização do termo fonologia em línguas de sinais, Quadros e Karnopp (2004) completa:

> "(...) As línguas de sinais são denominadas línguas de modalidade gestual-visual (ou espaço-visual), pois a informação linguística é recebida pelos olhos e produzida pelas mãos. Apesar da diferença existente entre línguas de sinais e línguas orais, no que concerne à modalidade de percepção e produção, o termo "fonologia" tem sido usado para referir-se também ao estudo dos elementos básicos das línguas de sinais. Historicamente, entretanto, para marcar a diferença entre esses dois tipos de sistemas linguísticos, Stokoe (1960) propôs o termo "quirema" às unidades formacionais dos sinais (configuração de mãos, locação e movimento) e, ao estudo de suas combinações, propôs o termo "quirologia" (do grego "mão"). Outros pesquisadores , incluindo Stokoe em edição posterior (1978) , têm utilizado os termos "fonema" e "fonologia", estendendo seus significados de modo a abarcar a realização linguística visual-espacial. (...)"

Assim, é na fonologia que estudam-se as unidades formacionais dos sinais que são produzidos basicamente pelos movimentos das mãos em conjunto com movimentos do corpo e da face. Seus principais parâmetros fonológicos inicialmente estudados em Stokoe foram a locação (ponto de articulação), movimento e configuração das mãos. Entretanto, em trabalhos posteriores e mais recentes, os parâmetros orientação da palma da mão e expressões não-manuais foram incluídos, pela sua importância na definição e distinção entre os sinais. A Figura 2 apresenta a marcação destes parâmetros.

Assim, os estudos recentes sobre a língua de sinais apresentam os cinco parâmetros fonológicos seguintes:

1. Locação (L), também chamado ponto de articulação (PA)

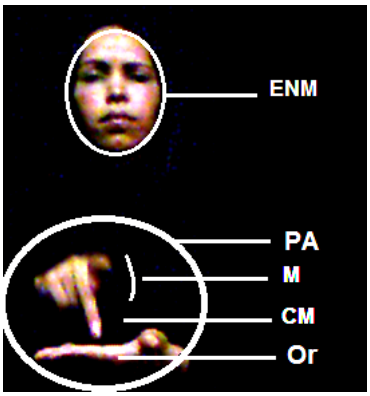


Figura 2: Parâmetros fonológicos da Língua Brasileira de Sinais (baseado em (QUADROS; KARNOPP, 2004)). ENM: expressões não-manuais. PA: ponto de articulação. M: movimento. CM: configuração de mão. Or: orientação da palma da mão.

2. Configuração de mão (CM)
3. Movimento (M)
4. Orientação da palma da mão (Or)
5. Expressões não-manuais (ENM).

As definições, os possíveis atributos e respectivos valores para cada um destes atributos são descritos a seguir. Primeiramente apresenta-se a abordagem linguística para cada parâmetro, conforme descrito em (QUADROS; KARNOPP, 2004) e, em seguida, as adaptações estabelecidas neste trabalho para simplificação dos parâmetros que são utilizados na implementação de um sistema de classificação destes.

2.3.2 Ponto de articulação (PA)

A locação (ou ponto de articulação), "é aquela área do corpo, ou no espaço de articulação definido pelo corpo em que ou perto do qual o sinal é articulado", conforme Friedman (1976 apud QUADROS; KARNOPP, 2004). Há um espaço de enunciação definido como uma área que contém todos os pontos dentro do raio de alcance das mãos em que os sinais são articulados.

De acordo com Ferreira-Brito (1995 apud QUADROS; KARNOPP, 2004) as locações principais são divididas em quatro regiões: cabeça, mão, tronco e espaço neutro. A Figura 3 apresenta estas locações. Os valores possíveis para cada uma delas são descritos na Tabela 4.

A adaptação realizada neste trabalho, para este parâmetro, consiste na divisão do espaço em três áreas, que são os atributos para este parâmetro: (i) cabeça, (ii) ombro e (iii) tronco. Para cada uma destas áreas são possíveis os valores direita, centro e esquerda, conforme pode-se ver na Figura 4, para o sinal "Verdade".

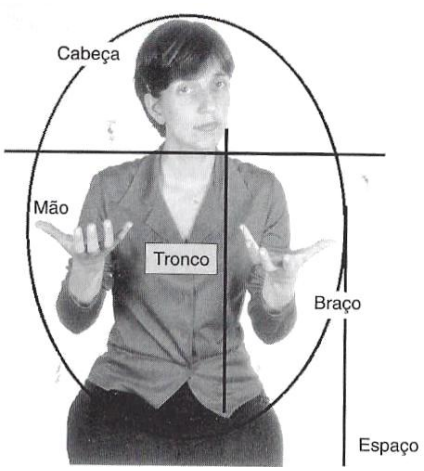


Figura 3: As quatro áreas principais de articulação dos sinais, conforme Battison (1974 apud QUADROS; KARNOPP, 2004, p. 49), no espaço de enunciação dos sinais.

Locação (Ponto de articulação)	
CABEÇA	{topo da cabeça, testa, rosto, parte superior do rosto, parte inferior do rosto, orelha, olhos, nariz, boca, bochechas, queixo}
TRONCO	{pescoço, ombro, busto, estômago, cintura, braços, braço, antebraço, cotovelo, pulso}
MÃO	{palma, costas das mãos, lado do indicador, lado do dedo mínimo, dedos, ponta dos dedos, dedo mínimo, anular, dedo médio, indicador, polegar}
ESPAÇO NEUTRO	

Tabela 4: Locações definidas por Ferreira-Brito (1995 apud QUADROS; KARNOPP, 2004).

2.3.3 Configuração de mão (CM)

O parâmetro Configuração de mão trata da forma tanto da mão direita quanto da mão esquerda em um sinal. Há sinais nos quais apenas uma das mãos é utilizada, assim como há sinais nos quais ambas as mãos produzem o movimento em conjunto. Tanto a mão direita quanto a esquerda podem ser a mão dominante em um sinal. Em geral, quando o sinalizador é destro, ele utiliza a mão direita como referência e, quando ele é canhoto, utiliza a mão esquerda. Portanto, na Língua Brasileira de Sinais, o sinal pode ser realizado com a mão dominante de preferência do sinalizador.

Não há um consenso entre os pesquisadores sobre a quantidade de formas possíveis para as mãos na Língua Brasileira de Sinais. Assim, de acordo com Quadros e Karnopp (2004) há 46 possíveis configurações de mão. Em seu livro, Qua-

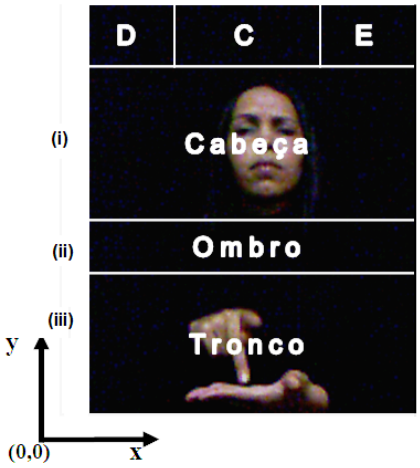


Figura 4: Atributos e valores para o parâmetro Ponto de articulação. Cada um dos 3 atributos (cabeça, ombro e tronco) pode assumir individualmente os valores D (direita), C (centro) e E (esquerda).

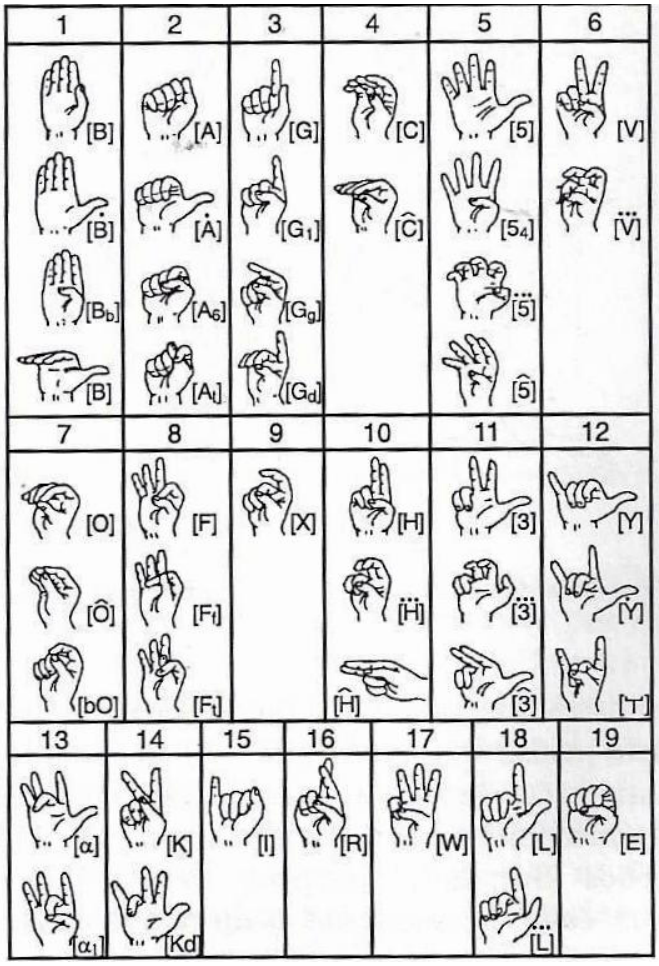


Figura 5: As 46 configurações de mão (CMs) da Língua Brasileira de Sinais de acordo com Ferreira-Brito (1995).

dros e Karnopp (2004) reproduzem, na realidade, a quantidade e as formas definidas em Ferreira-Brito (1995 apud QUADROS; KARNOPP, 2004). A Figura 5 apre-



Figura 6: Os 13 agrupamentos para as 134 configurações de mão propostas em (**??**). As configurações de mão são agrupadas por semelhança em sua forma.

senta cada uma destas configurações possíveis agrupadas verticalmente por semelhança de forma e com o respectivo fonema. Cada língua de sinais tem suas próprias formas para a configuração de mãos mas, neste caso, as configurações de mãos assemelham-se àquelas da Língua Americana de Sinais.

Em (**??**), os autores conduzem trabalhos mais recentes acerca da quantidade e dos agrupamentos possíveis para as configurações de mão. Em seu trabalho os autores têm como objetivo documentar sinais em um sistema computacional interativo, onde cada sinal terá sua configuração de mão descrita conforme 134 configurações propostas e organizadas em 13 grupos formados por similaridade de formato. Optou-se nesta tese pela utilização das configurações de mão descritas no trabalho de **??**) como referência, por haver mais formatos e por ser o trabalho mais recente que o descrito por Ferreira-Brito (1995). As 134 formas possíveis, organizadas em 13 grupos, podem ser vistas na Figura 6.

A abordagem proposta nesta tese para o reconhecimento do parâmetro Configuração de mão é definida pelos seguintes três atributos:

- (i) No primeiro atributo deve constar o grupo ao qual pertence a configuração de mão (G1 a G13).

- (ii) O alinhamento do punho ao longo da execução do sinal é o valor relativo ao segundo atributo. O punho pode estar alinhado de acordo com os eixos ordenados x, em y ou em z.

- (iii) Por fim, o terceiro atributo deve conter a informação de variação da configuração de mão durante a execução do gesto. Este atributo pode assumir, portanto, apenas dois valores: sim ou não.

2.3.4 Movimento (M)

Em (QUADROS; KARNOPP, 2004), o movimento é definido como "um parâmetro complexo que pode envolver uma vasta rede de formas e direções, desde os movimentos internos da mão, os movimentos do pulso e os movimentos direcionais no espaço."

Os movimentos identificados na Língua Brasileira de Sinais por Ferreira-Brito (1990) referem-se ao tipo, direcionalidade, maneira e frequência destes. Assim, o movimento pode estar nas mãos, pulsos e antebraço; os movimentos podem ser unidirecionais, bidirecionais ou multidirecionais; a maneira do movimento é a categoria que descreve a qualidade, a tensão e a sua velocidade e a frequência do movimento refere-se ao seu número de repetições. Os valores de cada um dos atributos deste parâmetro estão descritos na Tabela 5.

É possível notar nesta tabela a grande quantidade de detalhes somente para a definição do parâmetro fonológico Movimento. Nesta tese, a escolha realizada para o reconhecimento deste parâmetro teve por objetivo simplificá-lo de tal forma que um sistema computacional possa distinguir entre dois de seus atributos, os quais são descritos a seguir:

1. (i) O atributo direcional Tipo, com os seguintes valores possíveis: (1) para cima, (2) para baixo, (3) para a direita, (4) para a esquerda, (5) para dentro ou (6) para fora. Cada um dos seis valores possíveis para este atributo pode ser visto na Tabela 6. Nela, três quadros do vídeo gravado para o sinal são exibidos indicando o início, meio e final do movimento das mãos. Vale ressaltar que para os casos 2, 3, 5 e 6, sinais "Esperto", "Pessoa", "Ontem" e "Futuro", respectivamente, o movimento se refere à mão direita, uma vez que a mão esquerda não participa destes sinais. Já para o caso 1, sinal "Comemorar", o movimento para cima é realizado por ambas as mãos. E, por fim, para o caso 4, sinal "Engordar", o movimento é realizado para a esquerda pela mão esquerda e para a direita se a referência é a mão direita.

Categorias do parâmetro Movimento	
TIPO	
Contorno ou forma geométrica	{retilíneo, helicoidal, circular, semicircular, sinuoso, angular, pontual}
Interação	{alternado, de aproximação, de separação, de inserção, cruzado}
Contato	{de ligação, de agarrar, de deslizamento, de toque, de esfregar, de riscar, de escovar ou de pincelar}
Torcedura do pulso	{rotação, com refreamento}
Dobramento do pulso	{para cima, para baixo}
Interno das mãos	{abertura, fechamento, curvamento e dobramento (simultâneo/gradativo)}
DIRECIONALIDADE	
Direcional	{unidirecional: para cima, para baixo, para a direita, para a esquerda, para dentro, para fora, para o centro, para a lateral inferior esquerda, para a lateral inferior direita, para a lateral superior esquerda, para a a lateral superior direita, para específico ponto referencial}
	{bidirecional: para cima e para baixo, para a esquerda e para a direita, para dentro e para fora, para laterais opostas - superior direita e inferior esquerda}
Não-direcional	
MANEIRA	
Qualidade, tensão e velocidade	{contínuo, de retenção, refreado}
FREQUÊNCIA	
Repetição	{simples, repetido}

Tabela 5: Categorias do parâmetro Movimento na Língua Brasileira de Sinais segundo Ferreira-Brito (1990).

2. (ii) A frequência é o segundo atributo deste parâmetro, possuindo os seguintes valores: simples ou repetido.

2.3.5 Orientação da palma da mão (Or)

Nos estudos iniciais das línguas de sinais realizados por Stokoe, a orientação da palma da mão não foi considerada como um parâmetro cujos valores sejam capazes de distinguir os sinais. Posteriormente, entretanto, constatou-se que as mudanças de significado no sinal ocorriam também pela mudança de orientação da palma da mão. Assim, a orientação hoje é um dos parâmetros fonológicos das línguas de sinais. Ela é a direção para a qual a palma da mão aponta ao longo da produção do sinal. Na

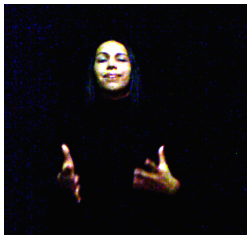
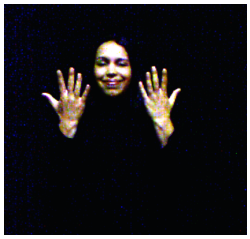
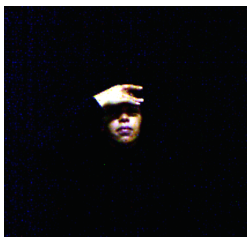
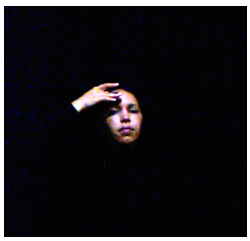
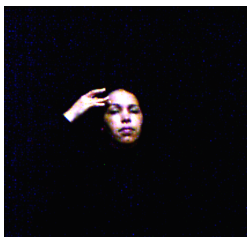
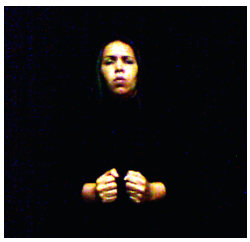
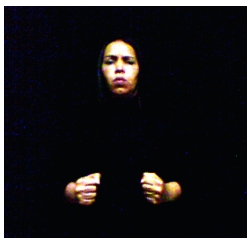
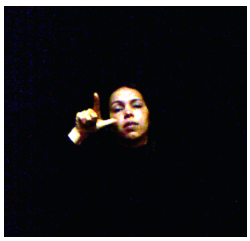
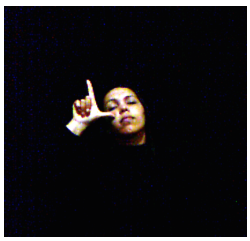
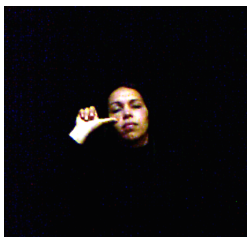
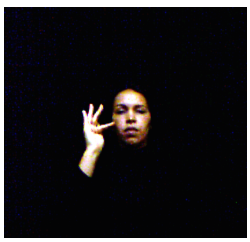
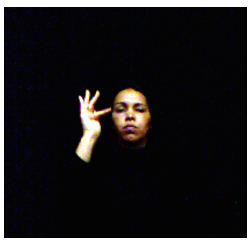
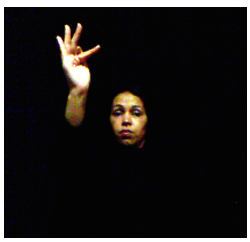
Tipo de Movimento			
1. Para cima			
2. Para baixo			
3. Para a direita			
4. Para a esquerda			
5. Para dentro			
6. Para fora			

Tabela 6: Atributos e valores para o parâmetro Movimento. (1) Para cima: sinal “Comemorar”. (2) Para baixo: sinal “Esperto”. (3) Para a direita: sinal “Pessoa”. (4) Para a esquerda: sinal “Engordar”. (5) Para dentro: sinal “Ontem”. (6) Para fora: sinal ”Futuro“.

Orientação da palma da mão	{para cima, para baixo, para dentro, para fora, para o lado (contralateral), para o lado (ipsilateral)}

Tabela 7: Valores possíveis para o parâmetro Orientação da palma da mão.

Figura 7 vê-se a imagem contendo todas as seis orientações possíveis da palma da mão. Já na Tabela 7 estes seis valores possíveis para o parâmetro são descritos.



Figura 7: Orientações da palma da mão, conforme Quadros e Karnopp (2004).

A adaptação realizada para este parâmetro fonológico constituiu-se em estabelecer valores para dois atributos da orientação da palma da mão. São eles:

- (i) O primeiro atributo, denominado orientação da palma da mão, difere dos valores linguísticos estabelecidos em (QUADROS; KARNOPP, 2004) por um valor. Seus possíveis valores são (1) para cima, (2) para baixo, (3) para dentro, (4) para fora, (5) para o lado. Não realiza-se neste trabalho a distinção do valor para o lado entre contralateral ou ipsilateral. As cinco variações de palma de mão possíveis para este atributo podem ser vistos na Figura 8. No caso 1, o sinal realizado, “Verdade”, apresenta a orientação da palma da mão esquerda. Para os outros casos, 2, 3, 4 e 5, respectivamente os sinais “Espalhar”, “Sumir”, “Olhar” e “Amar”, a mão cuja palma é apresentada é a direita.

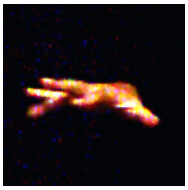
Orientação da palma da mão				
1. Para cima	2. Para baixo	3. Para dentro	4. Para fora	5. Para o lado
				

Tabela 8: Atributos e valores para o parâmetro Orientação da palma da mão. (1) Para cima: sinal “Verdade”. (2) Para baixo: sinal “Espalhar”. (3) Para dentro: sinal “Sumir”. (4) Para fora: sinal “Olhar”. (5) Para o lado: sinal “Amar”.

- (ii) O segundo atributo do parâmetro orientação da palma da mão possui valores que informam se há ou não mudança de orientação da palma da mão para aquele sinal ao longo de sua execução. Assim, este atributo pode assumir dois valores: sim ou não.

2.3.6 Expressões não-manuais (ENM)

O quinto parâmetro fonológico de Libras são as expressões não-manuais (ENM). Elas são formadas pelo movimento da face, dos olhos, da cabeça ou do tronco e possuem dois papéis na definição de cada sinal. As expressões não-manuais tanto podem marcar construções sintáticas quanto diferenciar itens lexicais. Como construção sintática elas podem marcar sentenças interrogativas, orações relativas, topicalizações, concordância e foco. Já com a função de diferenciar itens lexicais marcam referência específica, referência pronominal, partícula negativa, advérbio, grau ou aspecto.

Na Tabela 9 encontram-se identificados os valores possíveis para o parâmetro Expressões não-manuais. Neste caso é possível que dois valores ocorram simultaneamente, tal qual as marcas de interrogação e negação. As expressões não-manuais não são objeto de estudo desta tese.

2.3.7 Resumo dos atributos e valores para os quatro parâmetros adaptados

A Tabela 10 apresenta um resumo com os quatro parâmetros fonológicos linguísticos adaptados para o problema de extração de características descrito nesta tese. Cada um dos parâmetros possui diversos atributos próprios e cada atributo possui os valores descritos na citada tabela.

Expressões não-manuais		
ROSTO	Parte superior	{sobrancelhas franzidas, olhos arregalados, lance de olhos, sobrancelhas levantadas}
	Parte inferior	{bochechas infladas, bochechas contraídas, lábios contraídos e projetados e sobrancelhas franzidas, correr da língua contra a parte inferior interna da bochecha, apenas bochecha direita inflada, contração do lábio superior, franzir do nariz}
CABEÇA	{balanceamento para frente e para trás (sim), balanceamento para os lados (não), inclinação para a frente, inclinação para o lado, inclinação para trás}	
ROSTO E CABEÇA	{cabeça projetada para a frente, olhos levemente cerrados, sobrancelhas franzidas; cabeça projetada atrás e olhos arregalados}	
TRONCO	{para a frente, para trás, balanceamento alternado dos ombros, balanceamento simultâneo dos ombros, balanceamento de um único ombro}	

Tabela 9: Expressões não-manuais da Língua Brasileira de Sinais, definidas por Ferreira-Brito (1995).

	Parâmetros	**Atributos**	**Valores**
(i)	Ponto de articulação	cabeça	{*direita, centro, esquerda*}
		ombro	{*direita, centro, esquerda*}
		tronco	{*direita, centro, esquerda*}
(ii)	Configuração das mãos	grupo	{*1, ..., 13*}
		eixo de alinhamento	{*em x, em y, em z*}
		variação	{*sim, não*}
(iii)	Movimento	tipo	{*para cima, para baixo, para a direita,* {*para a esquerda, para dentro, para fora*}
		frequência	{*simples, repetido*}
(iv)	Orientação da palma da mão	orientação	{*para cima, para baixo, para dentro, para fora, para o lado*}
		variação	{*sim, não*}

Tabela 10: Parâmetros da estrutura fonológica da Libras originalmente proposta por Quadros e Karnopp (2004) e resumida nesta tese. Cada parâmetro consiste de um conjunto de atributos com os seus valores possíveis discriminados na coluna mais à direita.

2.4 Sinais selecionados

Após a descrição da estrutura fonológica da Língua Brasileira de Sinais conforme descrito em (QUADROS; KARNOPP, 2004) e da adaptação destes parâmetros com vistas à extração de características em visão computacional, foram selecionados sinais para composição de uma base de sinais de tal forma que experimentos de reconhecimento de parâmetros e sinais possam ser executados.

Existem mais de dez mil sinais presentes na Língua Brasileira de Sinais. Com o objetivo de selecionar sinais que sejam representativos de todo este corpo, utilizou-se

a classificação de cada sinal descrita no Dicionário Enciclopédico Trilíngue, conforme Capovilla, Raphael e Maurício (2012a), Capovilla, Raphael e Maurício (2012b). Neste dicionário são descritos 34 morfemas[4] que representam grupos de sinais similares em significado e forma. Um sinal foi escolhido em cada agrupamento de morfemas proposto para representar todo o agrupamento. O critério de seleção baseou-se em um cruzamento dos sinais do Dicionário Enciclopédico Trilíngue com o Dicionário da Língua Brasileira de Sinais, em sua versão *online* como em (SURDOS, 2008). Cada um dos 34 sinais selecionados para compor a base existe com o mesmo formato em ambas as bases.

A Figura 8 apresenta cada um destes 34 sinais selecionados que compõem a base de sinais criada nesta tese e descrita em maiores detalhes no Capítulo 3.

Cada um destes 34 sinais tem seus parâmetros fonológicos adaptados definidos tanto para a mão direita quanto para a mão esquerda. As Figuras 9 e 10 apresentam os atributos e respectivos valores para cada um dos 34 sinais. O valor do atributo de um parâmetro quando a mão não participa de determinado sinal está marcado por um asterisco. Isto acontece para a mão esquerda em 12 dos 34 sinais selecionados. Nestes casos, apenas o parâmetro Ponto de articulação é mantido, com o valor “à esquerda do tronco”. No mais, cada valor em cada atributo dos 4 parâmetros é marcado com um “x” em fundo preto.

As Figuras 9 e 10 apresentam, assim, os valores de saída desejados para cada sinal. Eles são a referência para o sistema supervisionado de classificação de parâmetros apresentado no Capítulo 6.

2.5 Considerações finais

A estrutura fonológica descrita em (QUADROS; KARNOPP, 2004) é bastante extensa e constituída de detalhes sutis. Realizar a classificação para cada sinal utilizando esta estrutura fonológica como se coloca em (QUADROS; KARNOPP, 2004) é uma tarefa complexa mesmo para especialistas em línguas de sinais e torna-se ainda mais complexa para sistemas computacionais automáticos, como proposto nesta tese.

Assim, a adaptação de cada um dos parâmetros fonológicos linguísticos realizado pautou-se na simplificação destes de tal forma que ainda seja possível fazer a distinção entre os 4 parâmetros fonológicos relacionados às mãos: o ponto de articulação, a configuração da mão, o movimento e a orientação da palma da mão.

[4] Um morfema é a menor unidade sublexical que codifica significado.

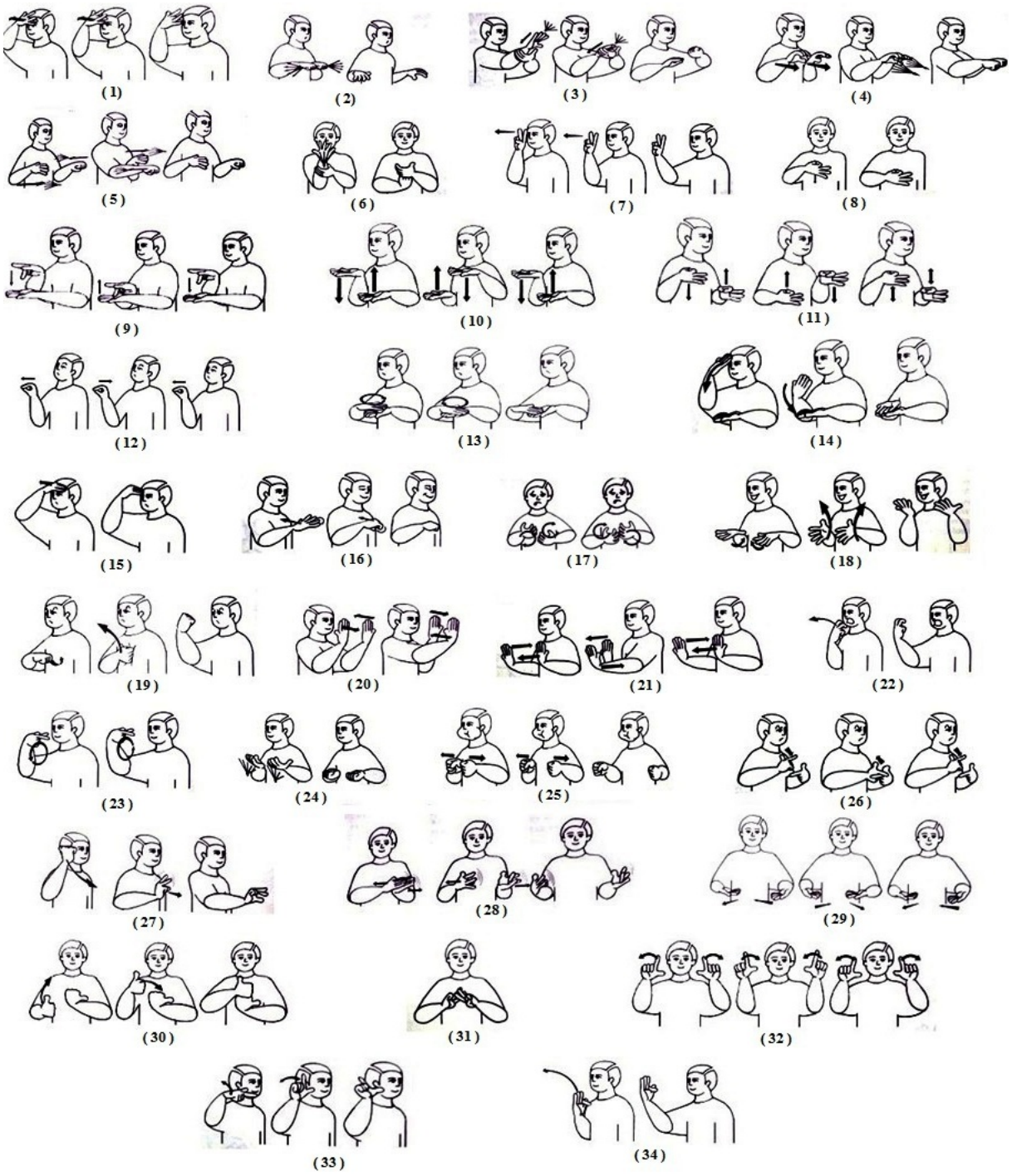


Figura 8: Os 34 sinais gravados na base de sinais e utilizados nesta tese. (1)“Pessoa”, (2)“Espalhar”, (3)“Copiar”, (4)“Agarrar”, (5)“Recolher”, (6)“Sumir”, (7)“Olhar”, (8)“Justo”, (9)“Verdade”, (10)“Peso”, (11)“Justiça”, (12)“Quem”, (13)“Nada”, (14)“Acreditar”, (15)“Esquecer”, (16)“Amar”, (17)“Angustiado”, (18)“Comemorar”, (19)“Rancor”, (20)“Assembleia”, (21)“Comparar”, (22)“Gritar”, (23)“Falar”, (24)“Absorver”, (25)“Engordar”, (26)“Brigar”, (27)“Esperto”, (28)“Brilhar”, (29)“Empregado”, (30)“Substituir”, (31)“Prisão”, (32)“Televisão”, (33)“Ontem”, (34)“Futuro”. As figuras de cada sinal foram editadas a partir de Capovilla, Raphael e Maurício (2012a), Capovilla, Raphael e Maurício (2012b).

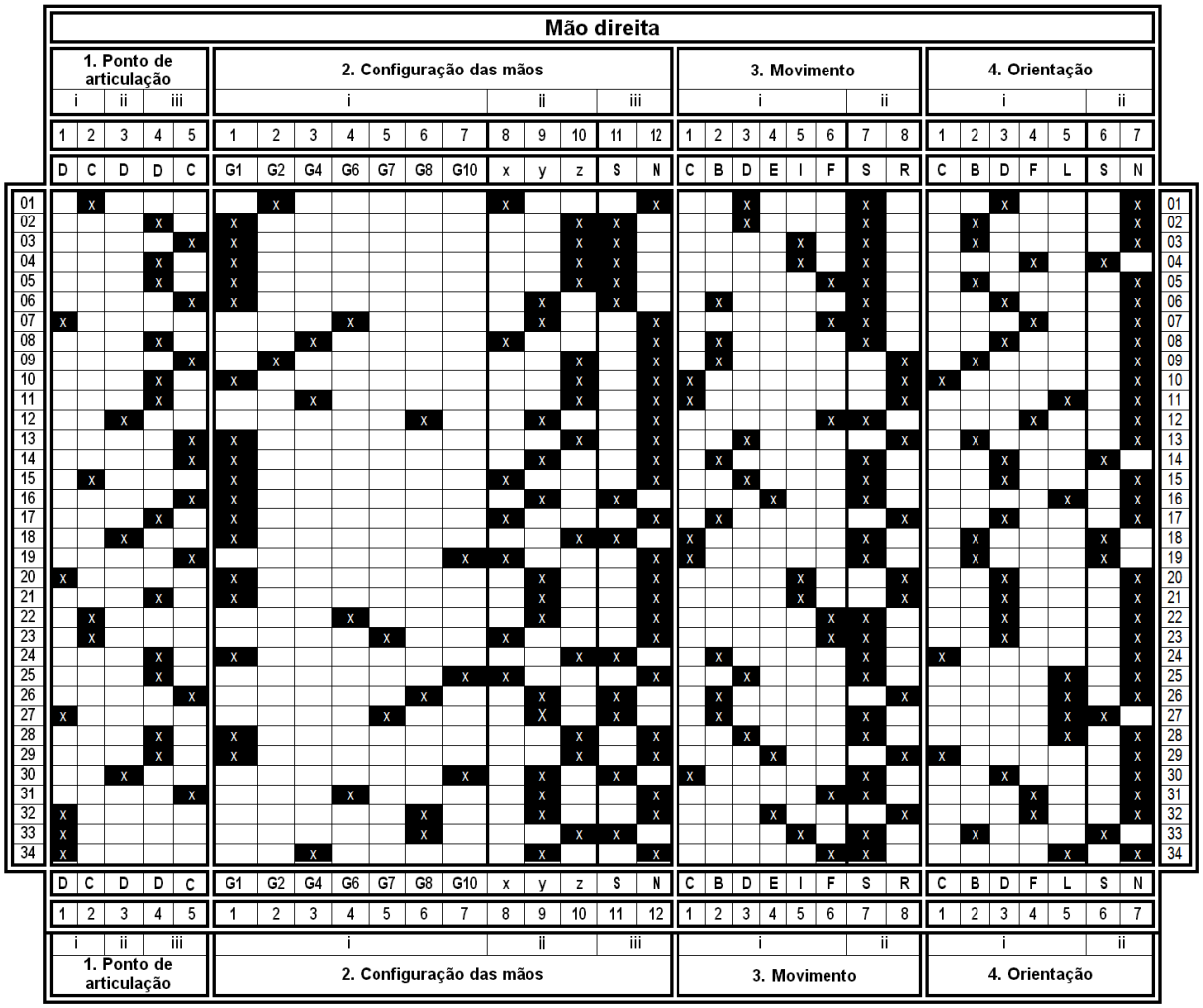


Figura 9: Valores dos atributos dos parâmetros fonológicos resumidos para a mão direita dos 34 sinais selecionados. (1) *Ponto de articulação*: (i) 1. À direita da cabeça (D). 2. Ao centro da cabeça (C). (ii) 3. À direita do ombro (D). (iii) 4. À direita do tronco (D). 5. Ao centro do tronco (C). / (2) *Configuração da mão*: (i) 1. G1. 2. G2. 3. G4. 4. G6. 5. G7. 6. G8. 7. G10. (ii) 8. Alinhamento no eixo x (x). 9. Alinhamento no eixo y (y) . 10. Alinhamento no eixo z (z). (iii) 11. Mudança na configuração da mão ou alinhamento: sim (S). 12. Mudança na configuração da mão ou alinhamento: não (N). / (3) *Movimento*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para a direita (D). 4. Para a esquerda (E). 5. Para dentro (I). 6. Para fora (F). (ii) 7. Frequência: simples (S). 8. Frequência: repetida (R). / (4) *Orientação da palma da mão*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para dentro (D). 4. Para fora (F). 5. Para o lado (L). (iii) 6. Variação: sim (S). 7. Variação: não (N).

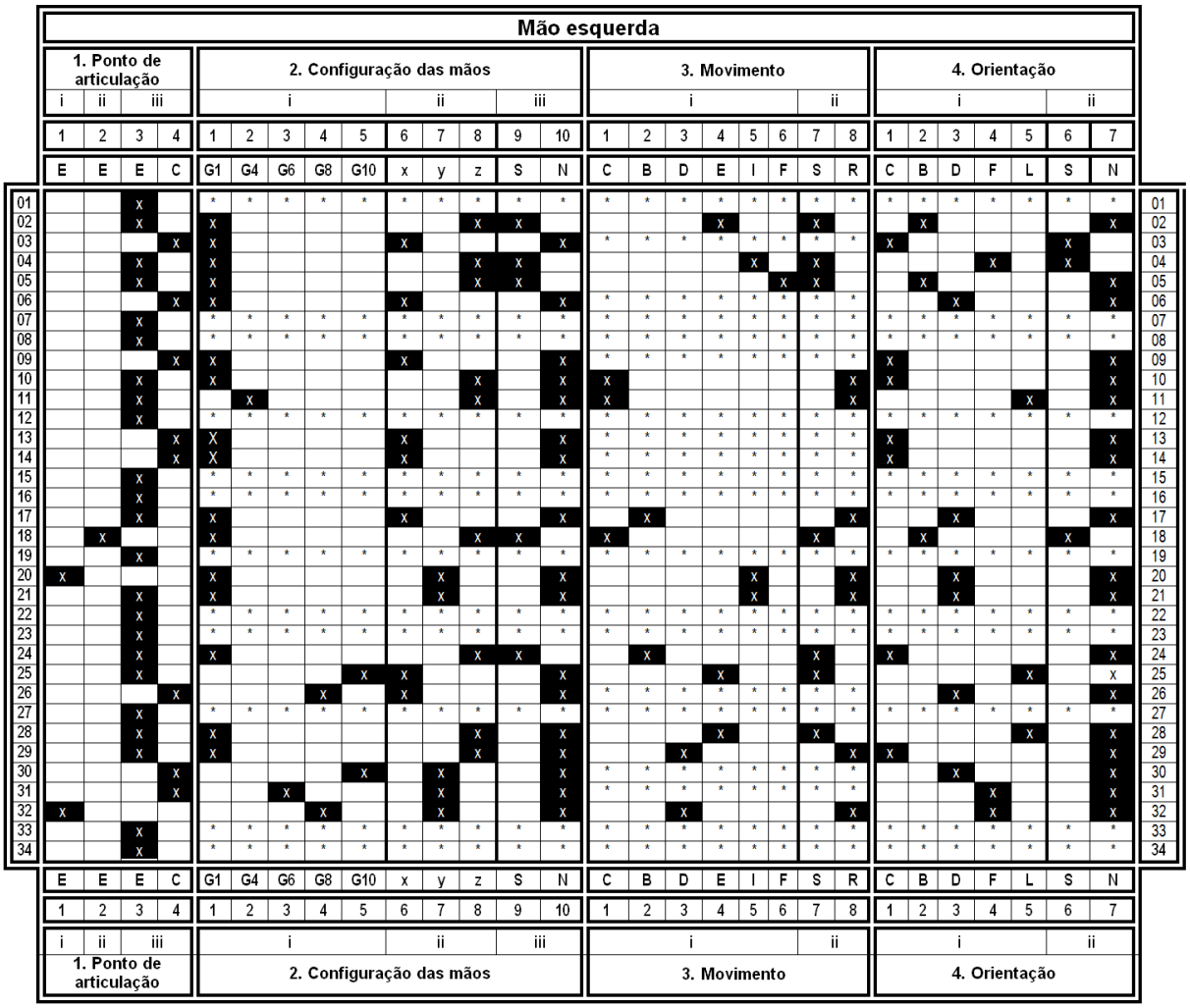


Figura 10: Valores dos atributos dos parâmetros fonológicos resumidos para a mão esquerda dos 34 sinais selecionados. (1) *Ponto de articulação*: (i) 1. À esquerda da cabeça (E). (ii) 2. À esquerda do ombro (E). (iii) 3. À esquerda do tronco (E). 4. Ao centro do tronco (C). / (2) *Configuração da mão*: (i) 1. G1. 2. G4. 3. G6. 4. G8. 5. G10. (ii) 6. Alinhamento no eixo x (x). 7. Alinhamento no eixo y (y). 8. Alinhamento no eixo z (z). (iii) 9. Mudança na configuração da mão ou alinhamento: sim (S). 10. Mudança na configuração da mão ou alinhamento: não (N). / (3) *Movimento*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para a esquerda (E). 4. Para a direita (D). 5. Para dentro (I). 6. Para fora (F). (ii) 7. Frequência: simples (S). 8. Frequência: repetida (R). / (4) *Orientação da palma da mão*: (i) 1. Para cima (C). 2. Para baixo (B). 3. Para dentro (D). 4. Para fora (F). 5. Para o lado (L). (ii) 6. Variação: sim (S). 7. Variação: não (N).

3 Criação da base de sinais

A Língua Brasileira de Sinais possui mais de dez mil verbetes documentados em (CAPOVILLA; RAPHAEL; MAURíCIO, 2012a; CAPOVILLA; RAPHAEL; MAURíCIO, 2012b), além de estar em constante mutação e expansão como toda língua viva. Como este trabalho tem por proposta realizar o reconhecimento de parâmetros por meio de Visão Computacional, houve a necessidade de se estabelecer sinais que representassem todo este corpo linguístico. O caminho natural foi primeiramente estabelecer quais são estes sinais. Isto foi realizado a partir da definição e classificação dos sinais em morfemas[1] no Dicionário Trilíngue Ilustrado de Capovilla, Raphael e Maurício (2012a), Capovilla, Raphael e Maurício (2012b). Após a escolha de quais sinais reconhecer, foi necessário construir uma base de sinais que os contenham, uma vez que não há, hoje, uma base destes sinais gravada com as ferramentas propostas neste trabalho. Assim, são descritos neste capítulo os *frameworks* que operam *software* e *hardware* permitindo a gravação em vídeo dos sinais selecionados. As especificações utilizadas para a criação desta base de sinais ora utilizada aqui é descrita ao final do capítulo.

3.1 Introdução

Os trabalhos em Visão Computacional apresentam uma variada gama de recursos que são combinados entre si de acordo com a aplicação que se pretende implementar. Para o caso específico do uso de Visão Computacional com o objetivo de reconhecimento de padrões, a seguinte divisão em tarefas pode ser descrita:

1. Aquisição da imagem
2. Segmentação da região/objeto de interesse
3. Extração das características sobre a região de interesse
4. Seleção de características
5. Classificação das imagens de acordo com as características extraídas

O primeiro item, a aquisição da imagem na entrada, é o tema tratado neste capítulo. Especificamente para este trabalho, a aquisição é a gravação dos vídeos

[1] Um morfema é a menor unidade sublexical que codifica significado.

dos sinais em Libras. Esta gravação envolve escolhas de ferramentas de *hardware* e *software* que são utilizadas tanto para pesquisas acadêmicas quanto no mercado. Há soluções que são propostas como *frameworks*, integrando diversas bibliotecas de comandos, linguagens de programação e sensores RGB-D. Nas seções a seguir, as soluções mais utilizadas são descritas e, posteriormente, as escolhas ferramentais envolvidas neste trabalho são citadas.

3.2 Ferramentas computacionais

Há uma variada gama de possibilidades para implementação de trabalhos em Visão Computacional. A decisão sobre qual solução utilizar depende do que cada uma delas oferece em termos de praticidade e interação com as demais ferramentas. Algumas possibilidades são apresentadas a seguir, dentre *frameworks*, bibliotecas, linguagens de programação, *hardwares* e *drivers*. Também é possível estabelecer conexões entre estas ferramentas para implementar as soluções desejadas.

3.2.1 Sensores RGB-D

Os sensores RGB-D são câmeras especiais com a capacidade de realizar duas gravações distintas. A primeira câmera, em geral, grava imagens de intensidade em formato RGB (Red, Green, Yellow). A segunda câmera é, na verdade, constituída de um emissor e receptor de infravermelho que permite a gravação de objetos em três dimensões (Depth). O emissor gera um sinal em determinado padrão e, de acordo com o sinal detectado pelo receptor, mapeiam-se os objetos existentes no ambiente. Há sensores RGB-D que permitem também a gravação de informações de áudio por meio de microfones embutidos, além de permitir o ajuste das câmeras de forma automática a partir de sinais de acelerômetros e motores instalados no equipamento.

A área de Interfaces Naturais ao Usuário (NUI - Natural User Interface) desenvolve soluções que utilizam estes sensores devido à facilidade de instalação e operação que eles oferecem quando utilizados em soluções de interface homem máquina. Dois sensores RGB-D de baixo custo têm sido utilizados para estas pesquisas de interfaces naturais dentro da Visão Computacional: (i) o sensor Kinect e (ii) o sensor Primesense Carmine. Há uma terceira família de sensores RGB-D, baseados no PrimeSense Carmine e fabricados pela ASUS [2], o ASUS Xtion Live e o ASUS Xtion. Ele são bastante similares ao PrimeSense e, portanto, serão citados apenas na comparação de especificações entre o Kinect e o PrimeSense Carmine.

[2] http://www.asus.com/Multimedia/Xtion_PRO/

3.2.1.1 O sensor Kinect

O sensor Kinect é um acessório do *videogame* XBOX produzido pela empresa Microsoft [3]. Na Figura 11a pode-se vê-lo com seus recursos de gravação de vídeo e áudio identificados. Ele foi lançado em novembro de 2010 como solução de captura de movimentos sem que o usuário precisasse utilizar fios ou outros acessórios para conectar-se ao *videogame*. A partir do ano de 2013, o sensor foi aprimorado e a Figura 11b apresenta esta nova versão [4], que tem como um dos pontos fortes a melhora na resolução das imagens gravadas.

Sensores RGB-D



(a) Kinect lançado em 2010.



(b) Kinect lançado em 2013.



(c) PrimeSense Carmine



(d) ASUS Xtion

Tabela 11: (a) Sensor Kinect produzido entre os anos de 2010 e 2014 e comercializado como acessório de movimento do videogame XBOX, da Microsoft. (b) Sensor Kinect produzido a partir de 2013 e atualmente comercializado como acessório de movimento do videogame XBOX, da Microsoft. (c) Sensor PrimeSense Carmine desenvolvido pela empresa de mesmo nome e vendido para a Apple Incorporation em novembro de 2013. (d) Sensor ASUS Xtion produzido pela empresa ASUS.

A integração entre *hardware* e *software* do Kinect o permitem realizar a detecção e gravação de até 20 pontos específicos do corpo humano. Na Figura 11 estão descritos todos os 20 pontos detectados no corpo humano por um conjunto *software* + Kinect.

Na verdade, quando operado por um *software*, o Kinect grava três informações: (i) o vídeo de intensidade de cores em RGB, (ii) o vídeo com imagens de profundidade

[3] http://www.microsoft.com/en-us/kinectforwindows/

[4] http://spectrum.ieee.org/automaton/robotics/robotics-hardware/heres-microsofts-new-kinect-sensor

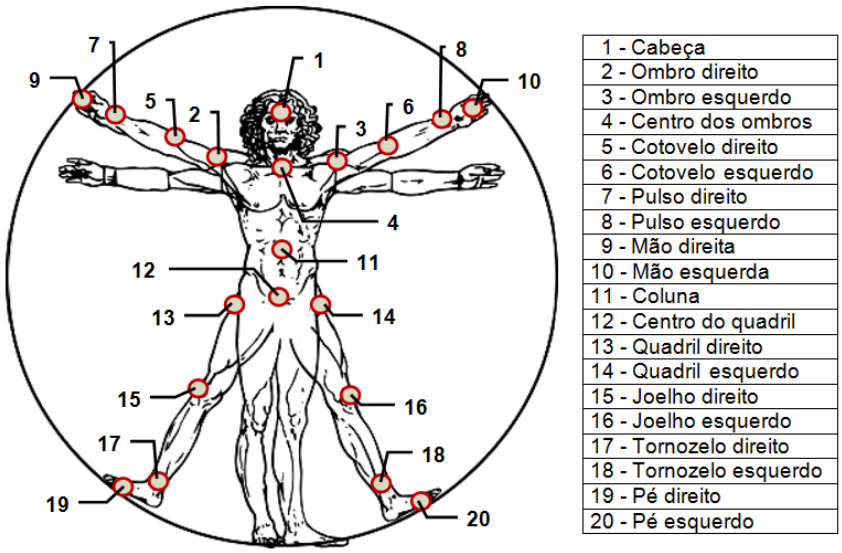


Figura 11: Pontos do corpo humano acessados pelos *softwares* que operam o Kinect.

e (iii) os valores das distâncias (x,y,z) de cada um dos 20 pontos que detecta. Estas especificações, propriedades e recursos do Kinect podem ser encontradas em (CRUZ; LUCIO; VELHO, 2012), em (ZHANG, 2012), em (MICROSOFT, 2013) e em (MANKOFF; RUSSO, 2013).

3.2.1.2 O sensor PrimeSense Carmine

A empresa PrimeSense desenvolveu um sensor RGB-D que pode ser visto na Figura 11c. A PrimeSense é uma empresa israelense especializada em sensores 3D, que foi adquirida pela *Apple Incorporation*[5] por cerca de 350 milhões de dólares em novembro de 2013.

Diversas soluções são oferecidas no desenvolvimento de aplicações de interfaces naturais utilizando o PrimeSense, que é consituído por um SoC (system on chip), uma câmera de intensidade, os sensores infravermelhos e um *middleware*[6] denominado *NiTE*.

De fato, tanto o Kinect quanto outros sensores RGB-D foram criados a partir do sensor PrimeSense, que implementou primeiramente a tecnologia de infravermelho com o objetivo de gravar a profundidade do ambiente. É esta profundidade que fornece as imagens em 3D tanto de objetos quanto de ambientes.

[5] http://appleinsider.com/articles/14/07/11/apples-secret-plans-for-primesense-3d-tech-hinted-at-by-new-itseez3d-ipad-app

[6] *Middleware* ou mediador, no campo da computação distribuída, é um programa de computador que faz a mediação entre um *software* e as demais aplicações de um sistema. É constituído por módulos com APIs (Application Program Interface) de alto nível que proporcionam a sua integração com aplicações desenvolvidas em diversas linguagens de programação e interfaces de baixo nível que permitem a sua independência relativamente ao dispositivo.

3.2.1.3 Sensor Kinect *versus* Xtion/PrimeSense Carmine

Na Tabela 12 a seguir são apresentados os pontos pró e contra de três famílias de sensores: o Microsoft Kinect, o PrimeSense Carmine e o ASUS Xtion. O ASUS Xtion é bastante similar ao PrimeSense Carmine, e ambos possuem características similares. Portanto, estão sendo contrapostos em conjunto ao Microsoft Kinect.

Microsoft Kinect	**ASUS Xtion / Primesense Carmine**
1. Alta qualidade dos *drivers*.	1. Baixa qualidade dos *drivers*.
2. Estabilidade na operação com diversos hardwares.	2.Trabalha somente com a USB 3.0 como controladora.
3. Possui um motor que pode ser controlado remotamente pela aplicação, que o posiciona de forma conveniente à esta.	3. Não possui motor para ajuste do posicionamento das câmeras.
4. Tamanho (12” x 3” x 2.5”) .	4. Não necessita de fonte de alimentação externa. Utiliza a USB.
5. Peso (3.0 lb).	5. Tamanho (7"x 2"x 1.5").
6. Precisa de fonte externa AC/DC.	6. Peso (0.5 lb).
7. Alta interferência com outro Kinect.	7. Baixa interferência com outros ASUS Xtion ou PrimeSense Carmine.
8. Baixa qualidade de imagem RGB se comparado ao novo Kinect.	8. Apresenta boa resolução de imagem RGB.

Tabela 12: Comparação entre o sensor Kinect, da Microsoft e os sensores PrimeSense Carmine / ASUS Xtion.

3.2.2 *Frameworks*

Há duas importantes soluções propostas como *frameworks* em Visão Computacional que trabalham diretamente na implementação de reconhecimento de padrões em vídeos. Nas seções a seguir, ambos os *frameworks* são descritos: primeiramente o OpenNI e, em seguida, o Microsoft SDK for Kinect.

3.2.2.1 OpenNI

OpenNI ou Open Natural Interaction é uma organização sem fins lucrativos liderada pela indústria que tem por objetivo melhorar e certificar a interoperabilidade de interfaces naturais (NUI - Natural User Interaction) e orgânicas (OUI - Organic User Interaction) para dispositivos e aplicações em interação homem - computador, conforme descrito em (OPENNI, 2014).

O *framework* disponibilizado pelo OpenNI (também conhecido por OpenNI SDK - Software Development Kit) fornece um conjunto de APIs (Application Program Interface), escritas em código aberto e com licença de utilização livre. Estas APIs fornecem suporte para a implementação de sistemas em reconhecimento de comandos de voz, reconhecimento de gestos e rastreamento de movimento do corpo. Ela permite a utilização das linguagens de programação C e C++.

A pretensão do OpenNI é que suas APIs possam tornar-se um padrão para aplicações que acessam dispositivos de interação natural. Um passo para alcançar esta pretensão é permitir a operação tanto do PrimeSense Carmine quanto do Kinect.

3.2.2.2 Microsoft Kinect for Windows SDK

Ao contrário da OpenNI, que trabalha com código aberto e licença livre, a Microsoft, empresa fabricante do sensor RGB-D Kinect, opera com o chamado *software* proprietário. No caso do Kinect, entretanto, os APIs que o operam foram liberados para a comunidade em geral em forma de SDKs (Software Development Kit) que são frequentemente atualizados por sua equipe. Estes SDKs permitem o uso das linguagens de programação C++, C# e Visual Basic para criar aplicações e experiências que auxiliem o reconhecimento de voz e de gestos. Em seu SDK para Kinect, a Microsoft inclui exemplos de aplicações com o código fonte completamente aberto e, de acordo com a própria fabricante, esta ação facilita o desenvolvimento de aplicações as mais diversas.

Este SDK, entretanto, opera somente sobre a plataforma Windows, para seus sistemas operacionais Windows 7 e Windows 8.

3.2.3 Bibliotecas de comandos em Visão Computacional

Em Visão Computacional há diversos procedimentos e técnicas aplicadas às imagens e vídeos que são comuns aos sistemas independente da aplicação. Duas importantes bibliotecas que implementam estas técnicas estão disponíveis para o desenvolvimento de sistemas em Visão Computacional: o OpenCV e o pacote de funções em Visão Computacional do Matlab. Ambos tem como objetivo facilitar a aplicação destas técnicas operando sobre imagens e vídeos. Um projeto que apresenta uma biblioteca de comandos específicos para o Kinect foi criado e também será descrito a seguir, o OpenKinect.

3.2.3.1 OpenCV

A biblioteca conhecida como OpenCV (Open Source Computer Vision Library) é uma biblioteca com o código fonte aberto. Ela é constituída de *software* de Aprendizagem de Máquina e Visão Computacional. Seu objetivo é fornecer uma infraestrutura comum para aplicações em Visão Computacional e aumentar a utilização de percepção de máquina em produtos comerciais. Ela é um produto com licença BSD (Berkeley Source Distribution) e isto faz com que o mercado tenha facilidade em utilizar e modificar o código disponibilizado.

De acordo com a documentação oficial do OpenCV, em OpenCV (2014), a biblioteca OpenCV é constituída por mais de 2.500 algoritmos otimizados, que incluem um conjunto tanto de soluções clássicas quanto soluções que são o estado da arte em Visão Computacional e Aprendizagem de Máquina.

A biblioteca oferece interfaces para operação em plataforma Windows, Linux, Android e Mac OS. Estas interfaces podem ser construídas em C++, C, Python, Java e Matlab.

3.2.3.2 Matlab

Matlab é uma linguagem de alto nível fabricada pela MathWorks, MathWorks (2012). Ele tem por proposta disponibilizas um ambiente interativo que executa tarefas computacionais de forma mais rápida que linguagens de programação como C, C++ e Fortran, segundo sua fabricante. O Matlab é bastante conhecido e utilizado especialmente nas diversas áreas da Engenharia por ser um *software* com recursos matemáticos e implementar técnicas em diversas áreas de conhecimento, por meio de seus *toolboxes*[7]. Além do mais, permite simulações de grande parte das técnicas dos *toolboxes* que disponibiliza, utilizando uma interface amigável, o ambiente *Simulink*. A área de Visão Computacional é contemplada por um *toolbox* específico de mesmo nome.

Tanto o sensor Kinect quanto o PrimeSense Carmine podem ser manipulados pelo Matlab, por meio de *drivers* próprios para este fim.

3.2.3.3 OpenKinect

O OpenKinect é uma comunidade aberta que se propõe a trabalhar especificamente com o sensor Kinect, produzindo *softwares* a partir de uma biblioteca de código aberto desenvolvida para uso do Kinect em plataformas Windows, Linux e Mac. Inicialmente criada para compartilhar soluções de operação no Kinect, atualmente perdeu força frente a outras soluções existentes, duas delas citadas nas seções anteriores: o OpenNI e o Microsoft Kinect for Windows SDK.

3.2.4 Soluções utilizadas

A escolha das ferramentas utilizadas neste trabalho tem por base os estudos das soluções diversas apresentadas para implementação de sistemas em Visão Computacional utilizando sensores RGB-D. Entre os sensores RGB-D, optou-se pelo Kinect, por seu baixo custo e facilidade de integração. Para operar o sensor, foi esco-

[7] *Toolboxes* são um conjunto de arquivos com a extensão ".m" criados a partir de funções intrínsecas. Há *toolboxes* disponíveis para diversas áreas do conhecimento no Matlab.

lhido um *software* desenvolvido para plataforma Windows 7, o *nuiCaptureAnalyze*[8]. Ele foi desenvolvido por meio do SDK para Kinect disponibilizado pela Microsoft e sua escolha foi estabelecida pela forma como grava as informações obtidas do Kinect. Enquanto a gravação de cada sinal é realizada, ele grava simultaneamente (i) o vídeo de intensidade RGB, (ii) o vídeo de profundidade, (iii) o vídeo do esqueleto e (iv) dados de posição dos 20 pontos definidos na Figura 11 em formato Matlab.

3.2.5 Resumo dos recursos

A tabela a seguir resume as ferramentas existentes em Visão Computacional e que combinam *frameworks* que operam sobre sensores RGB-D para aplicações diversas da área.

Framework	Biblioteca	Linguagens de programação	Plataforma
OpenNI	OpenCV, Matlab	C++, C#, Python, Java	Windows, Linux, Mac
OpenKinect	OpenCV, Matlab	C++, C#, Phyton, Java, ActionScript	Windows, Linux, Mac, Android
Microsoft SDK for Kinect	——	C++, C# Visual Basic	Windows 7 e 8 (Phone)

Tabela 13: Resumo das soluções apresentadas

3.3 A base dos sinais

Encontrar bases de sinais para uso da comunidade científica em Visão Computacional, especificamente para o reconhecimento de gestos e mais especificamente de sinais é, ainda, um desafio. Como as aplicações e ferramentas disponíveis são bastante diversas, torna-se realmente um desafio construir uma base de vídeos que possuam características padronizadas de luminosidade, fundo, distância, cor de pele, resolução da imagem, posição de câmeras de captura de movimentos, dentre outras e que, desta forma, possam ser utilizadas para treinamento em diversas situações. Chunli, Wen e Jiyong (2002), Dias, Souza e Pistori (2006) e Cooper e Bowden (2007) apresentam estudos sobre as especificidades na criação de bases de gestos e produzem algumas bases em língua de sinais. Escalera et al. (2013) visualizaram estas dificuldades e vêm, desde 2011, promovendo ações de coleta e padronização de grande quantidade de dados, de forma interativa e *online*, para gestos manuais em geral. Eles utilizam como câmera o sensor RGB-D Kinect e o *software* de simulação matemática Matlab para operação e gravação dos vídeos. Os desafios destas ações vão desde a documentação dos sinais em dicionários de acordo com a comunidade

[8]http://nuicapture.com/

que se pretende gravar até a padronização das ferramentas utilizadas para a gravação dos sinais. Questões acerca da representatividade dos sinais na língua em que se insere a comunidade são as mais diversas possíveis e algumas delas são discutidas em (SCHEMER, 2003), em (JOHNSTON, 2003) e em (CLEVE, 2003).

Diante destas dificuldades em encontrar uma base de sinais brasileira e que fosse constituída com os sinais que analisamos e com as ferramentas que dispomos, decidiu-se por construir uma base de sinais própria, com o cuidado de especificá-la de tal forma que adicionar outros sinais e outros sinalizadores a ela seja possível seguindo estas especificações. Portanto, a base aqui utilizada é um protótipo que vem sendo melhorado ao longo dos anos de desenvolvimento desta pesquisa.

Decidiu-se pela utilização de sinais que representem os 34 morfemas moleculares propostos nos trabalhos apresentados em (CAPOVILLA; RAPHAEL; MAURíCIO, 2012a; CAPOVILLA; RAPHAEL; MAURíCIO, 2012b). Um único sinal foi escolhido de cada um dos agrupamentos de morfemas moleculares. Temos, desta forma, os 34 sinais descritos na Figura 8, cada um representante de agrupamentos de morfemas em Língua Brasileira de Sinais e representando diversos outros sinais devido à sua similaridade apresentada em (CAPOVILLA; RAPHAEL; MAURíCIO, 2012a; CAPOVILLA; RAPHAEL; MAURíCIO, 2012b). Como não há, até o presente momento, uma base de dados com estes sinais, a opção pela criação nossa própria base de sinais mostrou-se adequada. Cada um dos 34 sinais foi gravado cinco vezes por um único sinalizador. Ao se estabelecer um único sinalizador, o estudo concentra-se em reconhecer as diferenças entre os sinais e não entre a forma como cada sinalizador os realiza.

Os 34 sinais foram gravados utilizando como câmera o sensor RGB-D Kinect em conjunto com o software *NuiCaptureAnalyze* e três importantes requisitos foram estabelecidos para a criação da base de dados. São eles:

1. cada vídeo deve conter o espaço do sinal, que é realizado, necessariamente, acima dos quadris. Assim, posições relativas às pernas não foram gravadas, uma vez que estas não compõem nossa região de interesse;

2. a distância do sinalizador ao sensor deve ser suficiente para capturar os movimentos dos braços;

3. a gravação dos vídeos de intensidade, vídeos de profundidade, vídeos do esqueleto e das informações de posição de partes do corpo deve acontecer de forma simultânea.

A distância utilizada entre o sinalizador e o sensor Kinect, baseando-se nestas diretrizes, foi de 1,9 metros (6,2 pés). Este valor está em acordo com a faixa de dis-

tância de gravação do sensor, que está entre 0,8 metros (2,6 pés) e 3,5 metros (11 pés).

Utilizando estas ferramentas e obedecendo a estas especificações, foram gravados cinco vídeos para cada amostra de cada um dos 34 sinais. Um arquivo em formato Matlab, MathWorks (2012), contendo os valores em milímetros, das distâncias (x,y,z) de cada ponto detectado do corpo ao sensor, para cada um dos vídeos gravados, é gerado.

A Figura 12 apresenta um esquemático da gravação do sinal “Comemorar” utilizando-se estas ferramentas. Em (a), vê-se um quadro colorido no formato RGB do vídeo gravado; em (b) um quadro de profundidade do vídeo de profundidade; em (c) um quadro do esqueleto do vídeo do esqueleto. A taxa de gravação dos vídeos é de 30 quadros por segundo e tanto o vídeo em RGB, o vídeo da profundidade e o vídeo do esqueleto foram gravados em formato AVI (formato de áudio e vídeo da Microsoft).

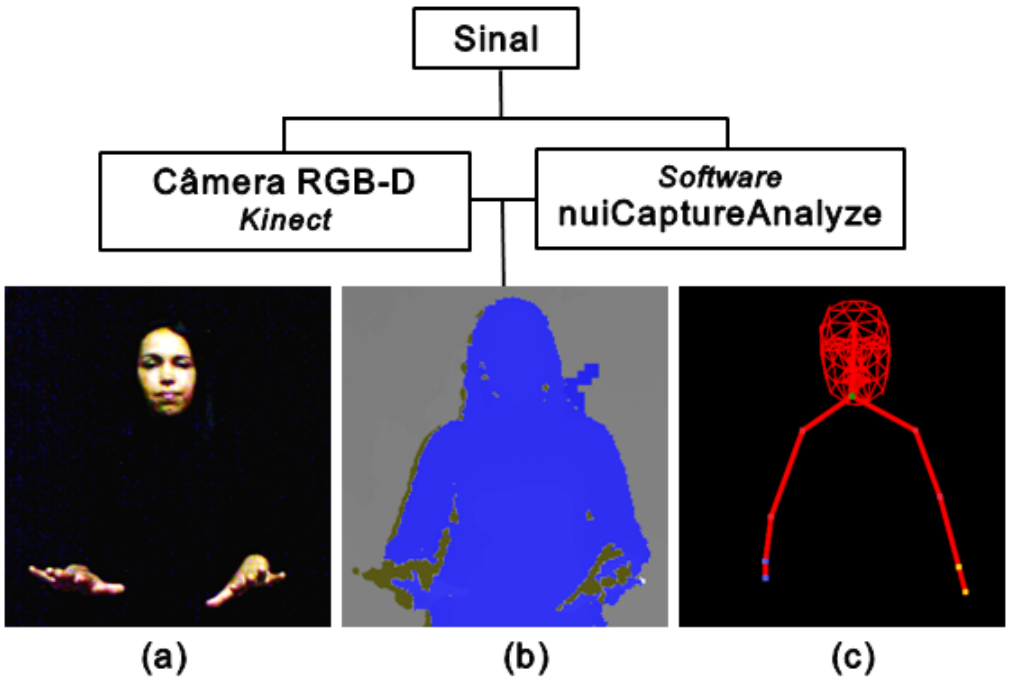


Figura 12: Formatos dos vídeos gravados para o sinal “Comemorar” utilizando o *software nuiCaptureAnalyze* operando o sensor Kinect: (a) Intensidade, (b) Profundidade e (c) Posição.

3.4 Considerações finais

Diante da grande quantidade de ferramentas integradas para criação de sistemas em Visão Computacional, diversas escolhas foram realizadas na criação de uma base de dados com sinais em Libras que fosse capaz de representar um conjunto de

sinais da língua. A questão da comercialização e baixo custo do sensor Kinect fez com que ele fosse utilizado, em conjunto com o *software nuiCaptureAnalyze* para a gravação dos sinais escolhidos.

Combinações diferentes de ferramentas levam à necessidade de construção de outras bases de sinais, pois a parametrização pode ser diferente, alterando brilho, distância ao sensor e formato de dados disponibilizado para utilização.

Assim, a construção de uma base de sinais como realizada neste trabalho fornece diretrizes gerais para futuras bases gravadas com sensores RGB-D mais modernos ou com características diferentes, mantendo-se as especificações aqui descritas.

4 Sumarização de vídeos

A sumarização é um problema recorrente em processamento de vídeos, uma vez que estes são, em geral, arquivos que contém grande quantidade de informação digital. A principal função da sumarização consiste em reduzir o tamanho dos vídeos com o mínimo de perda de informação contida nestes. Após uma breve discussão sobre técnicas aplicadas a redução de vídeos com diferentes finalidades, a abordagem realizada neste trabalho para reduzir as informações redundantes nos vídeos gravados para os sinais é detalhada e discutida. Ao final do capítulo, a detecção da região de interesse para este trabalho realizada a partir dos vídeos já sumarizados, tem seu algoritmo descrito.

4.1 Introdução

A sumarização de vídeo pode ser definida como o problema de encontrar-se uma sequência de quadros em um vídeo (com ou sem informação de áudio) que represente, da melhor forma possível e para a aplicação em questão, o conteúdo deste vídeo de forma concisa. A partir do vídeo reduzido deve ser possível procurar e recuperar informações tanto para vídeos com grandes sequências quanto para vídeos com sequências menores.

Dadas as diversas aplicações que se beneficiam diretamente da compactação resultante de uma sumarização, existe uma grande dificuldade em se padronizar as técnicas utilizadas, pois vídeos de diferentes gêneros (documentários, esportivos, filmes) solicitam diferentes técnicas que eliminem redundâncias em seus quadros. Isto deve-se ao fato de apresentarem dinâmicas bastante próprias. Portanto, a padronização nem sempre é possível, ou mesmo recomendável, uma vez que as técnicas de sumarização devem respeitar as características do vídeo e da forma como este é manipulado.

Esta diversidade de técnicas aplicadas em redução de vídeos aparece na grande variedade de pesquisas existentes. O método proposto em (MENDI; CLEMENTE; BAYRAK, 2013) baseia-se na estimação de uma métrica calculada por dois algoritmos de fluxo óptico distintos e é implementado para vídeos esportivos. Os métodos de Decomposição do Valor Singular (Single Value Decomposition), Simplificação das Curvas (Curve Simplification) e Modelo de Grafos (Graph Model) realizam sumarização conforme os trabalhos disponíveis em (GONG; LIU, 2000), em (DEMENTHON; KOBLA;

DOERMANN, 1998) e em (NGO; MA; ZHANG, 2005), respectivamente.

Grande parte das técnicas de sumarização exige algum nível de adaptação de parâmetros. Isto pode dar-se de forma automática, como em (REN; JIANG; FENG, 2010) e em (EJAZ; TARIQ; BAIK, 2012) ou de forma interativa com usuários, como em (ELLOUZE; BOUJEMAA; ALIMI, 2010). A avaliação visual como métrica para validação de métodos também é bastante utilizada nestes casos. Nos trabalhos de Gianluigi e Raimondo (2006) e Ma et al. (2002), Ma et al. (2005) a avaliação visual é realizada tanto diretamente, como no primeiro caso, quanto por meio do desenvolvimento de um chamado Modelo de Atenção ao Usuário, nos trabalhos seguintes.

A abordagem da sumarização como um problema clássico de otimização Min-Max pelos autores em (LI; SCHUSTER; KATSAGGELOS, 2005) e em (DOULAMIS; DOULAMIS; KOLLIAS, 2000) e de Algoritmos Evolucionários em (ELLOUZE; BOUJEMAA; ALIMI, 2010) mostram a possibilidade da utilização tanto de técnicas típicas de otimização quanto de técnicas inovadoras na solução deste problema.

Nesta tese nosso objeto de trabalho são os sinais, cada um deles gravado em vídeos. A necessidade da redução destes vídeos colocou-se inicialmente para a eliminação da redundância entre quadros e consequente aumento das distinções entre os vídeos de acordo com as características extraídas. Como consequência da redução de quadros nos vídeos dos sinais, a aplicação das técnicas para extrair as características de cada sinal tornou-se mais eficiente também em termos de tempo computacional de processamento.

4.2 Sumarização de vídeo

O problema de sumarização de vídeos é abordado aqui como um problema clássico de otimização conhecido como Problema da Diversidade Máxima (PDM). A descrição do Problema da Diversidade Máxima está presente, de forma resumida, no trabalho apresentado em Kuo, Glover e Dhir (1993). Ele consiste em encontrar um subconjunto $M(|M| = m)$ de um conjunto $N(|N| = n)$ de elementos tais que a diversidade dos *m* elementos seja maximizada.

No Problema da Diversidade Máxima calcula-se uma matriz chamada de diversidade que contém uma métrica de cálculo de diferenças que esteja de acordo com a aplicação em questão. Assim, diversas relações para o cálculo da diversidade podem ser utilizadas para definir os valores de diversidade d_{ij} (valor de diversidade entre o elemento i e j. O equacionamento do PDM é descrito na Equação 4.1, onde $x_i = 1$ significa que o elemento i está em um subconjunto de M e m é o parâmetro que define

a quantidade de quadros que deve ser selecionada entre os n quadros possíveis.

$$\max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} x_i x_j, \text{ sujeito a } \sum_{i=1}^{n} x_i = m \tag{4.1}$$

$$\text{onde } x_i \in \{0, 1\} \ \forall \ i = 1, \ldots, n$$

No presente caso calcula-se a diversidade entre o quadro i e o quadro j do vídeo de cada sinal. A métrica utilizada para o cálculo dos valores d_{ij} da matriz de diversidade $[\mathbf{D}]$ considera dois aspectos importantes nesta abordagem para vídeos: a distância temporal e a diferença de cores entre os quadros dos vídeos.

A distância temporal consiste dos valores de tempo em que cada quadro ocorre no vídeo. Assim, quadros muito próximos terão distâncias no tempo também próximas. Já a diferença de cores é calculada como a somatória do valor absoluto da diferença entre os quadros para os espaço de cores RGB. Quanto maior este valor, significa que há mais distância em termos de cores.

O objetivo aqui é, então, maximizar os valores de diversidade tanto para a métrica temporal quanto para a métrica de diferença de cores entre os quadros. Assim, a matriz de diversidade $[\mathbf{D}]$ implementada é dada por:

$$[\mathbf{D}] = [\mathbf{C}] + [\mathbf{T}] \tag{4.2}$$

onde
$\mathbf{D}$ é a matriz de diversidade, definida como

$$\begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{m1} & \cdots & d_{mm} \end{bmatrix}$$

$\mathbf{C}$ é a matriz diversidade para a diferença de cores e $\mathbf{T}$ é a matriz de diversidade para a distância temporal.

A matriz é de diversidade, desta forma, possui dimensão mxm e seus valores d_{ij} são números reais e positivos.

O Problema da Diversidade Máxima é um problema NP-difícil que pode ser solucionado por métodos heurísticos ou exatos. Neste caso, uma estratégia evolucionária foi empregada para a otimização e consequente solução do problema. Esta

estratégia, denominada MSES (Memetic Self-Adaptative Evolution Strategies) pode ser encontrada em (FREITAS et al., 2014). Nele, um algoritmo evolucionário com heurísticas de busca local foi projetado para resolver o Problema da Diversidade Máxima.

4.3 Sumarização dos sinais

O emprego da técnica de sumarização de vídeos nesta tese tem por objetivo reduzir a quantidade de quadros nos vídeos dos sinais linguísticos. Sumarizar estes vídeos permite que apenas os quadros que contêm mais informações distintas entre si sejam selecionados. Assim, a procura pela região de interesse nos vídeos e posterior extração de características é realizada somente nos quadros já selecionados.

Um número fixo de quadros representativos para cada sinal foi estabelecido neste trabalho em $m = 5$, a partir da avaliação das quantidades de quadros de cada vídeo. Há sinais linguísticos cuja duração gira em torno de 20 quadros, conforme vê-se na Tabela 16. Esta tabela apresenta também a quantidade média de quadros em cada sinal e o tempo médio de execução destes. Nota-se que o tempo médio máximo ocorre para o sinal “Justiça”, no valor de 2,91 segundos, o que corresponde a 87 quadros. Já o tempo mínimo acontece para o sinal “Esquecer”, no valor de 0,87 segundos e equivalente a 26 quadros em média. Nas Figuras 13 e 14 estão os quadros do vídeo “Justiça”. Ela está dividida em duas partes devido ao tamanho do vídeo, pois este é o sinal com média de maior duração. Já na Figura 15 são exibidos os quadros do sinal de menor duração em média, o sinal “Esquecer”. Já o fato do valor de m ser fixo acontece para que todos os sinais possam ser representados por vetores de características de mesmo tamanho.

Os parâmetros linguísticos reduzidos para os sinais “Justiça” e “Esquecer” são definidos no Capítulo 2 conforme exibido nas Figuras 9 para a mão direita e 10 para a mão esquerda. Nas Tabelas 14 e 15 estes parâmetros estão descritos.

Os parâmetros fonológicos de cada um dos sinais podem ser visualizados nos cinco quadros que são o resultado da sumarização.

Baseando-se nestes valores médios de tempo e de quantidade de quadros, nota-se também que a duração dos vídeos dos sinais acontece entre 0,87 e 2,91 segundos e o tamanho destes pode ser considerado pequeno se comparados a vídeos de outras aplicações de sumarização. Este tamanho, entretanto, não invalida a necessidade de redução de redundância nos sinais. Outras características dos vídeos dos sinais é que há maior movimento na região das mãos e não há grandes diferenças de fundo entre os quadros.

Parâmetros	**Justiça**	**Esquecer**
Ponto de articulação	À direita do tronco	Ao centro da cabeça
Configuração de mão	 (a) G4 - Forma 2 Alinhamento em z Não sofre alteração	 (b) G1 - Forma 3 Alinhamento em x Não sofre alteração
Movimento	Para cima Repetido	Para a direita Simples
Orientação da palma da mão	Para o lado Não sofre alteração	Para dentro Não sofre alteração

Tabela 14: Parâmetros fonológicos da mão direita dos sinais com maior e menor tempo de duração na média, respectivamente, sinais “Justiça” e “Esquecer”.

Parâmetros	**Justiça**	**Esquecer**
Ponto de articulação	À esquerda do tronco	À esquerda do tronco
Configuração de mão	 (a) G4 - Forma 2 Alinhamento em z Não sofre alteração	Não participa
Movimento	Para cima Repetido	Não participa
Orientação da palma da mão	Para o lado Não sofre alteração	Não participa

Tabela 15: Parâmetros fonológicos da mão esquerda dos sinais com maior e menor tempo de duração na média, respectivamente, sinais “Justiça” e “Esquecer”.

Sinais	Quantidade média de quadros	Tempo médio de execução
1. Pessoa	49,4	1,65
2. Espalhar	58,8	1,96
3. Copiar	45,4	1,51
4. Agarrar	41,8	1,39
5. Recolher	84,2	2,81
6. Sumir	55,6	1,85
7. Olhar	40,4	1,35
8. Justo	44,4	1,48
9. Verdade	53,2	1,77
10. Peso	86,0	2,87
11. Justiça	87,2	2,91
12. Quem	34,8	1,16
13. Nada	73,8	2,46
14. Acreditar	47,6	1,59
15. Esquecer	26,0	0,87
16. Amar	76,0	2,53
17. Angustiado	76,0	2,53
18. Comemorar	50,2	1,67
19. Rancor	55,0	1,83
20. Assembleia	73,0	2,43
21. Comparar	69,2	2,31
22. Gritar	37,4	1,25
23. Falar	43,4	1,45
24. Absorver	53,2	1,77
25. Engordar	41,4	1,38
26. Brigar	63,0	2,10
27. Esperto	47,6	1,59
28. Brilhar	48,8	1,63
29. Empregado	74,6	2,49
30. Substituir	59,6	1,99
31. Prisão	28,8	0,93
32. Televisão	81,8	2,83
33. Ontem	53,0	1,77
34. Futuro	45,0	1,50

Tabela 16: Quantidade média de quadros e tempo médio de execução em segundos para cada um dos 34 sinais da base de sinais utilizada. Cada sinal foi gravado cinco vezes a uma taxa de 30 quadros por segundo e por um único sinalizador.

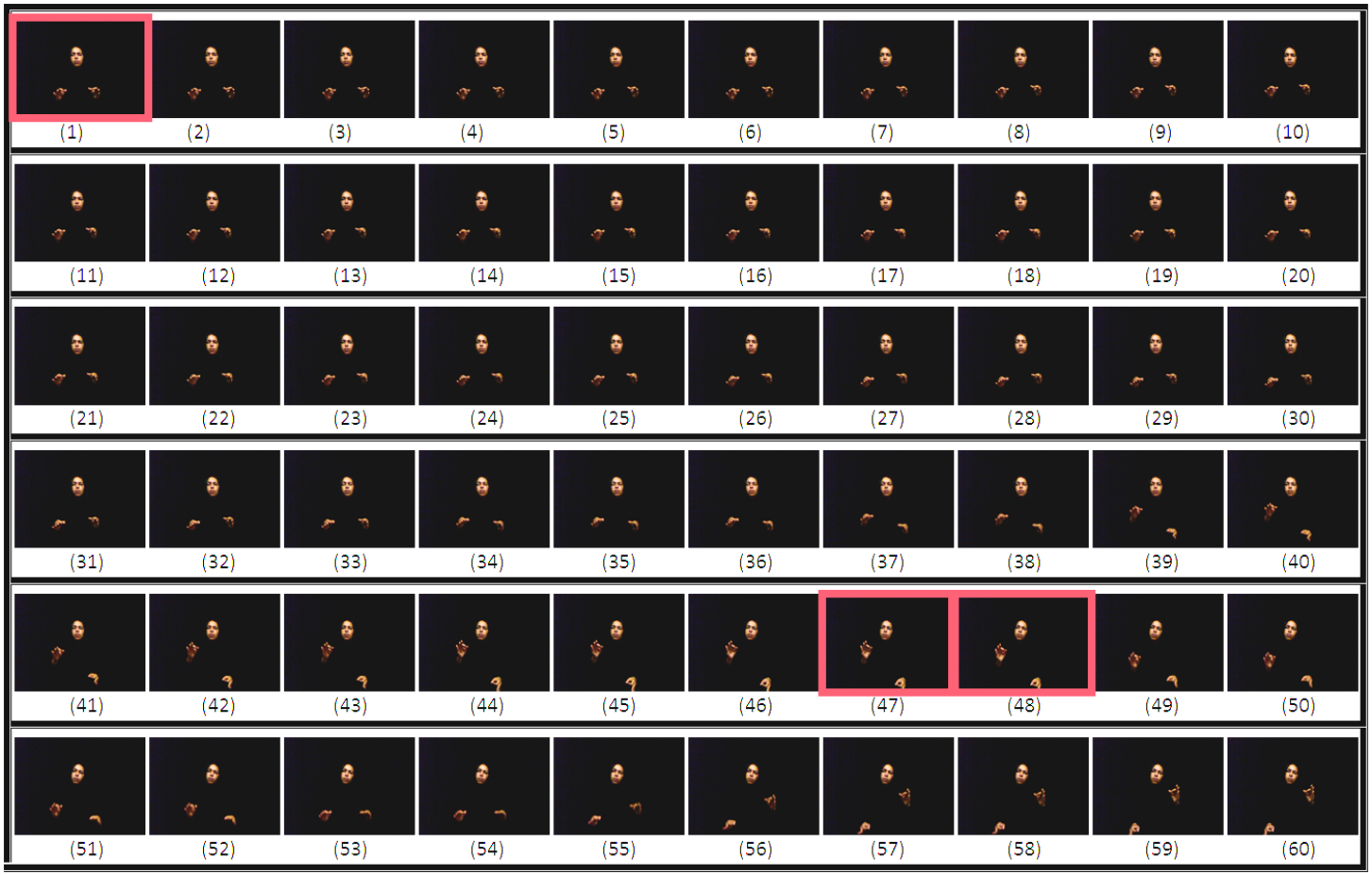


Figura 13: Primeira parte de duas do vídeo para o sinal “Justiça”: quadros 1 a 60 para a sua primeira gravação. Este sinal apresenta a maior quantidade de quadros na média entre as cinco gravações realizadas e, consequentemente, o maior tempo de execução. Os $m = 5$ quadros selecionados por meio da sumarização estão em destaque com moldura vermelha.

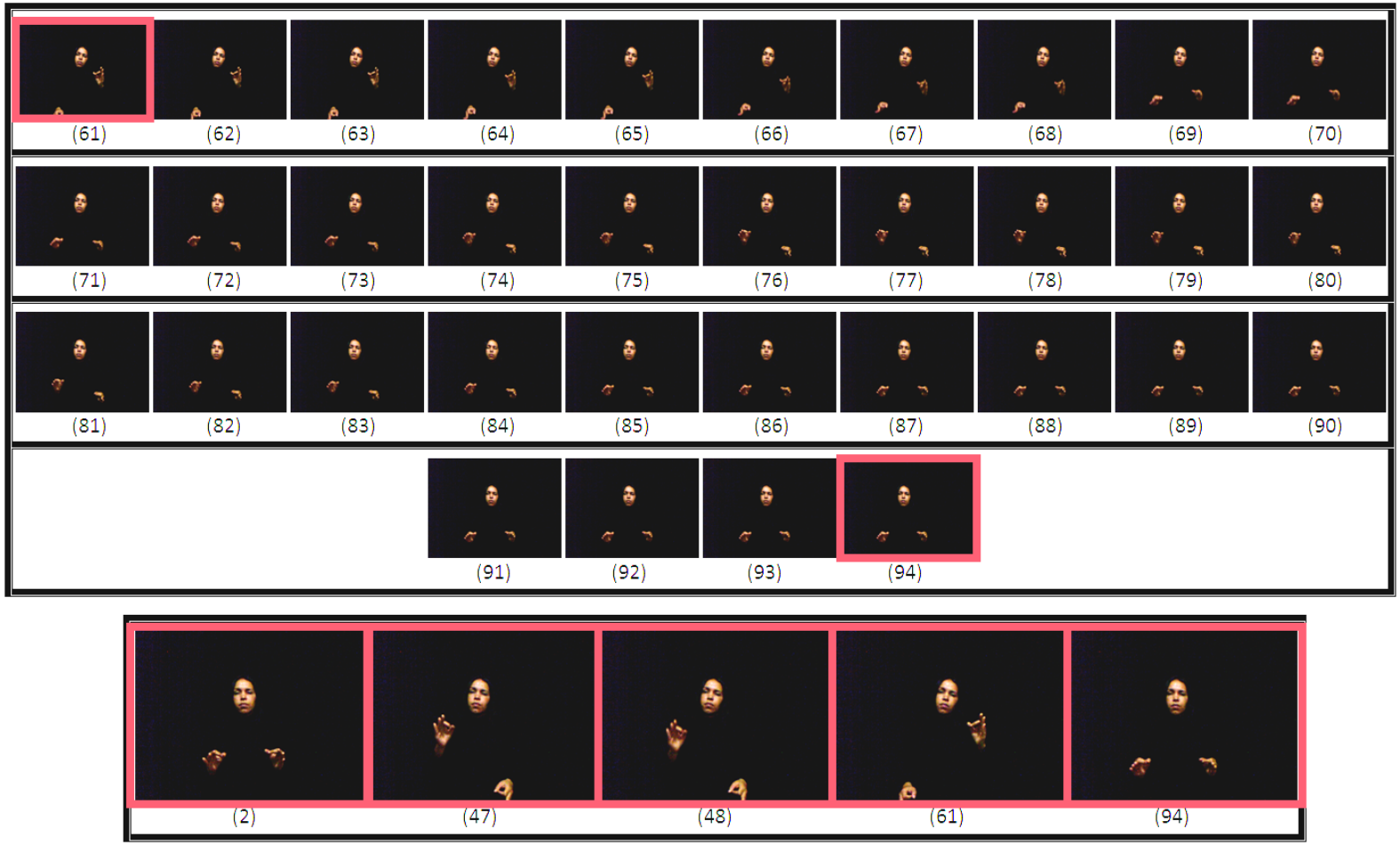


Figura 14: Segunda parte de duas do vídeo para o sinal “Justiça”: quadros 61 a 94 para a sua primeira gravação. Este sinal apresenta a maior quantidade de quadros na média entre as cinco gravações realizadas e, consequentemente, o maior tempo de execução. Os $m = 5$ quadros selecionados por meio da sumarização estão em destaque com moldura vermelha.

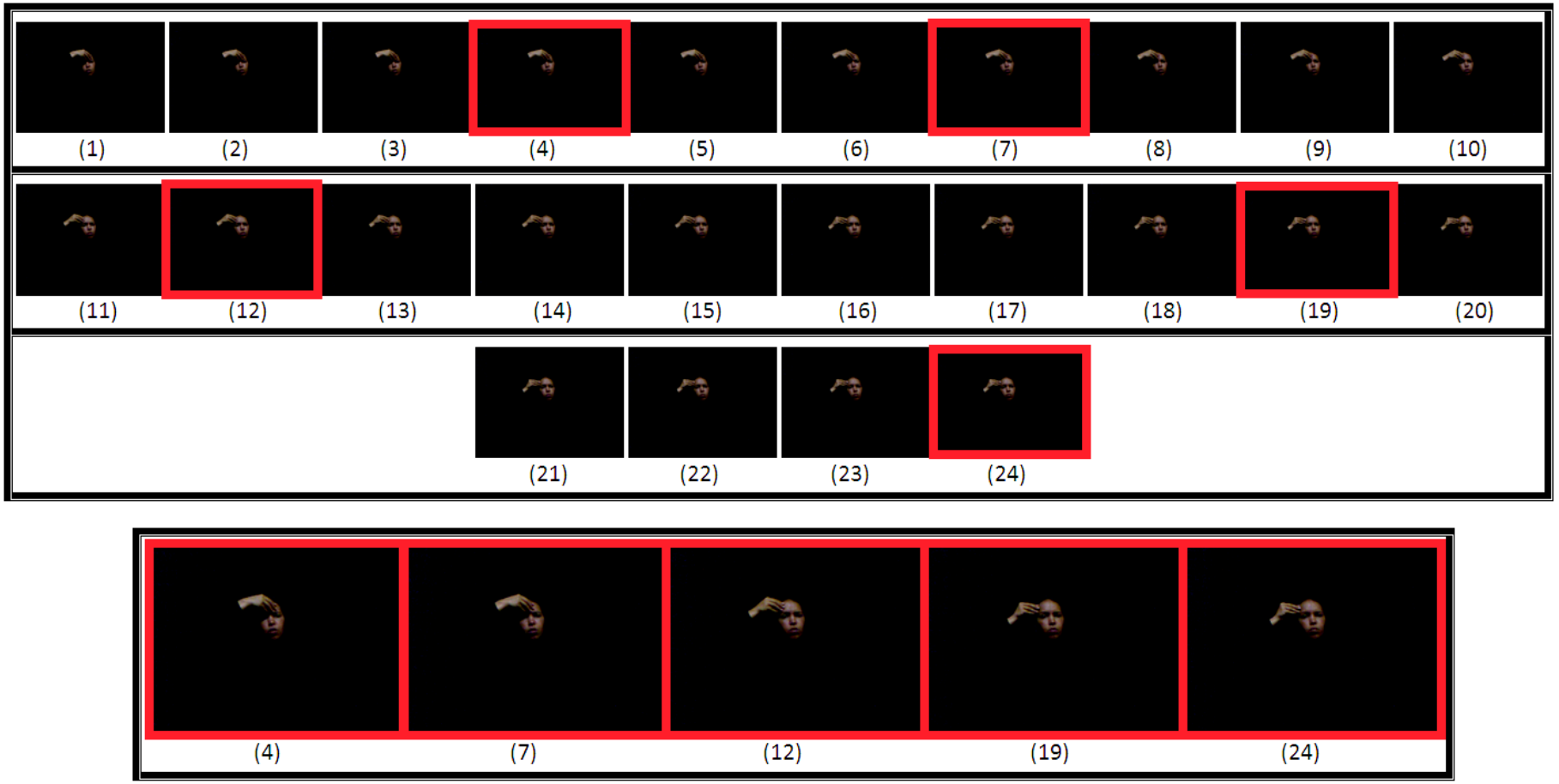


Figura 15: Quadros do sinal “Esquecer”, para a sua primeira gravação. Este sinal apresenta a menor quantidade de quadros na média entre as cinco gravações realizadas e, consequentemente, o menor tempo de execução. Os $m = 5$ quadros selecionados por meio da sumarização estão em destaque com moldura vermelha.

4.4 Detecção da região de interesse

A extração de características realizada nesta tese deve ser aplicada aos quadros dos vídeos sumarizados após a detecção da região de interesse (ROI). Há três regiões de interesse possíveis: a mão direita, a mão esquerda e o rosto. É por meio da análise da trajetória das mãos e da sua forma que os elementos fonológicos Ponto de articulação, Configuração de mão, Movimento e Orientação da palma da mão são detectados. Já para o elemento Expressões não-manuais é imprescindível a detecção do rosto.

Como a análise neste trabalho é realizada para os elementos fonológicos espaciais, a região de interesse é constituída pela região da mão direita e pela região da mão esquerda. O rosto também é detectado sempre que o sinal é realizado à sua frente, pois ele identifica a posição das mãos para alguns sinais.

Faz-se a descrição da extração desta região de interesse a partir do vídeo gravado pelo sensor Kinect operado por meio do *software nuiCaptureAnalyze* e já sumarizado.

As informações existentes para cada sinal, após a sumarização de seu vídeo RGB, são:

- 5 imagens da intensidade no padrão RGB;
- 5 imagens da profundidade;
- 5 imagens do esqueleto com os pontos das mãos, punhos, cabeça e ombros do corpo marcados.

A partir destas imagens as seguintes etapas são implementadas para que a detecção efetiva das mãos nas imagens seja possível.

1. Binarizar a imagem da profundidade: nesta imagem, o corpo humano fica em evidência em relação ao fundo gravado. O corpo fica gravado com as cores magenta, vermelho, verde e azul. O fundo é gravado em cinza, conforme Figura 16 (b). Assim, a binarização é realizada de tal forma que os *pixels* das cores citadas sejam representados pelo valor “1” e os *pixels* da cor cinza são representados pelo valor “0”.
2. Retirar o fundo da imagem: faz-se a operação de multiplicação entre a imagem RGB convertida para escala de cinza e a imagem binarizada da profundidade.
3. A partir do vídeo do esqueleto, representado na Figura 16 (c), são calculadas as coordenadas (x_d, y_d) para a mão direita e (x_e, y_e) para a mão esquerda. Este

cálculo é realizado por meio das médias dos valores das posições dos *pixels*, de acordo com o Algoritmo 1. A mão direita é marcada pela cor azul na figura do esqueleto e a mão esquerda pela cor alaranjada na mesma figura.

4. Duas segmentações são realizadas na imagem RGB tendo como ponto central estes valores de posições calculados. Para que toda a mão esteja incluída na imagem, o valor empírico de tamanho do quadro segmentado foi $141x141$ *pixels*. Este valor permitiu que a mão segmentada de maior tamanho fosse visível. Assim, a imagem com a ROI tem este tamanho e pode ser vista na Figura 16 (e) e Figura 16 (f) para as mãos direita e esquerda, respectivamente.

5. Finalmente utiliza-se um algoritmo para detecção da cor da pele para garantir que somente as mãos sejam detectadas. Esta imagem é, então, convertida para o formato Preto & Branco, conforme visto na Figura 16 (g) e Figura 16 (h), também para as mãos direita e esquerda, respectivamente.

Na Figura 16 apresenta-se o fluxograma do procedimento de detecção da região de interesse a partir das imagens gravadas pelo sensor RGB-D.

O algoritmo para o cálculo das posições (x_d, y_d) e (x_e, y_e) da região de interesse é descrito em Algoritmo 1.

Algoritmo 1: Cálculo das posições (x_d, y_d) e (x_e, y_e)

$x_d \Leftarrow 0$
$y_d \Leftarrow 0$
$x_e \Leftarrow 0$
$y_e \Leftarrow 0$
$tamanhoX \Leftarrow 640$
$tamanhoY \Leftarrow 480$
for i = 1 to tamanhoX **do**
  for j = 1 to tamanhoY **do**
    if cor == 'azul' **then**
      $x_d \Leftarrow x_d + 1$
      $y_d \Leftarrow y_d + 1$
    end if
    if cor == 'alaranjado' **then**
      $x_e \Leftarrow x_e + 1$
      $y_e \Leftarrow y_e + 1$
    end if
  end for
end for
$x_d \Leftarrow media\,(x_d)$
$y_d \Leftarrow media\,(y_d)$
$x_e \Leftarrow media\,(x_e)$
$y_e \Leftarrow media\,(y_e)$

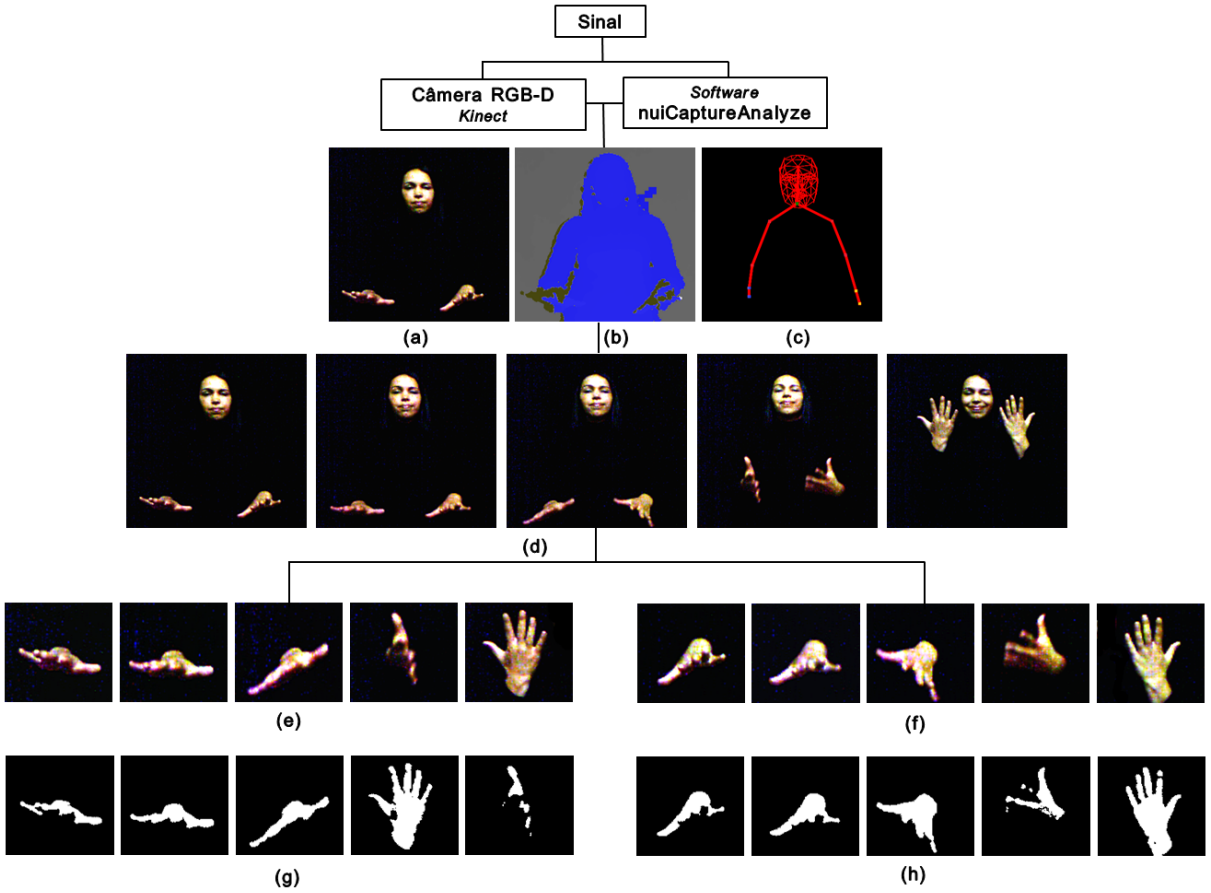


Figura 16: Detecção da região de interesse - mãos direita e esquerda para o sinal "Comemorar". Vídeos gravados utilizando o *software nuiCaptureAnalyze* operando o sensor Kinect. Em (a), vê-se um quadro colorido no formato RGB do vídeo gravado; em (b) um quadro de profundidade do vídeo de profundidade; em (c) um quadro do esqueleto do vídeo do esqueleto. Em (d) estão os quadros selecionados pela sumarização de vídeos por meio da solução do Problema da Diversidade Máxima. Já em (e) e (f) tem-se as regiões de interesse em RGB, detectada para as mãos direita e esquerda, respectivamente. Em (g) e (h) estas regiões de interesse estão detectadas em Preto & Branco por meio do algoritmo de detecção de cor da pele.

A detecção e extração da região de interesse representa a etapa final de processamento da imagem na abordagem para extração de características apresentada neste trabalho. São estas imagens que terão as características extraídas. No próximo capítulo, estas características e as técnicas empregadas para obtê-las são descritas.

4.5 Considerações finais

A maioria dos sinais em Língua Brasileira de Sinais possui movimento, o que permite sua gravação em vídeo. Como a gravação acontece a uma taxa de 30 quadros por segundo, em geral há quadros no vídeo que possuem informações muito próximas umas das outras. Estas informações tornam-se redundantes devido a esta

pouca distinção. Reduzir estes vídeos dos sinais a poucos quadros que representem as informações de parâmetro linguístico tornou-se uma necessidade para melhorar o desempenho da extração de característica em termos de eficiência computacional e em termos de informação real útil contida em cada vídeo.

Tratar o problema da sumarização neste trabalho como um problema de diversidade máxima e solucionar este problema por meio de um algoritmo evolucionário mostrou ser bastante efetivo. Uma alternativa utilizada é amostrar uma certa quantidade de quadros a cada período estabelecido. Entretanto, um sinal tem dinâmicas distintas entre si e pode possuir grande parte dos movimentos em qualquer momento ao longo da sua execução. Assim, o uso das métricas de tempo e diferença de cor para o cálculo da matriz de diversidade no problema da diversidade máxima considera as maiores diferenças entre quadros independente do momento em que estas acontecem.

5 Extração de características

O presente capítulo concentra a descrição e análise de cada uma das características extraídas dos sinais. Ao todo são sete características, apresentadas em seções próprias. Cada uma destas sete características possui relação com um ou mais parâmetros fonológicos da Língua Brasileira de Sinais. Descritas cada uma das características nas seções 5.2 a 5.8, a construção do vetor de características para cada sinal, que é a concatenação das características extraídas, é apresentada e discutida na seção 5.9 ao final do capítulo. Ressalta-se que as características aqui citadas são extraídas posteriormente à sumarização dos vídeos e para cada um dos cinco quadros representativos selecionados nesta sumarização.

5.1 Introdução

Na análise de diversas características que pudessem ser extraídas de sinais, foi dada especial atenção àquelas que apresentassem relação com a estrutura fonológica da Língua Brasileira de Sinais. Esta relação é apresentada, de forma resumida, na Tabela 17.

Seção	Características extraídas	Parâmetros em Libras
5.2	Distância bidimensional	Ponto de articulação
5.3	Distância tridimensional	Ponto de articulação
5.4	Velocidade	Movimento, Orientação da palma da mão
5.5	Área das mãos	Configuração da mão
5.6	Média das posições das quinas das mãos	Ponto de articulação, Movimento, Orientação da palma da mão
5.7	Linhas detectadas	Configuração da mão
5.8	Quantidade de pontos em comum	Movimento, Orientação da palma da mão

Tabela 17: As características extraídas da região de interesse são exibidas na coluna à esquerda. Os parâmetros fonológicos da Libras associados a estas características são exibidos na coluna da direita.

Três características são relacionadas aqui ao parâmetro Ponto de articulação: a distância bidimensional, a distância tridimensional e a média das posições das quinas das mãos. Neste caso, explora-se a particularidade do parâmetro Ponto de articulação de fornecer informações espaciais sobre os sinais.

Para os parâmetros fonológicos Movimento e Orientação da palma da mão há três características que apresentam relação direta: a velocidade, a média das posi-

ções das quinas das mãos e a quantidade de pontos em comum. Neste caso, ambos os parâmetros apresentam informações de deslocamento das mãos no espaço de execução dos sinais. O Movimento fornece informações para o rastreamento da mão ao longo do espaço, com referência no pulso. Já a Orientação da palma da mão, como o próprio nome sugere, diz respeito à posição das palmas das mãos no espaço.

Duas características extraídas possuem relação direta com o formato das mãos: (i) a área de cada uma delas e (ii) as linhas detectadas sobre seu formato. Portanto, elas contêm informações sobre a Configuração da mão.

A abordagem e a implementação da extração de cada uma das características na imagem são descritas em detalhes nas seções a seguir. As duas primeiras características não são extraídas diretamente dos quadros dos vídeos. A primeira distância é um cálculo realizado a partir de informações do esqueleto e a segunda distância é um valor gravado pelo *software* que opera o sensor Kinect.

A partir da seção 5.4, as características descritas são todas extraídas das imagens em Preto & Branco das mãos. A Figura 17 apresenta os cinco quadros para a mão direita de cada um dos três sinais que são utilizados aqui como exemplo para o cálculo das características. O primeiro sinal exemplo é o “Espalhar”. O segundo é o sinal “Agarrar” e o terceiro é o sinal “Futuro”. Os três possuem similaridades e diferenças e estas serão discutidas para cada característica ao longo das respectivas seções.

Os parâmetros fonológicos de cada um destes três sinais de exemplo são exibidos na Tabela 18. Há semelhanças e diferenças entre eles para diversos valores dos atributos dos parâmetros.

Parâmetros	**Espalhar**	**Agarrar**	**Futuro**
Ponto de articulação	À direita do tronco	À direita do tronco	À direita da cabeça
Configuração de mão	 (a) G1 - Forma 1 Alinhamento em z Sofre alteração	 (b) G1 - Forma 4 Alinhamento em z Sofre alteração	 (c) G4 - Forma 8 Alinhamento em y Não sofre alteração
Movimento	Para a direita Simples	Para dentro Simples	Para fora Simples
Orientação da palma da mão	Para baixo Não sofre alteração	Para fora Sofre alteração	Para o lado Não sofre alteração

Tabela 18: Parâmetros fonológicos dos sinais utilizados como exemplo. Todos os parâmetros são para a mão direita.

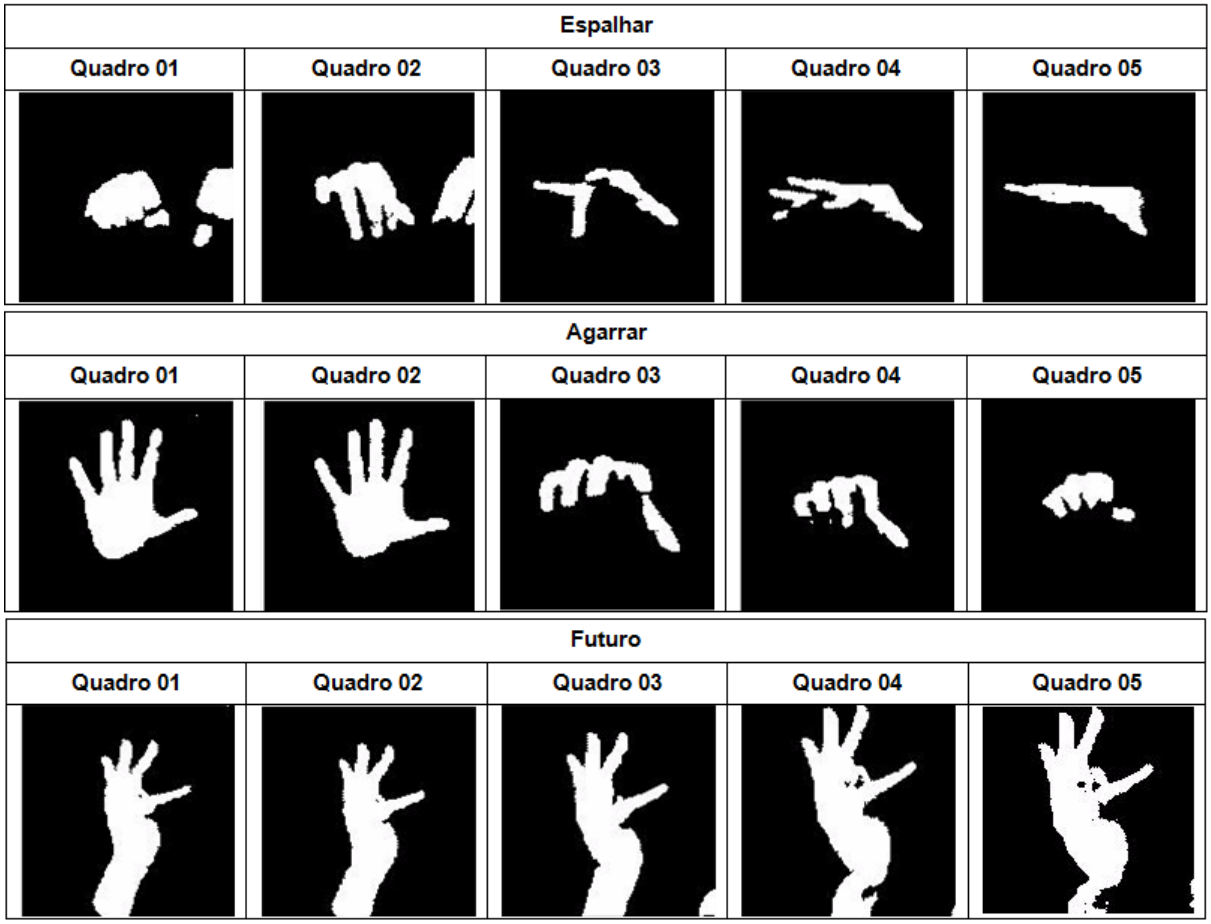


Figura 17: Sinais utilizados como exemplo para a extração das características. São exibidos os cinco quadros selecionados sequencialmente e todos para a mão direita. Os cálculos das características também são realizados para a mão esquerda, embora não apareçam nos exemplos.

5.2 Distância bidimensional

A distância bidimensional aqui calculada busca explorar a relação existente entre esta e o parâmetro Ponto de articulação em Libras, o qual tem por definição a posição do sinal no espaço. Esta distância é calculada entre as coordenadas das posições dos *pixels* que representam as mãos direita e esquerda e o ombro na imagem do esqueleto. A métrica utilizada é a distância euclidiana.

Calculam-se, de fato, dois valores: um é a distância bidimensional entre mão direita e o ombro e o outro é a distância bidimensional entre a mão esquerda e o ombro. A referência para o cálculo desta distância é centro do ombro, uma vez que ele possui alterações insignificantes de posição ao longo da execução de cada sinal, pois fica praticamente parado. Assim, quanto mais próximo a mão estiver do ombro, menor será este valor de distância.

Cada uma das destas três regiões (mão direita, mão esquerda e ombro) é representada por um pequeno conjunto de *pixels* no vídeo do esqueleto gravado para

cada sinal. Estas regiões diferenciam-se nas imagens pela cor. Para a mão direita os *pixels* são azuis. Já para a mão esquerda os *pixels* estão na cor alaranjada e, por fim, a cor verde é utilizada para marcar a posição do ombro nas imagens, conforme vê-se na Figura 18.

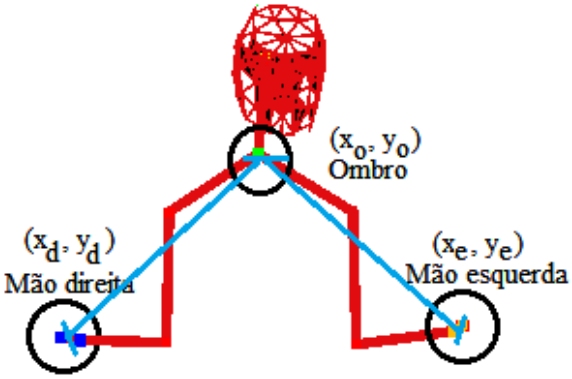


Figura 18: (a) Exemplo das posições (x, y) extraídas da imagem do esqueleto. A região da mão direita está gravada na cor azul; a região da mão esquerda está gravada na cor alaranjada. O ombro está gravado na cor verde.

Para o cálculo dos pontos dos *pixels* que representam as mãos e ombros, uma média das posições de cada cor em torno da região desejada é realizada, conforme Equações 5.1 e 5.2.

$$d_{d,o} = \sqrt{(x_d - x_o)^2 + (y_d - y_o)^2} \tag{5.1}$$

$$x_d = media\,(x_{azul})$$
$$x_o = media\,(x_{verde})$$

$$d_{e,o} = \sqrt{(x_e - x_o)^2 + (y_e - y_o)^2} \tag{5.2}$$

$$x_e = media\,(x_{alaranjado})$$
$$x_o = media\,(x_{verde})$$

onde:

$d_{d,o}$: distância euclidiana calculada entre as médias das coordenadas dos *pixels* da mão direita e do ombro.

$d_{e,o}$: distância euclidiana calculada entre as médias das coordenadas dos *pixels* da mão esquerda e do ombro.

(x_d, y_d): média das coordenadas dos *pixels* da mão direita no esqueleto.

(x_e, y_e): média das coordenadas dos *pixels* da mão esquerda no esqueleto.

(x_o, y_o): média das coordenadas dos *pixels* do ombro no esqueleto.

5.3 Distância tridimensional

A distância tridimensional é gravada diretamente pelo *software NuiCaptureAnalyze*.

Durante a gravação de cada sinal são gerados quatro tipos de informações distintas. Três destas informações são: (i) os vídeos de intensidade de cor (em RGB), (ii) os vídeos de profundidade e (iii) os vídeos do esqueleto com as posições detectadas marcadas por cores próprias. A quarta informação gravada é um arquivo no formato Matlab (MathWorks (2012)). Nele, estão todas as distâncias entre o sensor Kinect e os vinte pontos do corpo detectados conforme visto na Figura 11. O acesso a estas informações de posição é feito conforme a estrutura de programação encapsulada *Skeleton* descrita na Figura 19 e disponibilizada pelo fabricante do *software NuiCaptureAnalyze*[1].

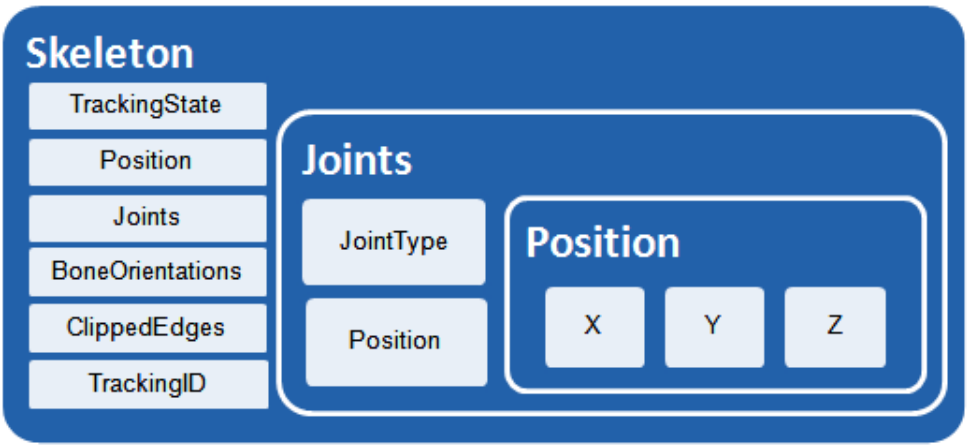


Figura 19: Estrutura computacional do arquivo em formato Matlab (MathWorks (2012) gravado pelo *NuiCaptureAnalyze*.

Estas posições gravadas podem ser acessadas, assim, pelos seguintes comandos:

- Skeleton.Joints.Position.X para o valor de x;
- Skeleton.Joints.Position.Y para o valor de y;
- Skeleton.Joints.Position.Z para o valor de z.

[1] <http://nuicapture.com/support/how-to-export/#reading-mat-files>.

Os valores estão gravados em milímetros para cada ponto em três dimensões (x, y, z). O eixo coordenado utilizado como referência é exibido na Figura 20.

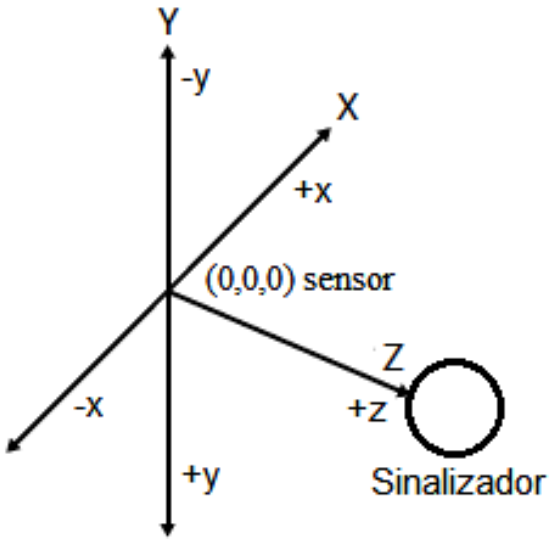


Figura 20: Sistema de coordenadas de referência para a gravação da distância tridimensional (x, y, z) utilizada como característica.

Estes valores de (x, y, z) foram utilizados diretamente como característica. Neste caso, são as distâncias entre o sensor Kinect e as regiões de interesse (mão direita e mão esquerda do sinalizador). Assim, tem-se três valores de distância para a mão direita e três para a mão esquerda, totalizando seis valores.

5.4 Velocidade

A velocidade do movimento de padrões de brilho é uma característica calculada sobre a imagem da região de interesse em Preto & Branco por meio da técnica de fluxo óptico. Esta técnica é definida por Horn e Schunck (1981) como “a distribuição de velocidades aparentes do movimento de padrões de brilho em uma imagem. O fluxo óptico surge do movimento relativo entre objetos e visualizadores. Consequentemente, o fluxo óptico pode fornecer informações importantes sobre a configuração espacial de objetos visualizados e a taxa de mudança destas configurações.”

Esta técnica aplicada a vídeos fornece informações importantes sobre a distribuição de objetos em uma cena e suas alterações ao longo do tempo. Os vetores de fluxo óptico são calculados a partir da derivação do brilho E da imagem no ponto (x, y) e tempo t, função definida como $E(x, y, t)$ por Horn e Schunck (1981). Cada conjunto de vetores é apresentado no quadro pertinente em vermelho para os três sinais exemplo na Figura 21.

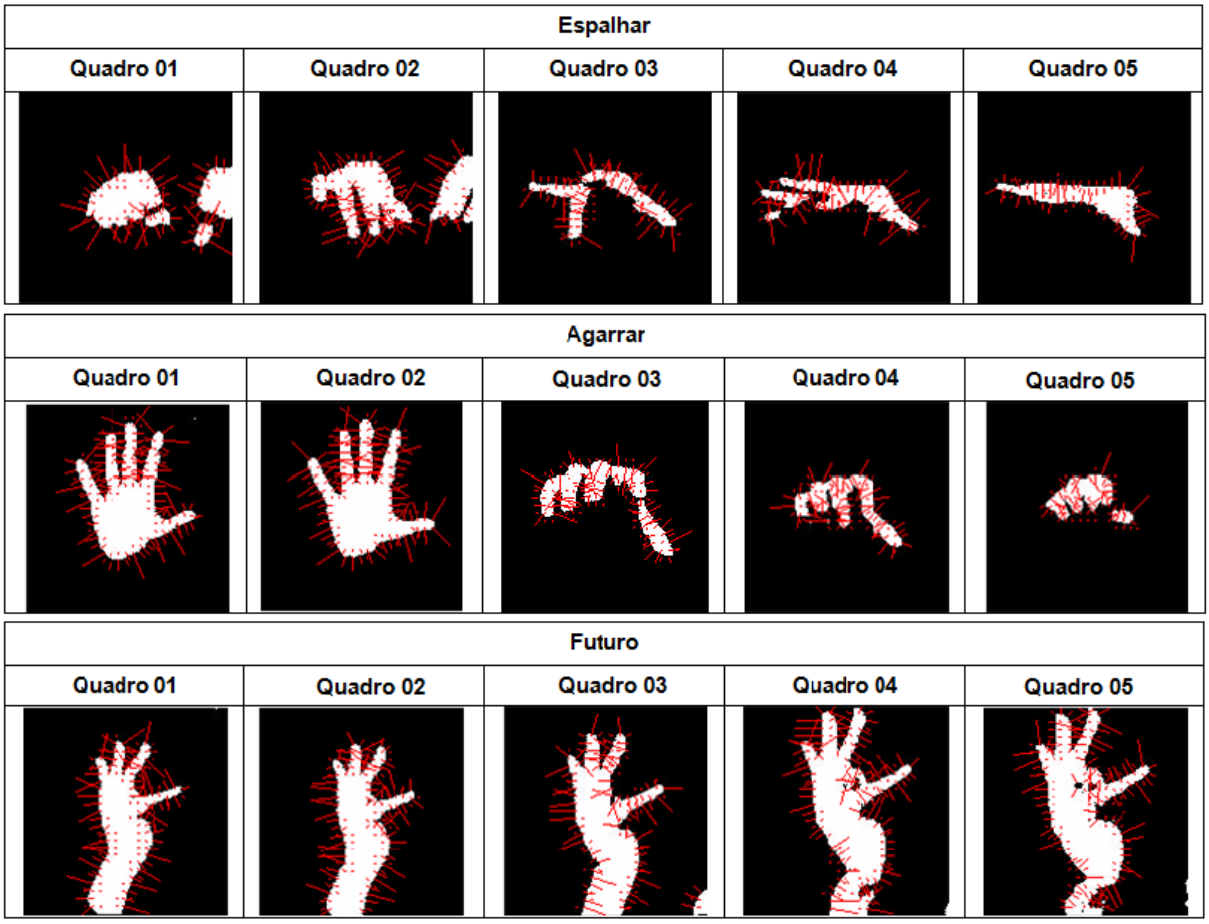


Figura 21: Vetores de movimento para cada uma das cinco imagens da mão direita dos três sinais exemplo: “Espalhar”, “Agarrar” e “Futuro”. Os vetores estão destacados em vermelho e posicionados onde há variação de brilho entre os quadros.

Um único valor de velocidade é utilizado para a 3a. característica neste trabalho em cada um dos cinco quadros de cada uma das mãos. Este valor é um limiar calculado a partir da matriz de velocidades da técnica de fluxo óptico. Esta matriz é composta por valores de vetores complexos. A magnitude dos vetores é calculada e seu valor, após ter sua média calculada, é utilizada como o limiar.

Como dito anteriormente, os vetores de velocidade são valores complexos. Eles estão plotados e exibidos na Figura 22, representando os vetores de fluxo óptico para cada quadro dos sinais “Espalhar”, “Agarrar” e “Futuro” na Figura 21. Os valores utilizados como característica são as magnitudes destes vetores.

Observa-se que cada sinal apresenta, para cada um dos quadros selecionados, vetores bem distintos. No primeiro caso, sinal “Espalhar”, a mão direita está aberta, mas com a palma para baixo e a imagem captura este sinal de frente. Para o sinal “Agarrar”, a mão direita começa aberta, com a palma para fora, e gradativamente fecha-se. Já para o sinal “Futuro”, a mão direita não está completamente aberta. Cada um dos dedos da mão direita encontra-se em uma direção. Este formato, entretanto,

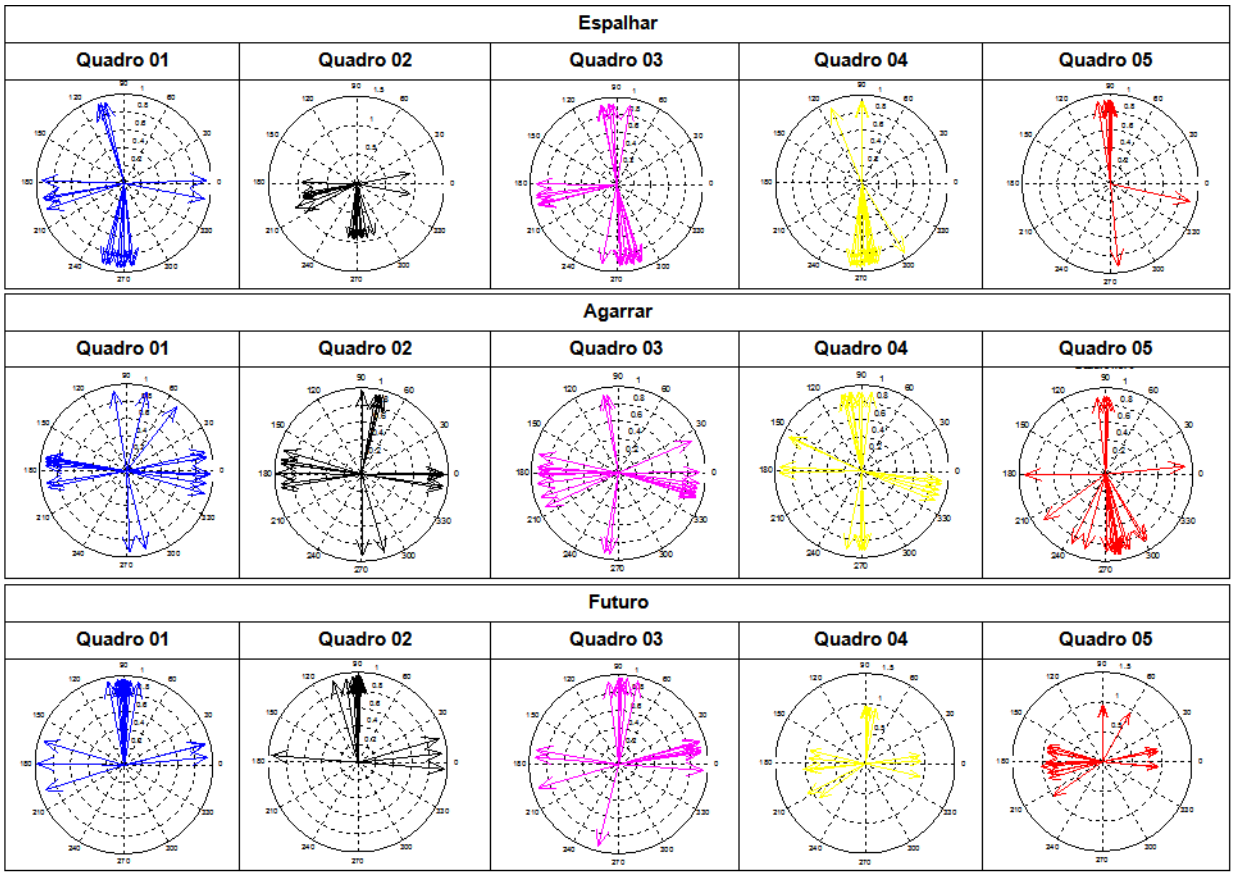


Figura 22: Vetores de velocidade calculados pelo fluxo óptico para cada um dos cinco quadros dos três sinais exemplo: “Espalhar”, “Agarrar” e “Futuro”.

é mantido com a mão movendo-se para a frente. As diferenças de brilho utilizadas para o cálculo desta característica são, portanto, bastantes diversas entre os sinais, gerando padrões próprios para cada um deles.

5.5 Área das mãos

A técnica de fluxo óptico também foi utilizada para a extração da característica “Área das mãos”. A mesma ideia de diferenciação da função brilho é válida neste cálculo. Para a característica velocidade os vetores de fluxo óptico fornecem informações de movimento. Já para a característica descrita aqui a diferença de brilho permite a segmentação de objetos em uma imagem. Nas imagens dos sinais, no máximo dois objetos são segmentados. Este valor pode acontecer quando ocorre no sinal uma aproximação entre as duas mãos em algum momento de sua execução.

Na Figura 23, retângulos em verde representam as áreas calculadas para cada um dos quadros selecionados nos três sinais exemplo: “Espalhar”, “Agarrar” e “Futuro”, mão direita. Nota-se que tanto o formato da mão quanto a forma como ela aparece na gravação podem alterar de forma significativa o valor destas áreas.

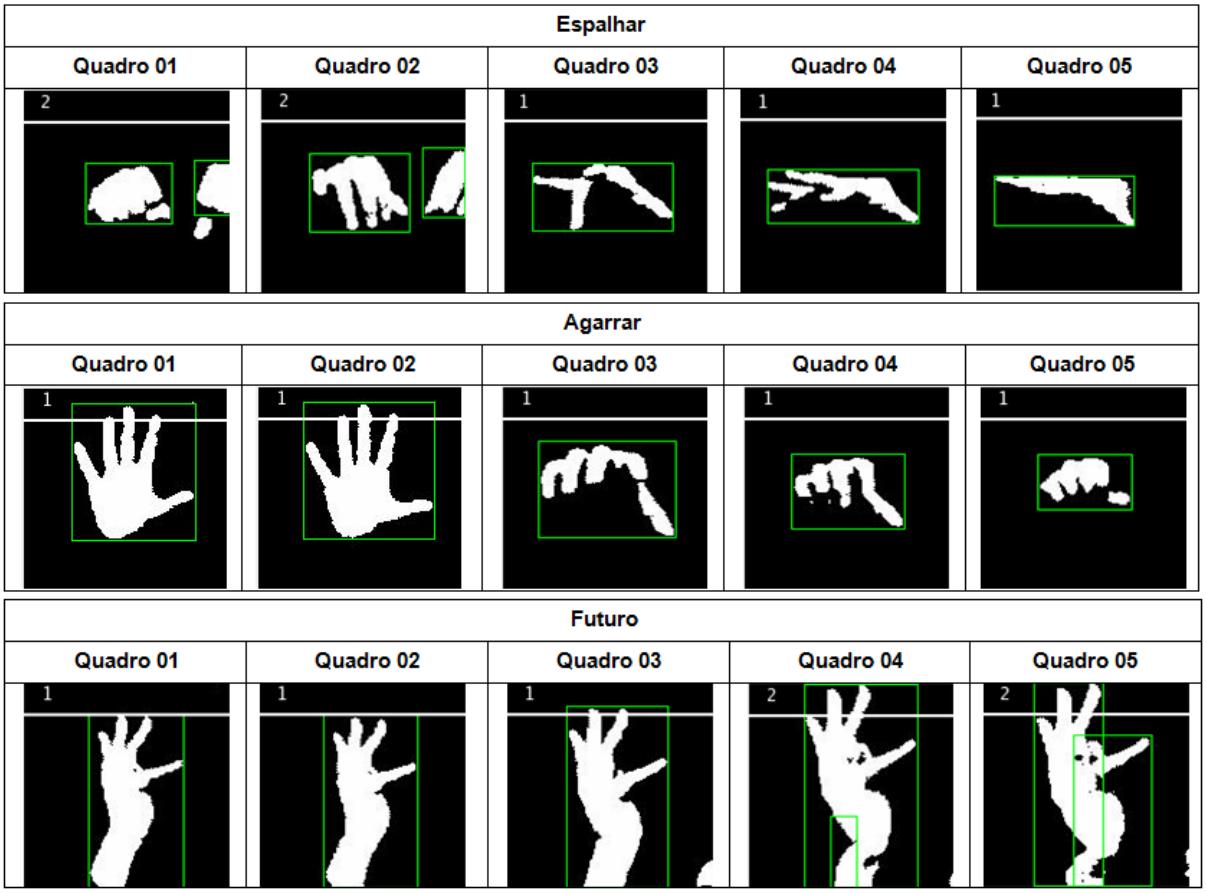


Figura 23: Áreas das mãos definidas pelo retângulo em cor verde em cada um dos cinco quadros para a mão direita dos sinais exemplo “Espalhar”, “Agarrar” e “Futuro”.

No gráfico da Figura 24 há uma comparação entre os valores de áreas das mãos direitas dos três sinais exemplo: “Espalhar”, “Agarrar” e “Futuro”. Cada imagem tem o tamanho 141 x 141 *pixels*. Assim, o valor máximo da área será de 19.881 $pixels^2$. Quanto maior o valor de área, mais a mão em questão está aberta e visível.

Para o primeiro sinal, “Espalhar”, vemos que a área primeiramente aumenta, em seguida se mantém estável e por fim diminui, o que pode ser comprovado pela Figura 23. No segundo sinal, “Agarrar”, primeiramente o valor da área aumenta e, em seguida, conforme a mão vai sendo fechada, ela diminui. Pode-se ver este padrão na Figura 23. Por fim, com relação ao terceiro sinal da comparação, “Futuro”, a mão fica na mesma posição e, portanto, a área permanece a mesma nos três primeiros quadros. No quarto quadro a área aumenta porque a mão dirige-se ao sensor e, no quinto quadro, a área cai porque a mão é segmentada pelo algoritmo em duas áreas distintas. Esta segmentação ao final em duas áreas acontece justamente pela aproximação da mão ao sensor. Embora a mão como um todo torne-se mais visível, os espaços entre os dedos também aumentam e acabam por ter tamanho suficiente para segmentar a mão em duas áreas.

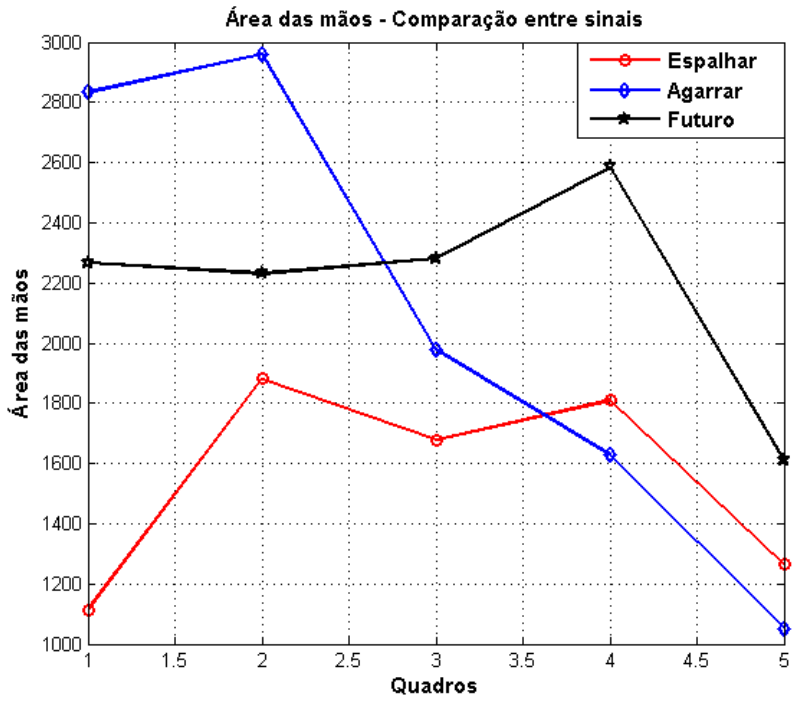


Figura 24: Valores das áreas das mãos definida pelo retângulo em *pixels*^{2} em cada um dos cinco quadros para a mão direita para os sinais “Espalhar”, “Agarrar” e “Futuro”.

5.6 Média das posições das quinas das mãos

Quinas podem ser definidas em processamento de imagens como a interseção de duas bordas. Elas também podem ser vistas como pontos para o quais há duas direções de bordas diferentes e dominantes na vizinhança local destes pontos. A localização de quinas em imagens informam sobre diferenças entre objetos e fundo e diferenças de profundidade. À medida em que um objeto move-se, suas quinas também sofrerão alterações. Assim, informações de deslocamento no espaço são possíveis de detectar também por meio da localização das quinas em imagens que compõem vídeos.

É possível encontrar diversas técnicas para o cálculo das quinas em uma imagem, como em (HARRIS; STEPHENS, 1988), em (SHI; TOMASI, 1994) e em (ROSTEN; DRUMMOND, 2006).

Na aproximação realizada neste trabalho as quinas são detectadas utilizando o detector de quinas de *Harris*, conforme descrito em (HARRIS; STEPHENS, 1988). Cada uma destas quinas possui valores de coordenadas de posição (x, y). O valor médio deste par ordenado é utilizado como característica extraída das imagens para cada quadro.

É possível ver todos os pontos que representam as quinas para cada um dos quadros dos sinais utilizados como exemplo na Figura 25, “Espalhar”, “Agarrar” e “Fu-

turo". O valor médio das posições destes pontos está destacado em vermelho.

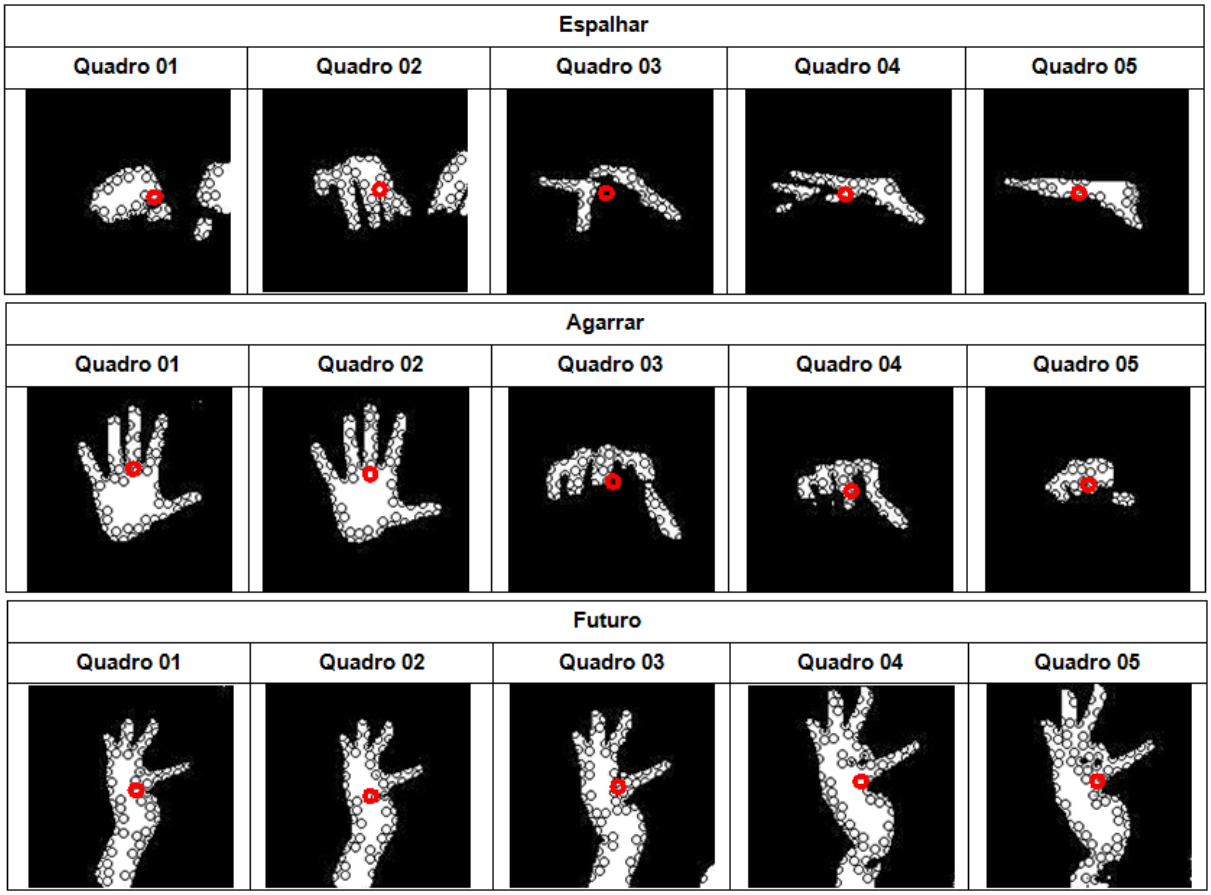


Figura 25: Característica extraída "Média das posições das quinas" entre os sinais "Espalhar", "Agarrar" e "Futuro".

O gráfico apresentado na Figura 26 apresenta o valor médio das posições das quinas para os sinais exemplos. A diferença entre cada um dos valores é razoavelmente estável. Entretanto, nota-se que cada um dos três sinais imprime seu trajeto de forma bem definida e distinta entre eles.

5.7 Linhas detectadas

A linha de maior valor encontrada nas imagens é utilizada como característica extraída neste caso. Cada linha é representada pelo par ordenado (ρ, θ) em coordenadas polares. Elas são calculadas para cada imagem pela aplicação da transformada de *Hough*, técnica esta descrita em (ILLINGWORTH; KITTLER, 1988). No trabalho aqui apresentado decidiu-se por utilizar ρ e θ da maior linha de cada um dos cinco quadros como os valores desta característica para cada sinal.

O algoritmo da transformada de *Hough* padrão (*STH - Standard Transform Hough*) é utilizado para o cálculo das linhas nas imagens. A STH utiliza uma repre-

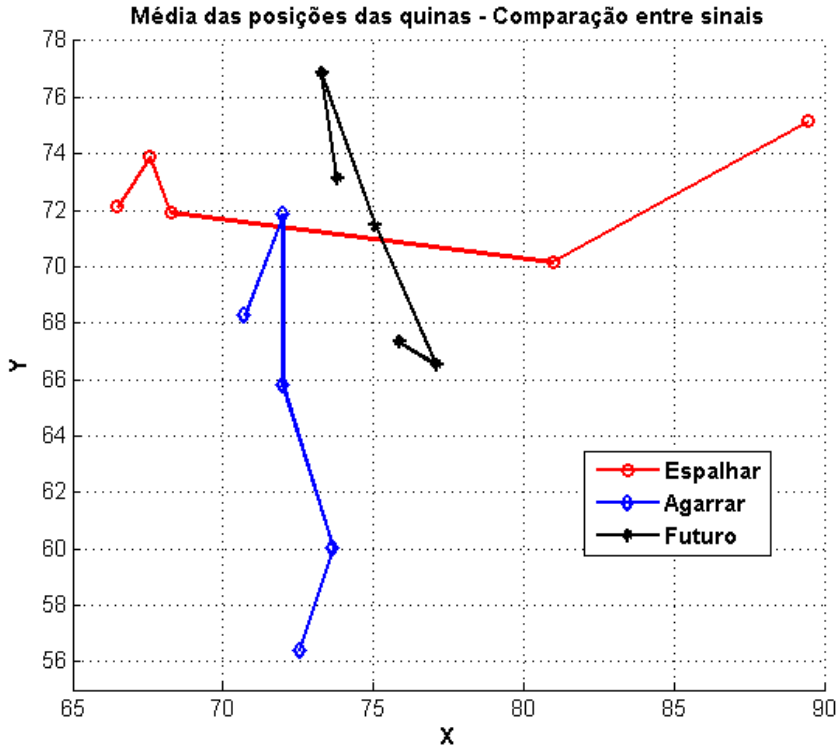


Figura 26: Valores da característica “Média das posições das quinas” entre os sinais “Espalhar”, “Agarrar” e “Futuro”.

sentação paramétrica das linhas conforme Equação 5.3.

$$\rho = x * cos(\theta) + y * sen(\theta) \tag{5.3}$$

onde:

ρ é a distância da origem dos eixos coordenados (x, y) até o cruzamento perpendicular à linha detectada pela transformada de *Hough*.

θ é o ângulo entre o eixo x e a linha definida por ρ, conforme Figura 27.

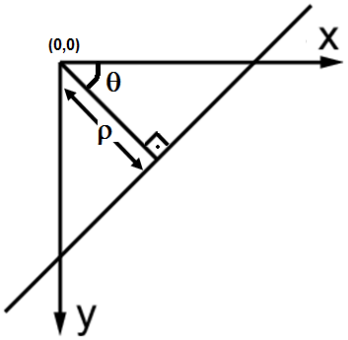


Figura 27: Definição das variáveis ρ e θ na detecção de linhas por meio da transformada de *Hough*.

Na Figura 28 as cinco maiores linhas detectadas para cada um dos cinco quadros dos sinais exemplo “Espalhar”, “Agarrar” e “Futuro” são apresentadas, respectivamente. A maior linha para cada quadro dentre as cinco maiores linhas apresentadas está na cor vermelha. A transformada de *Hough* é aplicada sobre a borda detectada em cada imagem e uma matriz *H* com os valores de todas as linhas é gerada. Optou-se por utilizar neste trabalho a detecção de borda de Canny, criada por Canny (1986). Os valores de características utilizadas aqui, para cada quadro, são o par ordenado (ρ, θ) da maior linha detectada em cada um deles.

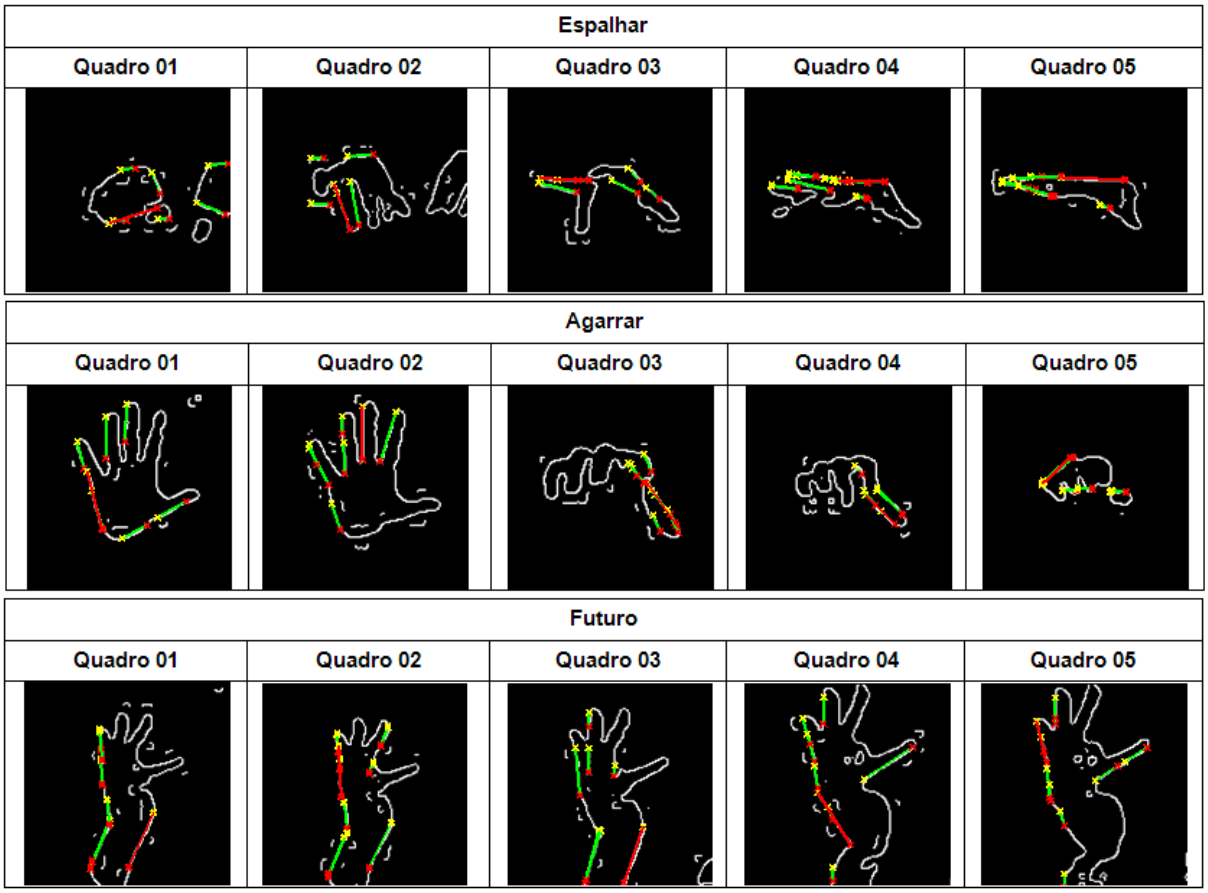


Figura 28: Característica “Linhas detectadas” para cada um dos quadros dos sinais exemplos “Espalhar”, “Agarrar” e “Futuro”. As cinco maiores linhas estão descritas em verde em cada quadro, com o ponto inicial em vermelho e o final em amarelo. A linha detectada é aquela de maior tamanho e seus valores de ρ e θ são as características extraídas. Elas estão destacadas em vermelho.

No caso do sinal “Espalhar” a mão encontra-se com a palma para baixo e orientada com o eixo x, da esquerda para a direita. Assim, possui linhas de maior valor neste eixo. Já para o sinal “Agarrar”, a mão encontra-se na maior parte do tempo, fechada e em movimento para a frente. As linhas aumentam seu valor ao longo da execução do sinal, pois a imagem fica maior à medida que o sinal evolui. No caso do sinal “Futuro”, as linhas são maiores no sentido y, uma vez que a mão em questão está com sua palma virada para o lado e alinhada de baixo para cima em y.

Como as coordenadas são polares, o gráfico representado na Figura 29 apresenta o valor ordenado (ρ, θ) das cinco linhas de maior valor detectadas em cada quadro, que podem ser vistas aplicadas sobre as imagens dos sinais na Figura 28. Como é de se esperar, o perfil dos valores do par ordenado é bastante distinto entre si.

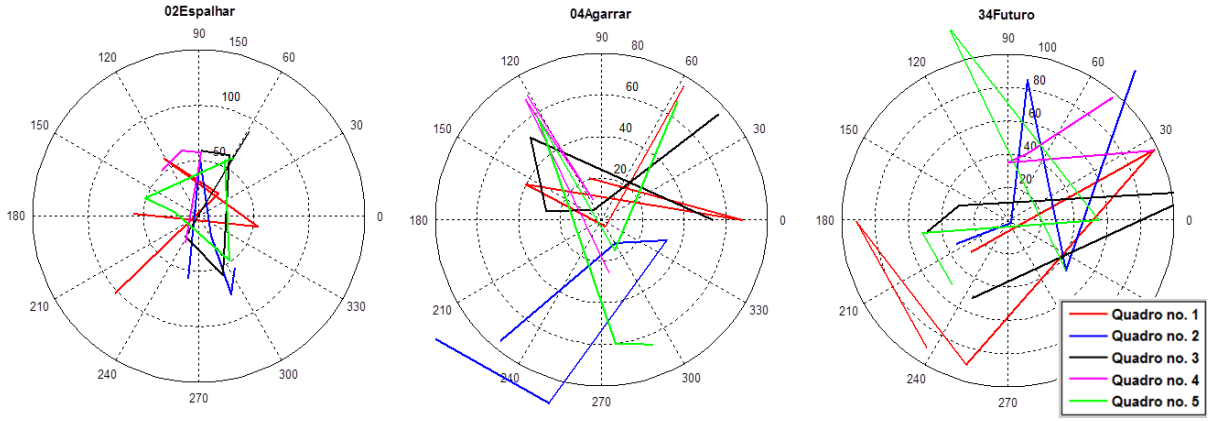


Figura 29: Valores dos pares ordenados (ρ, θ) das cinco maiores linhas detectadas em cada um dos cinco quadros dos sinais “Espalhar”, “Agarrar” e “Futuro.”

5.8 Quantidade de pontos comuns entre quadros

Esta característica extraída é composta pela quantidade de pontos que são comuns entre cada par de quadros subsequentes dos vídeos dos sinais. Como a comparação é uma diferença entre quadros, em ordem temporal, temos quatro valores calculados para cada sinal, e não cinco, como nos casos das características extraídas anteriores.

Para o cálculo destes pontos em comum foi utilizado o descritor SURF (Speeded Up Robust Features). Este descritor foi proposto por Bay et al. (2008) e é composto por funções de detecção, extração e casamento de características de interesse. Ele descreve uma distribuição da intensidade dentro dos *pixels* vizinhos ao ponto de interesse, o que também é realizado pela informação de gradiente em outro descritor bastante utilizado na área de visão computacional, o SIFT (Scale-Invariant Feature Transform). Este último é descrito em detalhes em Lowe (2004), assim como suas variações.

Especificamente sobre o descritor SURF, tem-se que este implementa transformadas *Wavelet* utilizando funções de *Haar* como funções base sobre as imagens, com o objetivo de reduzir o tempo para processamento da detecção, extração e casamento das imagens. Ele também propõe aumentar a robustez do reconhecimento de padrões. Sobre as propriedades das transformadas Wavelet, ver Daubechies (1990).

Os autores em (BAY et al., 2008) implementam o SURF em três etapas. São elas:

1. a primeira etapa consiste em fixar uma orientação reprodutível baseada em informação obtida de uma região circular em torno do ponto de interesse;
2. na segunda etapa, uma região quadrada é alinhada à orientação selecionada previamente e o descritor SURF é extraído desta região;
3. por fim, na terceira etapa, as características extraídas correspondentes em um par de imagens são obtidas.

Esta última etapa é chamada *Template matching* e sua descrição completa está disponível no trabalho apresentado pelo autor Brunelli (2009), em seu livro **Template Matching Techniques in Computer Vision: Theory and Practice.**

Na Figura 30, observam-se dois quadros sobrepostos a cada imagem, para os sinais exemplo “Espalhar”, ‘Agarrar” e “Futuro”. Em cada uma das imagens, os pontos em comum advindos da aplicação da técnica de *Template matching* estão marcados por círculos vermelhos e cruzes verdes e interligados por linhas de cor amarela.

No gráfico da Figura 31, os três sinais exemplo tem suas quantidades de pontos em comum exibidas. Observa-se aqui que esta característica também imprime um padrão distinto dos demais para cada sinal.

5.9 Construção do vetor de características

Um vetor de características é construído para cada parâmetro fonológico da Libras. O número de características em cada vetor depende da relação parâmetro fonológico *versus* característica que o representa. Baseando-se na Tabela 17, tem-se os seguintes vetores:

- Para o reconhecimento do parâmetro Ponto de articulação, foram utilizadas as características: (i) distância bidimensional, (ii) distância tridimensional e (iii) média das posições das quinas.
- Para o reconhecimento do parâmetro Configuração da mão, as características extraídas (i) área das mãos e (ii) linhas detectadas foram utilizadas.
- Para o parâmetro Movimento foram selecionadas as características (i) velocidade, (ii) média das posições das quinas e (iii) quantidade de pontos comuns entre quadros para construir o vetor correspondente.

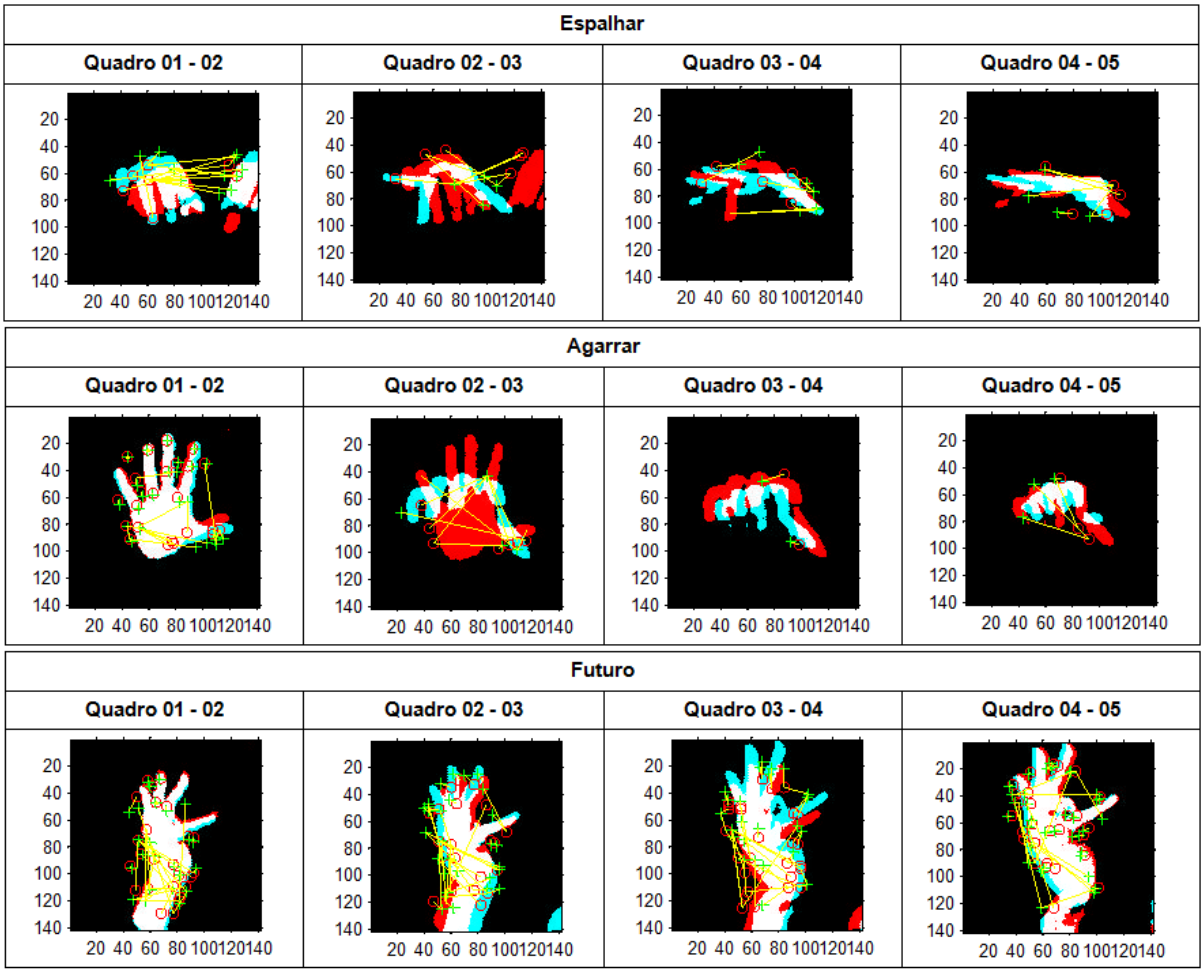


Figura 30: Diferenças entre os quadros 1-2, 2-3, 3-4 e 4-5 em sequência para os sinais exemplo “Espalhar”, “Agarrar” e “Futuro”.

- Finalmente, para o parâmetro Orientação da palma da mão, o vetor de carcterísticas é composto de (i) velocidade, (ii) média das posições das quinas e (iii) quantidade de pontos comuns entre quadros.

A estrutura destes quatro vetores é idêntica, alterando-se apenas a quantidade de características que tornará possível o reconhecimento de cada parâmetro fonológico da Libras. Esta estrutura geral para o vetor de características é exibida na Tabela 19.

As características *1 a M*, onde *M* é a quantidade de características para cada um dos parâmetros fonológicos, são extraídas do primeiro quadro representativo no vídeo $(n = 1)$ e, de acordo com o parâmetro em questão, seus valores são adicionados neste vetor. Isto é realizado para os n quadros em sequência. Ao final da sua construção, temos um vetor para cada um dos quatro parâmetros fonológicos da Libras em cada mão. Estes vetores são as entradas para o sistema de classificação de parâmetros apresentado no Capítulo 6 descrito a seguir.

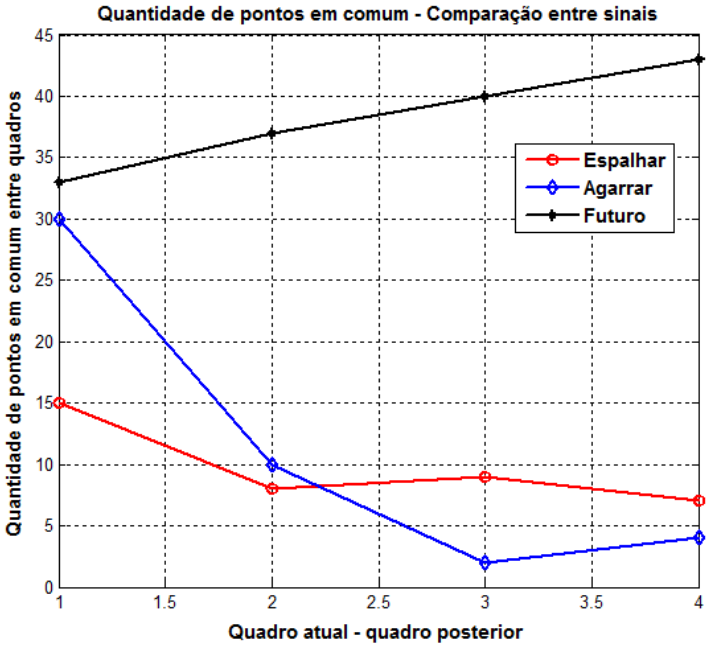


Figura 31: Quantidades de pontos em comum para cada uma das combinações 1-2, 2-3, 3-4 e 4-5 para os sinais “Espalhar”, “Agarrar” e “Futuro”.

Mão (direita ou esquerda)			
1	...	n	Quadros
1 a M	1 a M	1 a M	Características

Tabela 19: Estrutura do vetor de características para *M* características e *n* quadros. Neste estudo são analisadas $M = 1, 2, 3$ características, dependendo do parâmetro e $n = 5$ quadros.

Sendo o tempo uma importante variável no reconhecimento de sinais dinâmicos, adiciona-se seu valor de duas formas distintas aos vetores de características construídos. Assim, ao final deste capítulo, constrõem-se três tipos de vetores de características, conforme descrito a seguir:

1. o vetor de característica sem informação de tempo, conforme estrutura descrita na Tabela 19;

2. o vetor de característica com a variável tempo de cada quadro multiplicada pelos valores destas características, conforme estrutura descrita na Tabela 20.

3. o vetor de característica com os valores de tempo de cada quadro adicionado a ele, conforme estrutura descrita na Tabela 21;

Mão (direita ou esquerda)			
1	...	n	Quadros
1.t_1 a M.t$_1$	...	1.t_1 a M.t$_n$	Características

Tabela 20: Estrutura do vetor de características para *M* características e *n* quadros com a variável tempo multiplicada a cada característica.

Mão (direita ou esquerda)				
1	...	n		Quadros
1 a M	1 a M	1 a M	t_1 a t_n	Características

Tabela 21: Estrutura do vetor de características para *M* características e *n* quadros com a variável tempo adicionada ao final.

5.10 Considerações finais

Para que fosse possível extrair as sete características descritas neste capítulo, diversas etapas previamente precisaram ser implementadas, uma vez que a extração de características deve, necessariamente, ser realizada na região de interesse do problema.

As características de distâncias não operam diretamente sobre as imagens das mãos. A distância bidimensional utiliza informações da imagem do esqueleto que, por meio da marcação de regiões com cores distintas torna o seu valor limitado ao cálculo de distâncias euclidianas de valores médios. A distância tridimensional, embora também apresente o valor de uma distância, tem seu valor obtido diretamente por meio da solução escolhida para a gravação de sinais. Ao contrário da distância bidimensional, que compara valores de distâncias entre as mãos, a distância tridimensional consegue informações de profundidade no movimento das mãos porque grava a distância entre estas e o sensor, adicionando assim o eixo z ao valor.

Duas características são obtidas pela mesma técnica de fluxo óptico: a velocidade e a área das mãos. Embora sejam fruto da mesma teoria, elas fornecem informações relacionadas a parâmetros fonológicos distintos da Língua Brasileira de Sinais.

As técnicas utilizadas nos cálculos das características Média das posições das quinas, Linhas detectadas e Quantidade de pontos comuns entre quadros são, respectivamente, a detecção de quinas de *Harris*, a transformada de *Hough* e o descritor SURF. Todas elas são técnicas clássicas já bem consolidadas nos estudos de reconhecimento de padrões, dentre outras áreas. A forma como elas são aplicadas e interpretadas para que forneçam os valores de características relacionados a parâmetros da Libras é o diferencial.

Por fim, para que o tempo de execução de cada sinal possa ter sua importância

dimensionada por meio de experimentos, seu valor é incorporado ao vetor de características de duas formas distintas. Assim tem-se três vetores construídos, um no qual o tempo não é considerado e os outros dois com abordagens do tempo distintas.

No Capítulo 6 a seguir, cada um destes três vetores será utilizado como entrada em um sistema supervisionado de classificação de parâmetros da Língua Brasileira de Sinais.

6 Experimentos e resultados

O presente capítulo apresenta a descrição dos experimentos para a análise das características extraídas dos sinais. A ideia principal na elaboração destes experimentos é verificar e validar a relação existente entre cada uma das sete características extraídas dos sinais e os respectivos parâmetros fonológicos da Língua Brasileira de Sinais. O tempo de ocorrência de cada quadro dos sinais selecionados é uma informação adicional que também tem sua influência analisada em parte dos experimentos. Uma Máquina de Vetor de Suporte (SVM) foi implementada para a verificação destas relações por meio do reconhecimento automático dos parâmetros fonológicos de cada sinal. A aprendizagem é dita supervisionada, tendo como entrada as características extraídas e como valor desejado os valores pré-estabelecidos de cada atributo para cada um dos 34 sinais gravados. Realizada a classificação dos parâmetros para cada mão, os resultados são concatenados na formação de cada sinal. Esta saída obtida é comparada à saída desejada e um reconhecimento dos sinais é realizado. Embora este não seja o objetivo final específico deste trabalho, é importante analisar a classificação dos sinais resultante, inclusive com relação ao fato de que nem todos os parâmetros fonológicos da Língua Brasileira de Sinais estão contemplados neste trabalho.

6.1 Introdução

A classificação dos parâmetros de forma supervisionada foi utilizada para validar esta relação entre as sete características e quatro parâmetros fonológicos da Língua Brasileira de Sinais. A Máquina de Vetor de Suporte foi o sistema implementado para realizar esta classificação devido a sua capacidade de resolver problemas de classificação e regressão. Ela aprende na etapa de treinamento e seleciona dentre os dados de treinamento pontos que formarão um vetor de suporte para a classificação dos dados de teste. Desta forma, ela é capaz de generalizar e pode realizar, por exemplo, a classificação binária. É esta classificação que é realizada para cada um dos parâmetros fonológicos dos sinais.

Dois diferentes experimentos foram conduzidos neste trabalho. No primeiro, avaliou-se a interferência na mudança de *kernel* da Máquina de Vetor de Suporte na classificação de cada parâmetro dos sinais. Para tal, dois *kernels* foram utilizados: o linear e o RBF. No segundo experimento comparou-se o desempenho da classificação

dos parâmetros para os três vetores de características construídos no Capítulo 5: (i) o primeiro sem informação de tempo, (ii) o segundo com a informação de tempo multiplicada pelos valores das características extraídas e (iii) o terceiro com o valor do tempo adicionado ao final do vetor com as características, aumentando neste caso o tamanho final do vetor. Neste segundo experimento foi mantido o *kernel* que apresentou melhor desempenho: o RBF.

A análise de variância do desempenho do classificador foi realizada em ambos os experimentos. Desta forma, os resultados são apresentados em gráficos do tipo *box plot* e em todos os casos para as mãos direita e esquerda separadamente. A classificação final dos sinais a partir de todos os parâmetros reconhecidos é realizada por comparação entre o vetor de saída dos parâmetros obtidos pela Máquina de Vetor de Suporte e o vetor de parâmetros desejado para cada sinal. Após concatenar o resultado da classificação de parâmetros da mão direita e esquerda em um vetor que caracteriza o sinal, este teve o cálculo de distância para cada um dos sinais desejados realizada. A métrica utilizada foi a distância de *Hamming*. Os resultados são apresentados ao final do capítulo.

6.2 Experimentos

O diagrama da Figura 32 apresenta a estrutura utilizada para os experimentos de classificação supervisionada dos parâmetros linguísticos. Realiza-se esta classificação para a mão direita e para a mão esquerda de forma independente.

O sistema de classificação é construído por uma Máquina de Vetor de Suporte para cada valor possível de cada um dos parâmetros fonológicos. Assim, tem-se quatro blocos de Máquinas de Vetor de Suporte, cada um relacionado aos valores dos atributos do Ponto de articulação, Configuração de mão, Movimento e Orientação da palma da mão. Na Tabela 22 estes 32 valores dos atributos de cada um dos quatro parâmetros está descrito para a mão direita. O mesmo ocorre na Tabela 23 que apresenta 29 valores possíveis para a mão esquerda. São, portanto, 32 Máquinas de Vetor de Suporte para a classificação dos parâmetros observando a mão direita e 29 Máquinas de Vetor de Suporte para a classificação dos parâmetros da mão esquerda.

Cada um dos quatro blocos de máquinas de vetor de suporte implementado possui como entrada o vetor de características extraído dos sinais. São sete características ao todo. Entretanto, a quantidade de características M na entrada para cada um dos blocos depende da relação entre cada uma destas com os parâmetros fonológicos, conforme visto na Tabela 24.

Assim, o vetor de características fornecido como entrada para cada um dos blocos de máquinas de vetor de suporte é composto pelas M características pertinen-

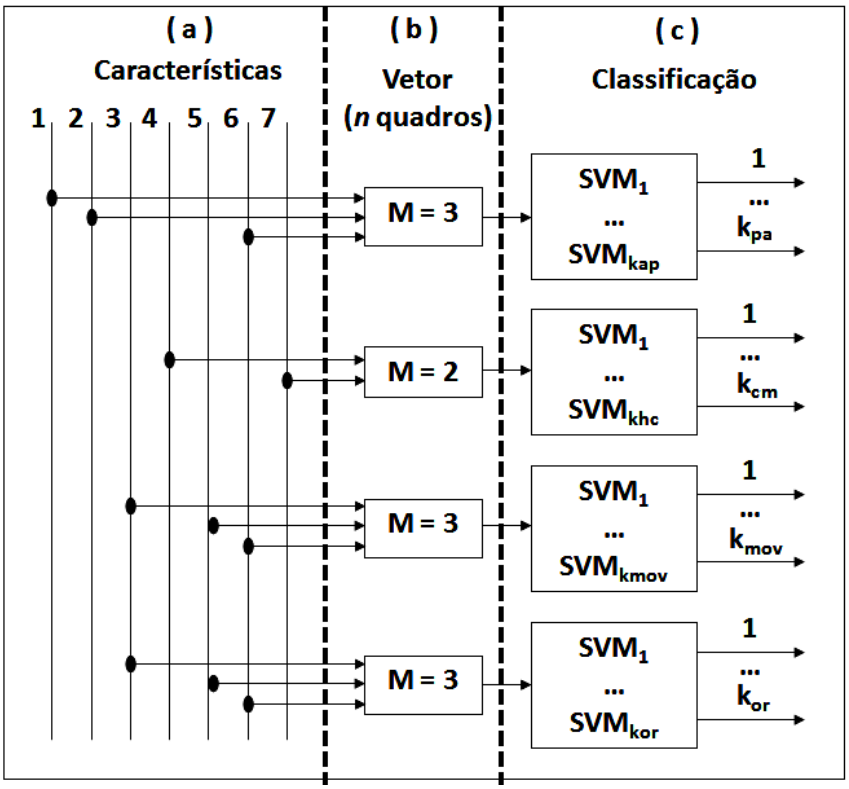


Figura 32: Esquemático do sistema implementado. (a) Características: sete características extraídas. (b) Vetor: veja Tabela 19 para a estrutura, $n = 5$ quadros selecionados e M = número de características para um dado elemento em Libras, veja a Tabela 17 para referência. (c) Classificação: k_{pa} saídas para o Ponto de articulação ($k_{pa} = 5$ para a mão direita e $k_{pa} = 4$ para a mão esquerda); k_{cm} saídas para a Configuração de mão ($k_{cm} = 12$ para a mão direita e $k_{cm} = 10$ para a mão esquerda); k_{mov} saídas para o Movimento ($k_{mov} = 8$ para as mãos direita e esquerda); k_{or} saídas para a orientação da palma da mão ($k_{or} = 7$ para ambas as mãos, direita e esquerda).

	Parâmetros	**Atributos**	**Valores**
(i)	Ponto de articulação	cabeça	{1. direita, 2. centro}
		ombro	{3. direita}
		tronco	{4. direita, 5. centro}
(ii)	Configuração de mão	grupo	{6. G1, 7. G2, 8. G4, 9. G6, 10. G7, 11. G8, 12. G10 }
		eixo de alinhamento	{13. em x, 14. em y, 15. em z}
		variação	{16. sim, 17. não}
(iii)	Movimento	tipo	{18. para cima, 19. para baixo, 20. para a direita, 21. para a esquerda, 22. para dentro, 23. para fora}
		frequência	{24. simples, 25. repetido}
(iv)	Orientação da palma da mão	orientação	{26. para cima, 27. para baixo, 28. para dentro, 29. para fora, 30. para o lado}
		variação	{31. sim, 32. não}

Tabela 22: Saídas das Máquinas de Vetor de Suporte para cada um dos quatro parâmetros fonológicos da Língua Brasileira de Sinais para a mão direta.

tes a cada parâmetro fonológico extraídas dos $n = 5$ quadros selecionados para cada sinal.

	Parâmetros	**Atributos**	**Valores**
(i)	Ponto de articulação	cabeça	{1. esquerda}
		ombro	{2. esquerda}
		tronco	{3. esquerda, 4. centro}
(ii)	Configuração de mão	grupo	{5. G1, 6. G4, 7. G6, 8. G8, 9. G10}
		eixo de alinhamento	{10. em x, 11. em y, 12. em z}
		variação	{13. sim, 14. não}
(iii)	Movimento	tipo	{15. para cima, 16. para baixo, 17. para a direita, 18. para a esquerda, 19. para dentro, 20. para fora}
		frequência	{21. simples, 22. repetido}
(iv)	Orientação da palma da mão	orientação	{23. para cima, 24. para baixo, 25. para dentro, 26. para fora, 27. para o lado}
		variação	{28. sim, 29. não}

Tabela 23: Saídas das Máquinas de Vetor de Suporte para cada um dos quatro parâmetros fonológicos da Língua Brasileira de Sinais para a mão esquerda.

	Parâmetros	**Características**	**M**
(i)	Ponto de articulação	Distância bidimensional (1) Distância tridimensional (2) Média das posições das quinas (5)	3
(ii)	Configuração de mão	Área da mão (4) Linhas detectadas (6)	2
(iii)	Movimento	Velocidade (3)	3
(iv)	Orientação da palma da mão	Média das posições das quinas (5) Quantidade de pontos em comum (7)	

Tabela 24: Características relacionadas a cada um dos quatro parâmetros fonológicos da Língua Brasileira de Sinais que compõem o vetor de entrada de cada bloco de máquinas de vetor de suporte.

Há duas fontes de aleatoriedade nos experimentos, além de um viés. As fontes de aleatoriedade são: (i) os dados de treinamento e os (ii) os dados de teste. Já os parâmetros dos *kernels* dos blocos de Máquina de Vetor de Suporte constituem o viés. Os dados de treinamento e teste são os valores de características extraídos. Em cada avaliação de reconhecimento de sinais o experimento de classificação foi executado 10 vezes, variando aleatoriamente as amostras utilizadas para treinamento e para teste. Cabe lembrar aqui que cada sinal foi gravado cinco vezes por um único sinalizador e que as sete características são extraídas dos quadros que compõem cada um destes sinais. Para o caso do treinamento foram utilizadas sempre três amostras dentre as cinco possíveis para cada sinal. Já para o caso do teste, duas amostras dentre as cinco foram aleatoriamente selecionadas. As médias e desvios padrões para cada resultado foram estabelecidos por meio da análise estatística de variância, a ANOVA (Analysis of Variance) e os resultados são apresentados em gráficos do tipo *box plot.* Estes gráficos são chamados diagrama de caixa e captam importantes aspectos dos conjuntos de dados, apresentando, dentre outros valores, os acertos mínimos e má-

ximos. Assim, além de analisar o classificador com relação a média, pode-se analisar também a sua variabilidade.

6.3 Classificação dos parâmetros fonológicos dos sinais

O esquema geral descrito na Figura 32 é a referência para a construção e implementação dos dois experimentos de classificação dos parâmetros dos sinais. A primeira variação é com relação ao *kernel* da Máquina de Vetor de Suporte. Dois *kernels* são comparados: o linear e o RBF. São apresentados resultados de comparação entre ambos para os três vetores de características distintos como entrada das Máquinas de Vetor de Suporte, descritos nas Tabelas 19, 20 e 21. Os *kernels* são descritos a seguir.

O *kernel* linear é um produto (*dot produt)*, como apresentado na Equação 6.1.

$$f(x_i, x_j) = \sum_{i=1}^{S} \sum_{j=1}^{S} x_i \cdot x_j \tag{6.1}$$

onde x_i e x_j são os vetores de características dos sinais do banco de sinais.

O *kernel* RBF para a Máquina de Vetor de Suporte é implementado conforme a Equação 6.2.

$$f(x_i, x_j) = \sum_{i=1}^{S} \sum_{j=1}^{S} \exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right) \tag{6.2}$$

onde x_i e x_j são os vetores de características dos sinais do banco de sinais.

No segundo experimento a variação avaliada está entre os três vetores construídos conforme as Tabelas 19, 20 e 21. Neste caso o *kernel* é o RBF, por ter obtido o melhor desempenho no primeiro experimento.

Cada uma das seções seguintes trata de um dos dois experimentos, apresentando seus resultados para os dados de teste. As comparações para cada experimento são feitas aos pares, para a mão direita primeiro e para a mão esquerda logo em seguida.

6.3.1 Comparação entre os *kernels* Linear e RBF

Nas comparações realizadas a seguir, duas questões devem ser observadas. O valor médio de acerto do sistema de classificação, por questões óbvias, é o primeiro. Quanto maior este valor, melhor o desempenho para o experimento em questão. A

segunda observação é quanto à variação em torno desta média. Se este valor for muito elevado, ou mais elevado que o valor sob comparação, a indicação é o desempenho melhor é do experimento com menor variação.

As Figuras 33 e 34 apresentam os resultados para a mão direita e para a mão esquerda, respectivamente. A entrada é o vetor de características nomeado como 1 conforme Tabela 19. Neste caso, o valor do tempo de execução para cada um dos quadros selecionados não está presente.

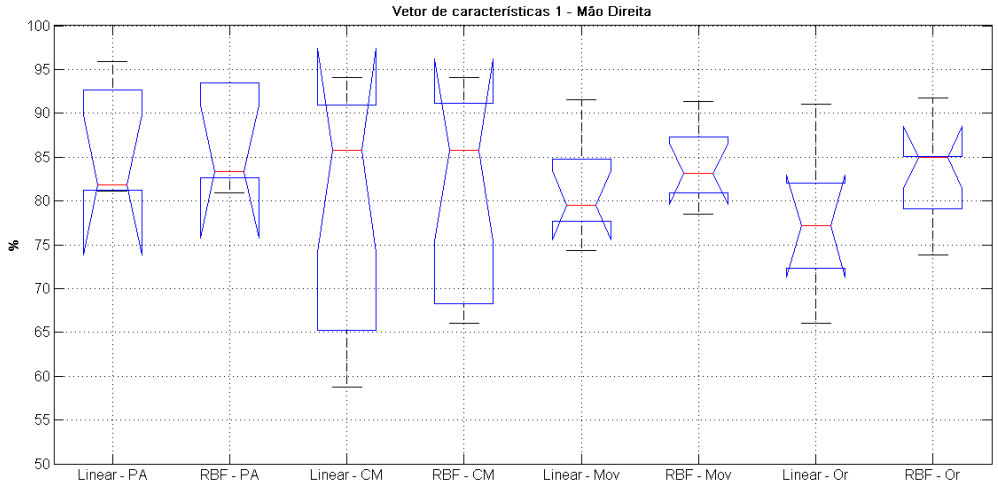


Figura 33: Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 19.

Na comparação entre *kernels*, o *kernel* RBF apresentou maior desempenho na classificação dos parâmetros para todos os quatro parâmetros. No caso do parâmetro Configuração de mão, este desempenho melhor do RBF em termos da média é bastante sutil. Entretanto, a variabilidade do desempenho para a Configuração de mão é visivelmente maior no caso do *kernel* linear. O que deseja-se é que a variabilidade seja a menor possível. Portanto, de fato, o *kernel* RBF tem melhor desempenho, que está entre 83,1 a 85,8%.

Para o caso da mão esquerda, tem-se que os resultados do *kernel* RBF são melhores para os parâmetros Configuração de mão e Orientação da palma da mão. Por consequência, para os parâmetros Ponto de Articulação e Movimento, o *kernel* linear possui valores mais altos. Entretanto, o alto valor de variação nos valores do Movimento indicam que o *kernel* RBF é melhor. Neste caso, o desempenho do sistema de classificação de parâmetros está entre 85,5 e 93,41%.

As Figuras 35 e 36 apresentam os resultados da classificação o vetor construído conforme Tabela 20, para as mãos direita e esquerda, respectivamente. Neste

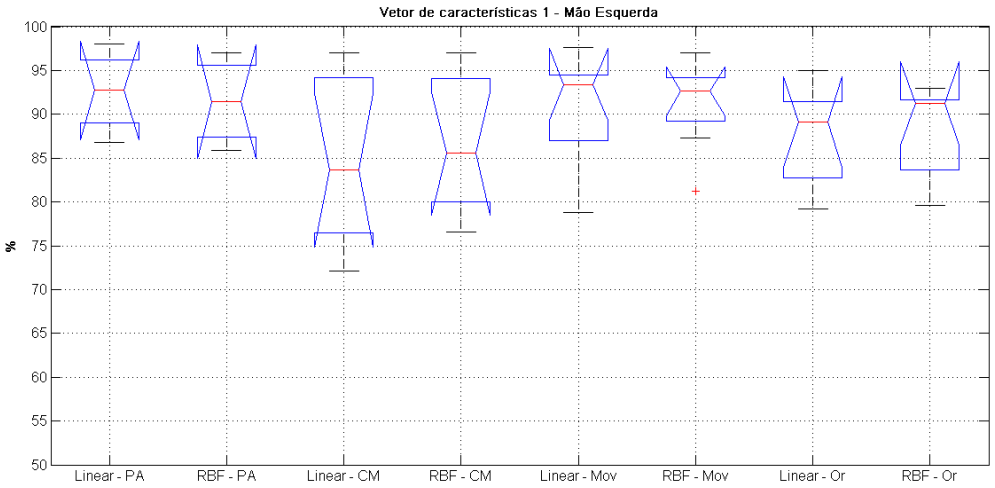


Figura 34: Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 19.

caso, a informação de tempo está presente, multiplicado por cada valor do vetor de características.

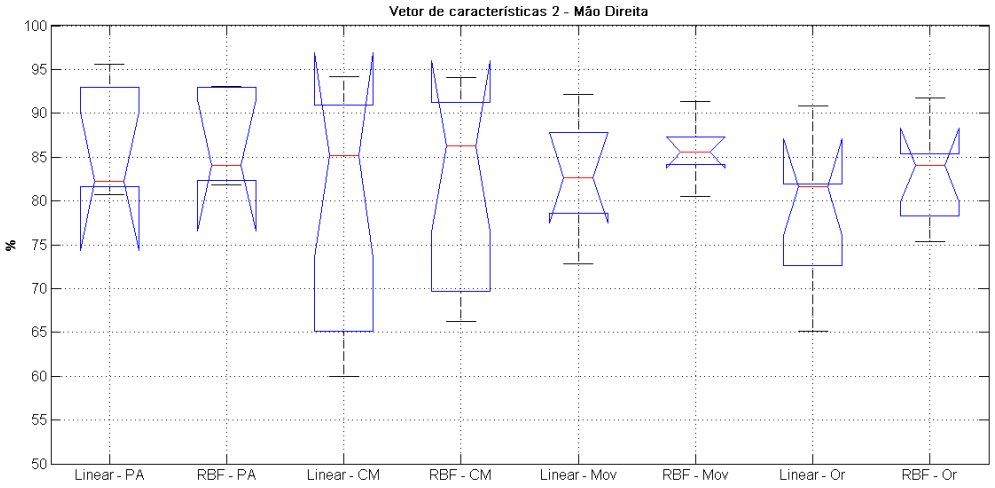


Figura 35: Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 20.

Para o caso da mão direita, tem-se que o melhor desempenho ocorre para o *kernel* RBF em todos os quatro parâmetros fonológicos. Neste caso, ao comparar-se a variação dos valores em torno da média, tem-se novamente que o uso do *kernel* RBF apresentou melhores resultados com menor variação. Os acertos médios estão

entre 84,1 e 86,3%.

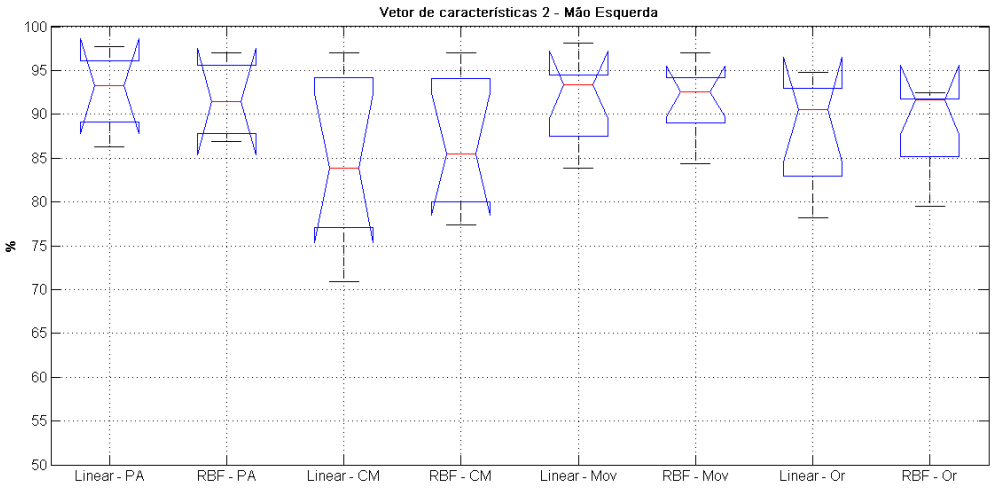


Figura 36: Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 20.

Vamos ao caso da mão esquerda para este vetor de características. O resultado parece-se bastante com o resultado apresentado na comparação para o vetor 1. Os parâmetros Ponto de articulação e Movimento com melhor desempenho utilizando o *kernel* linear e os parâmetros Configuração de mão e Orientação da palma da mão com valores mais altos para o *kernel* RBF. Em termos de variação, o parâmetro Movimento apresenta melhor resultado utilizando-se o *kernel* RBF. Os valores de desempenho estão entre 85,5 e 93,4% de acerto.

Já o par de Figuras 37 e 38 apresenta os resultados de classificação de parâmetros obtidos com o vetor de característica construído conforme Tabela 21 tanto para a mão direita quanto para a mão esquerda. Neste terceiro tipo de vetor, a informação do tempo é concatenada ao seu final. Portanto, o tamanho deste vetor de características, se comparado aos outros dois citados anteriormente, possui 5 valores a mais.

Os valores médios de acerto na classificação dos quatro parâmetros fonológicos apresenta resultados de desempenho visivelmente maiores quando se utiliza o *kernel* RBF. Ainda utilizando-se o *kernel* RBF, tem-se uma menor variação nas médias e, consequentemente, melhor desempenho. Estes valores de médias estão entre 81,6 e 85,5% de acerto na classificação dos parâmetros fonológicos.

Por fim, comparando-se os valores de acerto a partir dos *kernels* linear e RBF, para este vetor de características, tem-se que o uso do *kernel* RBF apresentou melhor desempenho. Este resultado também aparece para a variação em torno das médias,

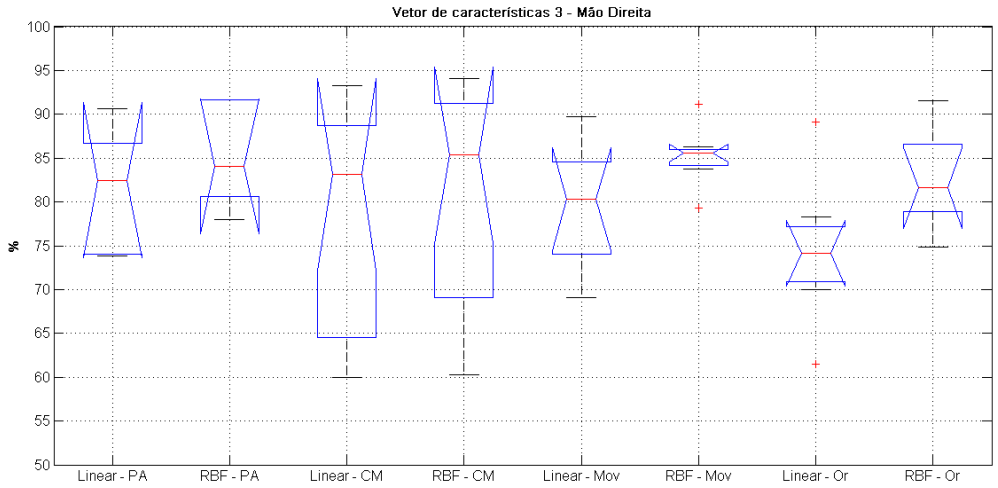


Figura 37: Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21.

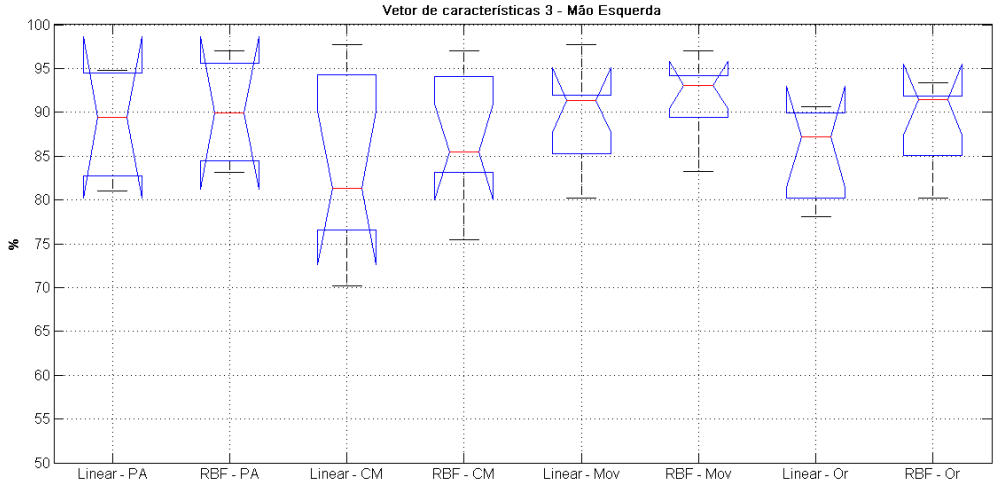


Figura 38: Comparação entre os *kernels* linear e RBF na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov) and (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21.

que é maior para o *kernel* linear. Os valores médios dos acertos, neste caso, estão entre 85,5 e 93,1%.

6.3.2 Comparação entre os três vetores de características propostos

A comparação de desempenho entre os três vetores de características obtidos conforme Tabelas 19, 20 e 21, mantendo o *kernel* RBF na Máquina de Vetor de Suporte, são apresentadas nesta seção.

Neste caso, o par de Figuras 39 e 40 apresenta, respectivamente, os resultados para a mão direita e para a mão esquerda. Os três vetores já são comparados em cada uma das figuras.

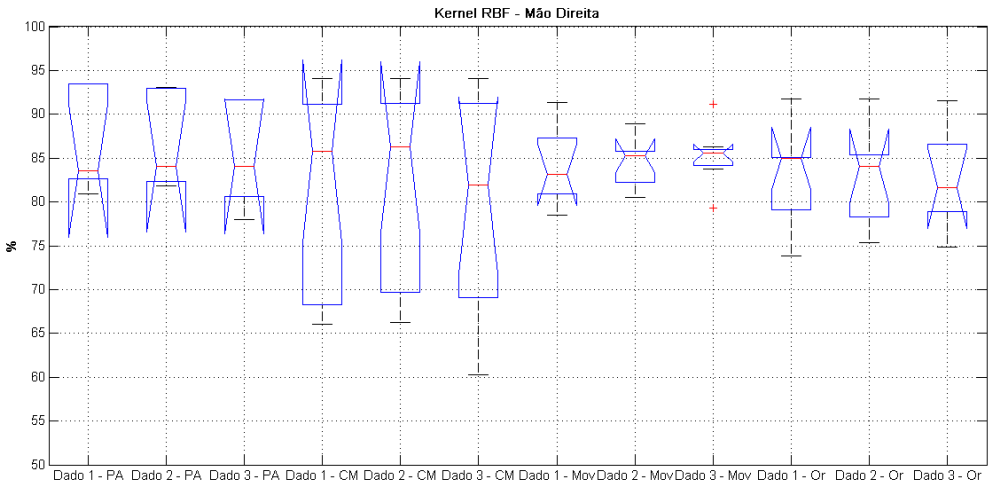


Figura 39: Comparação entre os vetores de características na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov), (iv) Orientação da palma da mão (Or) para a mão direita. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21.

Para a mão direita, o vetor de características com o tempo de informação adicionado (o vetor 3) teve uma precisão levemente superior aos outros dois vetores, especialmente para o parâmetro Configuração de mão. Para os outros parâmetros não houve nenhuma diferença estatisticamente significativa em adicionar ou não a informação de tempo ao vetor de características.

Para o caso da mão esquerda, comparando-se os 3 diferentes vetores, percebe-se que, exceto pelo parâmetro Ponto de articulação, cujo vetor 3 tem o pior desempenho visível, não há diferença significativa no desempenho dos parâmetros Configuração de mão, Movimento e Orientação da palma da mão.

Se retornarmos, entretanto, ao experimento anterior, é possível notar que o caso do vetor 3 foi a única situação na qual o desempenho foi melhor para os quatro parâmetros na mão esquerda.

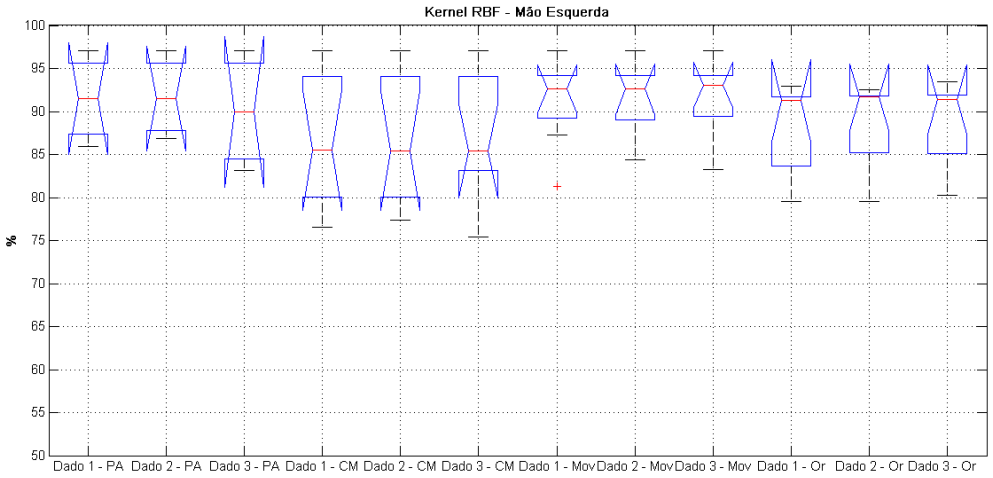


Figura 40: Comparação entre os vetores de características na classificação dos parâmetros fonológicos da Libras: (i) Ponto de articulação (PA), (ii) Configuração da mão (CM), (iii) Movimento (Mov), (iv) Orientação da palma da mão (Or) para a mão esquerda. Os resultados foram obtidos com o vetor de característica construído conforme Tabela 21.

É importante salientar também que a mão esquerda mantém-se parada em 12 dos 34 sinais selecionados para gravação, e fica junto ao tronco do lado esquerdo. Portanto, só há o parâmetro de treinamento supervisionado para ela com relação ao Ponto de articulação. As outras posições do vetor de características assumem o valor "0", indicando que não há valor naquele atributo e para aquele parâmetro. Isto torna a gama tanto de dados de treinamento quanto de dados de teste empobrecida, pois diminui a quantidade de valores a ser comparada e o sistema tem maior dificuldade em diferenciar as situações para cada sinal.

6.4 Reconhecimento dos sinais

Os experimentos nas seções anteriores mostram que a eficácia na identificação e classificação dos valores nos parâmetros fonológicos foi alta. Entretanto, esta informação deve ser integrada de alguma forma para reconhecer um sinal de forma única dentre outros. Isto pode ser realizado de diferentes formas. Nesta seção foram realizados experimentos utilizando um método simples para combinar esta informação: as saídas do classificador de parâmetros são comparadas para cada valor de cada parâmetro fonológico em ambas as mãos para os valores pré-estabelecidos e desejados nas Figuras 9 e 10.

A distância *Hamming* é calculada entre os vetores de saída (que são os parâmetros classificados) e todos os sinais das Figuras 9 e 10. Assim, tem-se uma

distância entre os parâmetros de um sinal com relação a todos os outros. A menor distância deve acontecer entre os vetores que representam os mesmos sinais. No caso aqui apresentado se, entre os *k* pontos mais próximos (com $k = 1, 3, 5$) estiver o sinal selecionado, então consideramos que houve acerto. Os valores de acerto são apresentados, portanto, para os valores de $k = 1, 3, 5$.

Sinais	**k = 1**	**k = 3**	**k = 5**
1. Pessoa	95 %	100 %	100 %
2. Espalhar	51 %	76 %	93 %
3. Copiar	67 %	79 %	93 %
4. Agarrar	0 %	6 %	7 %
5. Recolher	63 %	80 %	87 %
6. Sumir	94 %	97 %	100 %
7. Olhar	98 %	100 %	100 %
8. Justo	74 %	100 %	100 %
9. Verdade	43 %	84 %	96 %
10. Peso	88 %	98 %	99 %
11. Justiça	24 %	65 %	72 %
12. Quem	66 %	100 %	100 %
13. Nada	81 %	96 %	98 %
14. Acreditar	80 %	92 %	97 %
15. Esquecer	75 %	100 %	100 %
16. Amar	20 %	100 %	100 %
17. Angustiado	43 %	79 %	91 %
18. Comemorar	7 %	16 %	21 %
19. Rancor	20 %	99 %	100 %
20. Assembleia	7 %	51 %	70 %
21. Comparar	30 %	79 %	92 %
22. Gritar	98 %	100 %	100 %
23. Falar	70 %	100 %	100 %
24. Absorver	86 %	98 %	97 %
25. Engordar	17 %	43 %	55 %
26. Brigar	18 %	45 %	69 %
27. Esperto	22 %	67 %	100 %
28. Brilhar	88 %	100 %	100 %
29. Empregado	35 %	64 %	70 %
30. Substituir	38 %	66 %	78 %
31. Prisão	37 %	49 %	66 %
32. Televisão	0 %	4 %	6 %
33. Ontem	0 %	2 %	5 %
34. Futuro	1 %	1 %	1 %

Tabela 25: Eficácia média para o reconhecimento de cada um dos 34 sinais na base de sinais. Cada vetor de saída da Máquina de Vetor de Suporte é comparado às saídas desejadas apresentadas nas Figuras 9 e 10. A distância de *Hamming* é calculada entre estes vetores e os resultados são apresentados para os casos em que o acerto ocorre entre os **k = 1, 3, 5** vizinhos mais próximos.

7 Conclusões

7.1 Introdução

A extração de características, que é uma das etapas de reconhecimento de padrões, leva a uma série de desafios e possibilidades, em especial quando se trata de reconhecer sinais em uma língua de estrutura diversa e complexa como uma língua de sinais. Portanto, uma série de conclusões adquiridas ao longo do desenvolvimento deste trabalho de extração de características serão colocadas, além de propostas de ideias que direcionem trabalhos futuros.

7.2 Conclusões

No presente trabalho foi apresentada uma abordagem para extração de características de sinais da Língua Brasileira de Sinais. Esta metodologia envolve desde a escolha de quais sinais são representativos na língua até a implementação de um classificador automático de sinais.

Para que o objetivo final de classificar parâmetros fonológicos da Língua Brasileira de Sinais fosse alcançado, alguns desafios surgiram e foram modelados para o auxílio à classificação dos parâmetros.

O primeiro desafio foi criar uma base de sinais em Língua Brasileira de Sinais. Para isto, duas frentes de trabalho distintas surgiram. A primeira envolveu conhecimentos da estrutura da Língua Brasileira de Sinais. Diante de mais de 10.000 palavras documentadas em um dos muitos dicionários existentes, quantos e quais sinais podem ser tidos como representativos? A opção foi, a partir da referência de Capovilla, Raphael e Maurício (2012a), Capovilla, Raphael e Maurício (2012b), escolher um sinal de cada um dos agrupamentos morfêmicos descritos em seu trabalho. O segundo desafio foi a gravação destes sinais. A escolha por apenas um sinalizador foi feita na expectativa de que neste momento sejam analisadas as diferenças entre parâmetros fonológicos entre os sinais. Assim, a variabilidade nas formas como diferentes sinalizadores executariam os sinais, além de diferenças de altura e cor dos sinalizadores, não afetam a análise realizada nesta tese. Como a ideia é utilizar um sistema tido como uma interface o mais natural possível para a interação homem-máquina, o uso de um sensor que gravasse informações de profundidade sem a necessidade de se calçar luvas ou vestir roupas foi uma escolha natural foi possível por estar disponível

o acesso ao sensor Kinect de forma barata e de fácil manipulação, no momento em que iniciávamos esta pesquisa. A escolha da solução de *hardware* e *software*, dentre tantas combinações possíveis, pautou-se pela procura do mais simples.

Com a base de sinais gravada com quatro tipos de informações visuais distintas, o próximo desafio foi eliminar a redundância de informações presente nos vídeos. De fato, como o fundo dos vídeos gravados é preto, a variação entre quadros de um mesmo vídeo está em grande parte no movimento dos braços. No caso de alguns sinais, esta movimentação é pequena, inclusive. A opção aqui seria entre uma amostragem linear, que não nos atenderia, uma vez que o sinal não é executado de forma linear entre começo, meio e fim. Portanto, partiu-se para a utilização de uma técnica de sumarização recém publicada. Nela, a redução do vídeo é tratada como um problema de diversidade máxima. A otimização deste problema por meio de um algoritmo evolucionário aplicado a vídeos mostrou ser uma solução promissora e atendeu ao objetivo aqui proposto.

Quanto ao núcleo deste trabalho, a extração de características, foi possível selecionar sete, dentre diversas características estudadas. Estas sete possuem o que procurou-se neste trabalho, que é uma relação com a estrutura fonológica de parâmetros da Língua Brasileira de Sinais. Elas contêm em seus valores informações sobre os parâmetros espaciais Ponto de articulação, Movimento e Orientação da palma da mão e sobre o parâmetro de forma Configuração da mão.

A validação das características extraídas, tendo-as como entrada em um sistema supervisionado de aprendizagem capaz de classificar cada um dos quatro parâmetros fonológicos linguísticos objetos deste trabalho, foi então realizada por dois experimentos. No primeiro, variou-se o *kernel* do sistema de classificação escolhido, a Máquina de Vetor de Suporte. No segundo experimento, a análise comparativa foi realizada sobre diferentes vetores de características que contemplavam a informação de tempo (ou não) em seus valores.

Finalmente, foi implementado um sistema de classificação por cálculo de distâncias entre as saídas do classificador de parâmetros e a saída desejada destes parâmetros para cada sinal. Para alguns sinais o resultado foi bastante promissor. Para outros o desempenho foi ruim. Se considerarmos que há um outro parâmetro fonológico não utilizado neste trabalho, as expressões não-manuais e que também contribuem para a construção de cada sinal, os resultados de classificação de sinais é promissor.

Por fim, tem-se que o reconhecimento de língua de sinais é um problema bastante complexo e desafiador, necessitando de integração entre técnicas de extração de características, de tecnologia e de métodos de classificação os mais diversos. Além do mais, os experimentos realizados mostram que extrair características e relacioná-

las com os parâmetros fonológicos da língua auxilia o reconhecimento de sinais. As seções anteriores mostraram que a eficácia em classificar características de parâmetros estruturais é alta, o que significa que a metodologia de extração de característica proposta é útil no reconhecimento dos parâmetros que compõem os sinais. Entretanto, a integração desta informação para o atual sistema de reconhecimento é ainda um desafio. Embora alguns sinais estejam corretamente reconhecidos com uma alta taxa, outros sinais ainda necessitam de informação adicional tal que seu reconhecimento correto seja alcançado. Uma das informações que são necessárias para a composição do sinal é o seu parâmetro expressões não-manuais.

Se um novo sinal for adicionado ao sistema de reconhecimento implementado, será necessário apenas acrescentar este sinal à base de sinais e sua descrição em termos de parâmetros estruturais como nas Figuras 10 e 9. A quantidade de atributos e valores é sempre fixa, portanto o número de classes no sistema de classificação não aumenta com a adição de novos sinais. Portanto, nossa abordagem é escalável e pode, em princípio, ser manejável dada que a Libras possui mais de 10.000 sinais. Espera-se que, à medida em que novos sinais sejam adicionados ao classificador de parâmetros, o desempenho melhore, pois ter-se-á mais amostras para o treinamento da Máquina de Vetor de Suporte.

7.3 Trabalhos futuros

Esta metodologia apresenta grande potencial em Reconhecimento de Sinais em Língua Brasileira de Sinais e também muitos desafios. A estrutura das línguas de sinais é muito complexa, como descrito em Quadros e Karnopp (2004). Embora a tenhamos simplificado neste trabalho, os valores que devem ser detectados por sistemas computacionais são bastante sutis. Mesmo especialistas em línguas de sinais têm dificuldade para definir o valor de cada atributo de cada parâmetro para um sinal.

Deve-se explorar mais detalhamente esta estrutura fonológica, junto a especialistas em línguas de sinais, buscando características que distinguam os sinais e que possam ser detectadas por sistemas computacionais. Com uma gama maior de sinais cujas características serão extraídas, técnicas de seleção destas podem ser implementadas para classificar quantitativamente o desempenho de cada uma delas em relação ao reconhecimento de parâmetros.

O uso do sensor RGB-D é bastante promissor, uma vez que melhorias têm sido incorporadas a estes. Uma destas melhorias impacta diretamente na extração de características, que é o aumento da resolução da imagem em RGB e de profundidade. O alinhamento entre os vídeos destas gravações também é um trabalho em desenvolvimento nos *softwares* que operam estes sensores.

Outras possíveis técnicas que podem ser utilizadas na integração dos parâmetros classificados com o objetivo de reconhecer sinais são as Cadeias Ocultas de *Markov*. Elas podem absorver as características dos vídeos cronologicamente.

Por fim, a metodologia, os experimentos e os resultados descritos de forma detalhada nesta tese foram publicados de forma resumida nos trabalhos em (ALMEIDA; GUIMARÃES; RAMÍREZ, 2013) e em (ALMEIDA; GUIMARÃES; RAMÍREZ, 2014).

Referências Bibliográficas

AGRIS, U. von; BLOMER, C.; KRAISS, K. F. Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, mllr, and map. In: IEEE. *19th International Conference on Pattern Recognition*. [S.l.], 2008. (ICPR 2008), p. 1–4. Citado na página 31.

ALMEIDA, S. G. M.; GUIMARÃES, F. G.; RAMÍREZ, J. A. A methodology for feature extraction in brazilian sign language recognition. In: *Proceedings of the IX Workshop de Visão Computacional*. [S.l.: s.n.], 2013. (WVC 2013, v. 1), p. 1–6. Citado na página 115.

ALMEIDA, S. G. M.; GUIMARÃES, F. G.; RAMÍREZ, J. A. Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors. *Expert Systems with Applications*, v. 41, n. 16, p. 7259–7271, 2014. Citado na página 115.

ARAN, O.; AKARUN, L. A multi-class classification strategy for fisher scores: Application to signer independent sign language recognition. *Pattern Recognition*, v. 43, n. 5, p. 1776–1788, 2010. Citado na página 31.

ARAN, O. et al. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition*, v. 42, n. 5, p. 812–822, 2009. Citado na página 30.

BATTISON, R. Phonological deletion in american sign language. *Sign Language Studies*, v. 5, n. 1, p. 1–19, 1974. Citado 2 vezes nas páginas 8 e 43.

BAY, H. et al. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, v. 110, n. 3, p. 346–359, 2008. Citado 2 vezes nas páginas 93 e 94.

BRASHEAR, H. et al. Using multiple sensors for mobile sign language recognition. In: *Proceedings of the 7th IEEE International Symposium on Wearable Computers*. [S.l.: s.n.], 2003. (ISWC 2003), p. 45–52. Citado na página 32.

BRUNELLI, R. Template matching techniques in computer vision: Theory and practice. In: *Template Matching Techniques in Computer Vision*. [S.l.]: John Wiley & Sons, Ltd, 2009. p. i–x. Citado na página 94.

CANNY, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 8, n. 6, p. 679–698, 1986. Citado na página 92.

CAPOVILLA, F. C. Principais achados e implicações do maior programa do mundo em avaliação do desenvolvimento de competências linguísticas de surdos. In: *Transtornos de aprendizagem: da avaliação à reabilitação.* [S.l.]: Artes Médicas, 2008. p. 151–163. Citado na página 27.

CAPOVILLA, F. C. et al. Brazilian sign language lexicography and technology: Dictionary, digital encyclopedia, chereme-based sign retrieval, and quadriplegic deaf communication systems. *Sign Language Studies*, Project MUSE, v. 4, n. 3, p. 393–430, 2003. Citado na página 26.

CAPOVILLA, F. C.; RAPHAEL, W. D.; MAURíCIO, A. C. L. *Novo Deit-Libras: Dicionário Enciclopédico Ilustrado Trilíngue da Língua Brasileira de Sinais (Libras) baseado em Linguística e Neurociências Cognitivas, Volume I: Sinais de A a H.* [S.l.]: Edusp, 2012. Citado 8 vezes nas páginas 8, 25, 26, 52, 53, 56, 64 e 112.

CAPOVILLA, F. C.; RAPHAEL, W. D.; MAURíCIO, A. C. L. *Novo Deit-Libras: Dicionário Enciclopédico Ilustrado Trilíngue da Língua Brasileira de Sinais (Libras) baseado em Linguística e Neurociências Cognitivas, Volume II: Sinais de I a Z.* [S.l.]: Edusp, 2012. Citado 8 vezes nas páginas 8, 25, 26, 52, 53, 56, 64 e 112.

CARIDAKIS, G. et al. Non parametric, self organizing, scalable modeling of spatiotemporal inputs: The sign language paradigm. *Neural Networks*, v. 36, p. 157–166, 2012. Citado na página 31.

CHAVEIRO, N. et al. Instruments in brazilian sign language for assessing the quality of life of the deaf population. *Revista Saúde Pública*, v. 47, p. 616–623, 2013. Citado na página 27.

CHUNLI, W.; WEN, G.; JIYONG, M. A real-time large vocabulary recognition system for chinese sign language. In: *Gesture and Sign Language in Human-Computer Interaction*. [S.l.]: Springer Berlin Heidelberg, 2002. v. 2298, p. 86–95. Citado na página 63.

CLEVE, J. V. V. Lexicography and the university: Making the gaulladet dictionary of american sign language. *Sign Language Studies*, v. 4, n. 3, p. 487–500, 2003. Citado 2 vezes nas páginas 27 e 64.

COOPER, H.; BOWDEN, R. Large lexicon detection of sign language. In: *Proceedings of the 2007 IEEE international conference on Human-computer interaction*. [S.l.: s.n.], 2007. (HCI'07), p. 88–97. Citado na página 63.

COOPER, H. et al. Sign language recognition using sub-units. *Journal of Machine Learning Research*, v. 13, n. 1, p. 2205–2231, 2012. Citado na página 30.

CRUZ, L.; LUCIO, D.; VELHO, L. Kinect and rgbd images: Challenges and applications. In: *Proceedings of the 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*. [S.l.: s.n.], 2012. (SIBGRAPI-T '12), p. 36–49. Citado na página 59.

CUI, Y.; WENG, J. Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, v. 78, n. 2, p. 157–176, 2000. Citado na página 31.

DAUBECHIES, I. The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on*, v. 36, n. 5, p. 961–1005, 1990. Citado na página 93.

DEMENTHON, D.; KOBLA, V.; DOERMANN, D. Video summarization by curve simplification. In: *Proceedings of the sixth ACM International Conference on Multimedia*. [S.l.: s.n.], 1998. (MULTIMEDIA '98), p. 211–218. Citado na página 68.

DIAS, J.; SOUZA, K. P. de; PISTORI, H. Conjunto de treinamento para algoritmos de reconhecimento de libras. In: *Proceedings of the II Workshop de Visão Computacional*. [S.l.: s.n.], 2006. (WVC 2006). Citado na página 63.

DIPIETRO, L.; SABATINI, A. M.; DARIO, P. A survey of glove-based systems and their applications. *Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, v. 38, n. 4, p. 461–482, 2008. Citado na página 30.

DOULAMIS, A. D.; DOULAMIS, N. D.; KOLLIAS, S. D. A fuzzy video content representation for video summarization and content-based retrieval. *Signal Processing*, v. 80, n. 6, p. 1049–1067, 2000. Citado na página 68.

DRYER, M. S.; HASPELMATH, M. *The World Atlas of Language Structures Online*. 2014. Disponível em: <http://wals.info/>. Citado 2 vezes nas páginas 8 e 23.

DUDUCHI, M.; CAPOVILLA, F. C. Buscasigno: a construção de uma interface computacional para o acesso ao léxico da língua de sinais brasileira. In: *Proceedings of VII Brazilian Symposium on Human Factors in Computing Systems*. [S.l.: s.n.], 2006. (IHC '06), p. 21–30. Citado na página 26.

EJAZ, N.; TARIQ, T. B.; BAIK, S. W. Adaptive key frame extraction for video summarization using an aggregation mechanism. *Journal of Visual Communication and Image Representation*, v. 23, n. 7, p. 1031–1040, 2012. Citado na página 68.

ELLOUZE, M.; BOUJEMAA, N.; ALIMI, A. M. Im(s)2: Interactive movie summarization system. *Journal of Visual Communication and Image Representation*, v. 21, n. 4, p. 283–294, 2010. Citado na página 68.

ESCALERA, S. et al. *Multi-modal Challenge*. 2013. Disponível em: <http://gesture.chalearn.org/>. Citado na página 63.

FELIPE, T. A. Os processos de formação de palavra na libras. *ETD - Educação Temática Digital*, v. 7, n. 6, p. 200–217, 2006. Citado na página 27.

FERREIRA-BRITO, L. Uma abordagem fonológica dos sinais da lscb. *Informativo Técnico-Científico do INES, Rio de Janeiro*, v. 1, n. 1, p. 20–43, 1990. Citado 3 vezes nas páginas 14, 46 e 47.

FERREIRA-BRITO, L. Sistema ferreira brito-langevin de transcrição de sinais. *Por uma gramática de língua de sinais*, 1995. Citado 7 vezes nas páginas 8, 14, 42, 43, 44, 45 e 51.

FLASIŃSKI, M.; MYŚLIŃSKI, S. On the use of graph parsing for recognition of isolated hand postures of polish sign language. *Pattern Recognition*, v. 43, n. 6, p. 2249–2264, 2010. Citado na página 31.

FRATI, V.; PRATTICHIZZO, D. Using kinect for hand tracking and rendering in wearable haptics. In: *World Haptics Conference (WHC), 2011 IEEE*. [S.l.: s.n.], 2011. p. 317–321. Citado na página 31.

FREITAS, A. R. R. et al. Memetic self-adaptive evolution strategies applied to the maximum diversity problem. *Optimization Letters*, p. 1–10, 2014. Citado na página 70.

FRIEDMAN, L. A. *The manifestation of subject, object, and topic in the American Sign Language*. [S.l.]: Academic Press, 1976. Citado na página 42.

GEETHA, M.; MANJUSHA, U. C. A vision based recognition of indian sign language alphabets and numerals using b-spline approximation. *International Journal on Computer Science & Engineering*, v. 4, n. 3, 2012. Citado na página 31.

GIANLUIGI, C.; RAIMONDO, S. An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing*, v. 1, p. 69–88, 2006. Citado na página 68.

GONG, Y.; LIU, X. Video summarization using singular value decomposition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2000. (CVPR 2000, v. 2), p. 174–180. Citado na página 67.

HAN, J.; AWAD, G.; SUTHERLAND, A. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognition Letters*, v. 30, p. 623–633, 2009. Citado na página 30.

HARRIS, C.; STEPHENS, M. A combined corner and edge detection. In: *Proceedings of The Fourth Alvey Vision Conference*. [S.l.: s.n.], 1988. p. 147–151. Citado na página 89.

HORN, B. K. P.; SCHUNCK, B. G. Determining optical flow. *Artificial Intelligence*, v. 17, n. 1-3, p. 185–203, 1981. Citado na página 85.

HUANG, C.-L.; HUANG, W.-Y. Sign language recognition using model-based tracking and a 3d hopfield neural network. *Machine Vision Applications*, v. 10, n. 5-6, p. 292–307, 1998. Citado na página 31.

ILLINGWORTH, J.; KITTLER, J. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, v. 44, n. 1, p. 87–116, 1988. Citado na página 90.

INCERTIS, I. G.; GARCIA-BERMEJO, J. G.; CASANOVA, E. Z. Hand gesture recognition for deaf people interfacing. In: *18th International Conference on Pattern Recognition.* [S.l.: s.n.], 2006. (ICPR 2006, v. 2), p. 100–103. Citado na página 31.

JOHNSTON, A. T. Language standardization and signed language dictionaries. *Sign Language Studies*, v. 4, n. 3, p. 431–468, 2003. Citado 2 vezes nas páginas 27 e 64.

KARAMI, A.; ZANJ, B.; SARKALEH, A. K. Persian sign language (psl) recognition using wavelet transform and neural networks. *Expert Systems with Applications*, v. 38, n. 3, p. 2661–2667, 2011. Citado na página 31.

KARMOKAR, B. C.; ALAM, K. R.; SIDDIQUEE, K. Bangladeshi sign language recognition employing neural network ensemble. *International Journal of Computer Applications*, v. 58, n. 16, p. 43–46, 2012. Citado na página 31.

KIM, J.; JANG, W.; BIEN, Z. A dynamic gesture recognition system for the korean sign language (ksl). *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, v. 26, p. 354–359, 1996. Citado na página 30.

KULKARNI, V. S.; LOKHANDE, S. D. Appearance based recognition of american sign language using gesture segmentation. *International Journal on Computer Science & Engineering*, v. 2, n. 3, 2010. Citado na página 31.

KUO, C.-C.; GLOVER, F.; DHIR, K. S. Analyzing and modeling the maximum diversity problem by zero-one programming*. *Decision Sciences*, v. 24, n. 6, p. 1171–1185, 1993. Citado na página 68.

LAHAMY, H.; LICHTI, D. D. Towards real-time and rotation-invariant american sign language alphabet recognition using a range camera. *Sensors*, v. 12, n. 11, 2012. Citado na página 31.

LEWIS, M. P. et al. *Ethnologue: Languages of the world*. [S.l.]: SIL international Dallas, TX, 2009. Citado 3 vezes nas páginas 14, 21 e 22.

LI, Y. Hand gesture recognition using kinect. In: *2012 IEEE 3rd International Conference on Software Engineering and Service Science*. [S.l.: s.n.], 2012. (ICSESS 2012), p. 196–199. Citado na página 31.

LI, Z.; SCHUSTER, G.; KATSAGGELOS, A. Minmax optimal video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 15, n. 10, p. 1245–1256, 2005. Citado na página 68.

LOEDING, B. L. et al. Progress in automated computer recognition of sign language. In: *Computers Helping People with Special Needs*. [S.l.]: Springer Berlin Heidelberg, 2004, (Lecture Notes in Computer Science, v. 3118). p. 1079–1087. Citado na página 28.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, v. 60, n. 2, p. 91–110, 2004. Citado na página 93.

MA, Y.-F. et al. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, v. 7, n. 5, p. 907–919, 2005. Citado na página 68.

MA, Y.-F. et al. A user attention model for video summarization. In: *Proceedings of the tenth ACM international conference on Multimedia*. [S.l.: s.n.], 2002. (MULTIMEDIA '02), p. 533–542. Citado na página 68.

MANKOFF, K. D.; RUSSO, T. A. The kinect: a low-cost, high-resolution, short-range 3d camera. *Earth Surface Processes and Landforms*, v. 38, n. 9, p. 926–936, 2013. Citado na página 59.

MARAQA, M. et al. Recognition of arabic sign language (arsl) using recurrent neural networks. *Journal of Intelligent Learning Systems and Applications*, v. 4, p. 41–52, 2012. Citado na página 30.

MATHWORKS. *Matlab R2012a (7.14.0.739)*. 2012. Citado 4 vezes nas páginas 10, 62, 65 e 84.

MENDI, E.; CLEMENTE, H. B.; BAYRAK, C. Sports video summarization based on motion analysis. *Computers & Electrical Engineering*, v. 39, n. 3, p. 790–796, 2013. Citado na página 67.

MERRIAM-WEBSTER. *An Encyclopaedia Britannica Company Merriam-Webster*. 2014. Disponível em: <http://www.-webster.com/dictionary/quadriplegic>. Citado na página 26.

MICROSOFT. *Microsoft Kinect for Windows*. 2013. Disponível em: <http://www.microsoft.com/en-us/kinectforwindows/>. Citado 2 vezes nas páginas 31 e 59.

NGO, C.-W.; MA, Y.-F.; ZHANG, H.-J. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 15, n. 2, p. 296–305, 2005. Citado na página 68.

ONG, E.-J. et al. Sign language recognition using sequential pattern trees. p. 2200–2207, 2012. Citado 2 vezes nas páginas 30 e 31.

ONG, S. C. W.; RANGANATH, S. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions Pattern Analysis and Machine Intelligence*, v. 27, n. 6, p. 873–891, 2005. Citado na página 28.

OPENCV. *OpenCV - Open Source Computer Vision*. 2014. Disponível em: <http://opencv.org/>. Citado na página 62.

OPENNI. *The OpenNI Project*. 2014. Disponível em: <http://www.openni.org/>. Citado na página 60.

OSZUST, M.; WYSOCKI, M. Polish sign language words recognition with kinect. In: *The 6th International Conference on Human System Interaction*. [S.l.: s.n.], 2013. (HSI 2013), p. 219–226. Citado na página 31.

OTINIANO, K. *Reconhecimento de Sinais Estáticos a partir de Informação RGB-D usando um Descritor Kernel*. Dissertação (Mestrado) — Universidade Federal de Ouro Preto, 2013. Citado na página 31.

PARTON, B. S. Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence. *Journal of Deaf Studies and Deaf Education*, v. 11, n. 1, p. 94–101, 2006. Citado na página 29.

PARVINI, F. et al. An approach to glove-based gesture recognition. In: *Proceedings of the 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques*. [S.l.: s.n.], 2009. (HCI 2009), p. 236–245. Citado na página 30.

PHADTARE, L.; KUSHALNAGAR, R.; CAHILL, N. Detecting hand-palm orientation and hand shapes for sign language gesture recognition using 3d images. In: *Western New York Image Processing Workshop*. [S.l.: s.n.], 2012. (WNYIPW 2012), p. 29–32. Citado na página 31.

QUADROS, R. M. Linguistic policies, linguistic planning, and brazilian sign language in brazil. *Sign Language Studies*, v. 12, n. 4, p. 543–564, 2012. Citado na página 27.

QUADROS, R. M. d.; KARNOPP, L. B. *Língua de sinais brasileira: estudos linguísticos.* [S.l.]: Artmed, 2004. Citado 12 vezes nas páginas 8, 14, 40, 41, 42, 43, 44, 46, 49, 51, 52 e 114.

REN, J.; JIANG, J.; FENG, Y. Activity-driven content adaptation for effective video summarization. *Journal of Visual Communication and Image Representation*, v. 21, n. 8, p. 930–938, 2010. Citado na página 68.

ROSTEN, E.; DRUMMOND, T. Machine learning for high-speed corner detection. In: *Proceedings of the 9th European conference on Computer Vision - Volume Part I*. [S.l.: s.n.], 2006. (ECCV'06), p. 430–443. Citado na página 89.

SACKS, O. *Vendo Vozes: uma viagem ao mundo dos surdos*. [S.l.]: Companhia das Letras, 1990. Citado 2 vezes nas páginas 21 e 28.

SCHEMER, G. M. From variant to standard: An overview of the standardization process of the lexicon of sign language of the netherlands. *Sign Language Studies*, v. 4, n. 3, p. 469–486, 2003. Citado 2 vezes nas páginas 27 e 64.

SHI, J.; TOMASI, C. Good features to track. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 1994. (CVPR 1994), p. 593–600. Citado na página 89.

SOLE, M. M.; TSOEU, M. S. Sign language recognition using the extreme learning machine. In: *Africon, 2011*. [S.l.: s.n.], 2011. p. 1–6. Citado na página 31.

STARNER, T.; PENTLAND, A. Real-time american sign language recognition from video using hidden markov models. In: *Proceedings of the International Symposium on Computer Vision*. [S.l.: s.n.], 1995. p. 265–270. Citado na página 30.

STARNER, T.; WEAVER, J.; PENTLAND, A. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, p. 1371–1375, 1998. Citado na página 32.

SURDOS, I. N. de Educação de. *Dicionário da Língua Brasileira de Sinais*. 2008. Disponível em: <http://www.acessobrasil.org.br/libras/?wmode=transparent>. Citado na página 52.

TRINDADE, P.; LOBO, J.; BARRETO, J. Hand gesture recognition using color and depth images enhanced with hand angular pose data. In: *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*. [S.l.: s.n.], 2012. (MFI 2012), p. 71–76. Citado na página 31.

VERDU, A. C. M. A. et al. Recognition of altered segments in brazilian sign language. *Psychology & Neuroscience*, v. 5, p. 239–245, 2012. Citado na página 28.

WANG, Q. et al. Viewpoint invariant sign language recognition. *Computer Vision and Image Understanding*, v. 108, n. 1-2, p. 87–97, 2007. Citado na página 30.

WATSON, R. *A Survey of Gesture Recognition Techniques.* [S.l.], 1993. Citado na página 28.

WILCOX, S. The multimedia dictionary of american sign language: Learning lessons about language, technology, and business. *Sign Language Studies*, v. 4, n. 3, p. 379–392, 2003. Citado na página 27.

YIN, P. et al. Learning the basic units in american sign language using discriminative segmental feature selection. In: *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2009. (ICASSP '09), p. 4757–4760. Citado na página 30.

ZAFRULLA, Z. et al. American sign language recognition with the kinect. In: *Proceedings of the 13th international conference on multimodal interfaces*. [S.l.: s.n.], 2011. (ICMI '11), p. 279–286. Citado na página 31.

ZAHEDI, M.; KEYSERS, D.; NEY, H. Appearance-based recognition of words in american sign language. In: *Pattern recognition and image analysis*. [S.l.: s.n.], 2005. p. 511–519. Citado na página 31.

ZAKI, M. M.; SHAHEEN, S. I. Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, v. 32, n. 4, p. 572–577, 2011. Citado na página 31.

ZHANG, X. et al. A framework for hand gesture recognition based on accelerometer and emg sensors. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, v. 41, n. 6, p. 1064–1076, 2011. Citado na página 30.

ZHANG, Z. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, v. 19, n. 2, p. 4–10, 2012. Citado na página 59.

ZIEREN, J.; KRAISS, K.-F. Robust person-independent visual sign language recognition. In: *Pattern recognition and image analysis*. [S.l.]: Springer, 2005. p. 520–528. Citado na página 31.

ANEXO A – Anexo 1

A.1 Periódico - *Experts Systems With Applications* - 2014

Expert Systems with Applications 41 (2014) 7259–7271

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors

Sílvia Grasiella Moreira Almeida [a,b], Frederico Gadelha Guimarães [c,*], Jaime Arturo Ramírez [c]

[a] Graduate Program in Electrical Engineering, Federal University of Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, Minas Gerais, Brazil
[b] Federal Institute of Minas Gerais, Rua Pandiá Calógeras 898, Ouro Preto, Minas Gerais, Brazil
[c] Department of Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Available online 28 May 2014

Keywords:
Brazilian Sign Language Recognition
RGB-D sensors
Feature extraction

ABSTRACT

In contrast to speech recognition, whose speech features have been extensively explored in the research literature, feature extraction in Sign Language Recognition (SLR) is still a very challenging problem. In this paper we present a methodology for feature extraction in Brazilian Sign Language (BSL, or LIBRAS in Portuguese) that explores the phonological structure of the language and relies on RGB-D sensor for obtaining intensity, position and depth data. From the RGB-D images we obtain seven vision-based features. Each feature is related to one, two or three structural elements in BSL. We investigate this relation between extracted features and structural elements based on shape, movement and position of the hands. Finally we employ Support Vector Machines (SVM) to classify signs based on these features and linguistic elements. The experiments show that the attributes of these elements can be successfully recognized in terms of the features obtained from the RGB-D images, with accuracy results individually above 80% on average. The proposed feature extraction methodology and the decomposition of the signs into their phonological structure is a promising method to help expert systems designed for SLR.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Languages are complex systems of communication that humans use to express themselves, to manipulate objects and ideas and to foster cooperation and social bonds. The estimated number of languages in the world amount to a few thousands, the majority of which are vocal languages, also termed oral or spoken languages. Nevertheless, natural languages can also be signed ones, which is the most natural modality for deaf people. Both vocal and sign languages are composed by a limited set of building blocks called phonemes; in sign languages a limited set of shapes, orientations, locations and movements of the hands are combined to make up the words or morphemes of the language. The grammatical rules of the language link those morphemes into phrases and units of discourse. Many countries in the world have an official sign language, some examples are American Sign Language (ASL), French Sign Language (FSL), Italian Sign Language (ISL), Polish Sign Language (PSL), Japanese Sign Language (JSL), among others. In Brazil, the official sign language is the Brazilian Sign Language (BSL),[1] which was developed in the 19th century as a combination of the ancient Brazilian Sign Language and the French Sign Language. BSL is even different from the Portuguese Sign Language, even though Brazil and Portugal have the same vocal language, the Portuguese. Nowadays, BSL is recognized as an official language in Brazil with an estimated number of 5 million speakers, which is greater than the number of speakers of many vocal languages in the world.

Sign languages are visual-space languages, sharing spatial and visual elements such as shape of the hands, location and orientation of the signs, and the movements of the hands and the body together to convey ideas. All sign languages are characterized by these building blocks, differing basically in variations of these spatial and visual elements. In general, sign languages share a common phonological structure, with five elements[2]: (i) articulation point; (ii) Configuration of the hands; (iii) type of movements of the hands; (iv) orientation and (v) facial and body expressions. Each sign in each language is composed by a combination of these building blocks. These elements represent an important aspect of the language and can be exploited in automatic expert systems aimed at Sign Language Recognition (SLR). In contrast to speech recognition, whose speech features have been extensively explored in the

* Corresponding author. Tel.: +55 31 3409 3419.
E-mail addresses: silvia.almeida@ifmg.edu.br (S.G. Moreira Almeida), frederico-guimaraes@ufmg.br (F.G. Guimarães), jramirez@ufmg.br (J. Arturo Ramírez).

[1] In Portuguese, BSL is known as LIBRAS.

[2] Complete phonological, semantics, grammar and spelling structures defined in BSL, can be found in Quadros and Karnopp (2004).

http://dx.doi.org/10.1016/j.eswa.2014.05.024
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

research literature, feature extraction in SLR is still a very challenging problem, still lacking effective features for recognition that are independent from signers. Feature extraction in SLR requires different methods and approaches than in speech recognition. Most of the research literature focuses on specialized techniques of computer vision for extracting features but it is possible to find other approaches relying on gloves, sensors on fingers and hands and, more recently, RGB-D sensors, which capture not only color information but also depth infrared images. In a wider context, research on SLR has a potential impact on human–computer interaction, since both modalities of natural languages (vocal and signed) could in principle be used as an interface to interact with computers.

In this paper we present a methodology for feature extraction in Brazilian Sign Language that explores the phonological structure of the language and relies on an RGB-D sensor for obtaining data. Our goal is to extract features that would allow, in future work, the creation of an expert system capable of automatically translating signs from BSL to vocal Portuguese. To avoid retraining the system to every new sign and to overcome the scalability challenge related to the high number of different signs in the language, we relate these features to the phonological structure of the BSL. The use of an RGB-D sensor is very convenient in this application, since it is a low cost sensor with relatively easy programming. Therefore, we propose herein an innovative methodology in which the computational system should be guided by the linguistic system in the feature extraction process.

From the RGB-D images we obtain seven vision-based features. Each feature is related to one, two or three structural elements in BSL. We investigate this relation between extracted features and structural elements based on shape, movement and position of the hands. We consider that exploiting this relation is an important contribution of our approach when compared to other work on SLR. Since each sign is composed by these elements, the main idea is that establishing these feature-structure relations will improve the pattern recognition system. Moreover, news signs can be easily included in the implemented recognition system using their structural elements description, instead of having to train the whole system for each specific sign (or word or morpheme). Currently, BSL has a universe of around 10,000 signs according to Capovilla, Raphael, and Maurício (2012a, 2012b), thus it would be impractical to develop an SLR system for recognizing each individual sign. In principle, the ideas advanced here could be applied to other systems in SLR elsewhere, just requiring adaptations to the specificities of the sign language of interest.

In our results, to illustrate the proposed methodology, we have selected a set of 34 signs from Capovilla et al. (2012a, 2012b) as shown in Fig. 1. These signs represent 34 distinct morphemes in BSL. They were not chosen at random, but with the advise of a BSL expert, in such a way that these signs represent a wide range of structural elements of the language, and therefore are very representative of the universe of signs in BSL.

These 34 signs are recorded in a database using RGB-D sensors. The following information about each sign is recorded: (i) intensity image; (ii) depth image; (iii) skeleton image and (iv) positions of the body. The authors will make this dataset available in public domain for other researchers. After feature extraction, we employ Support Vector Machines (SVM) for pattern recognition and classification. We used two different kernels for classification and results between 70% and 97% in accuracy were obtained for each relation "extracted feature/structural element in BSL".

The remainder of the paper is organized as follows: first, in Section 2, we present an overview of related work. In Sections 3–5, the structure of BSL, our methodology and our system classification, respectively, are described. Section 6 shows the experiments and results. Finally, in Section 7 we present our conclusions.

2. Related work

Sign Language Recognition (SLR) is an important part of the larger application field of Hand Gestures Recognition (HGR) in Human–Computer Interaction. Complete works describing applications and techniques in HGR can be seen in Watson and College (1993), Chakraborty, Sarawgi, Mehrotra, Agarwal, and Pradhan (2008), Suarez and Murphy (2012), Chen, Wei, and Ferryman (2013), Palacios, Sagüés, Montijano, and Llorente (2013), Zhang, Yang, and Tian (2013) and Ren, Yuan, Meng, and Zhang, 2013 over the last two decades. But when SLR is the subject, some specific issues are recurrent, regardless of the methodology used. Designing solutions that achieve good performance despite all the constraints and difficulties imposed by the complexity of the problem is the real challenge.

Issues related to sign composition, interaction, relationships between hands, different classes of signs, lexicon complexity and non-standard sign translation and other sign language paradigms are discussed in Bossard, Braffort, and Jardino (2003) and Caridakis, Karpouzis, Drosopoulos, and Kollias (2012). Also, in Bossard et al. (2003), authors propose an SLR system design with the following system architecture in three levels: (i) sign recognition, (ii) sign selection and (iii) detection of relationships between signs. Each level itself can involve a very complex system design.

In Parton (2006), the author presents an overview about how techniques in different areas of Artificial Intelligence deal with SLR and translations. Designing SLR systems is a multidisciplinary task that can involve areas such as robotics, virtual reality, computer vision, neural networks, Virtual Reality Modeling Language (VRML), three-dimensional (3D) animation, natural language processing and intelligent computer-aided design. In Loeding, Sarkar, Parashar, and Karshmer (2004) and Ong and Ranganath (2005) we can find discussions about the progress and future of automated systems for SLR. Most recently, authors in Futane, Dharaskar, and Thakare (2012), Ong, Cooper, Pugeault, and Bowden (2012) and Cooper, Ong, Pugeault, and Bowden (2012) present a comparative study of different approaches in Sign Language Recognition. Out of possible approaches described in their work, such as glove based techniques, vision-based techniques and analysis of drawing gestures, we believe that vision-based techniques are the most natural way of constructing a human–computer interface and tackling the related challenges, in this case (i) segmentation of moving hands, (ii) tracking and analysis of hand motion and (iii) recognition itself.

The first issue to address when designing systems for SLR is related to the interaction between the user doing the signs and the computational interface. This interaction can employ glove-based systems or vision-based systems. In the first case, extensive overviews about gloves equipment and sensors in many applications in HGR can be found in Dipietro, Sabatini, and Dario (2008) and Parvini et al. (2009). In these two papers, the authors discuss many applications with data gloves. In Starner and Pentland (1995), Starner, Weaver, and Pentland (1998), Ong et al. (2012) and Cooper et al. (2012), the authors use colored gloves as markers. In Brashear, Starner, Lukowicz, and Junker (2003) multiple sensors are used in a system for mobile sign recognition and in Zhang et al. (2011) multiple sensors are used in a framework for Sign Language Recognition. In both glove-based systems and vision-based systems, sensors and markers are used to provide information about velocity, direction, position, orientation and angles of each hand. Glove-based systems can provide better precision and efficacy in detecting and tracking hands. However, they might be more expensive and require additional equipment for the user.

On the other hand, vision-based systems can be much cheaper and more comfortable for users. Therefore, vision-based systems have received increased attention from researchers and

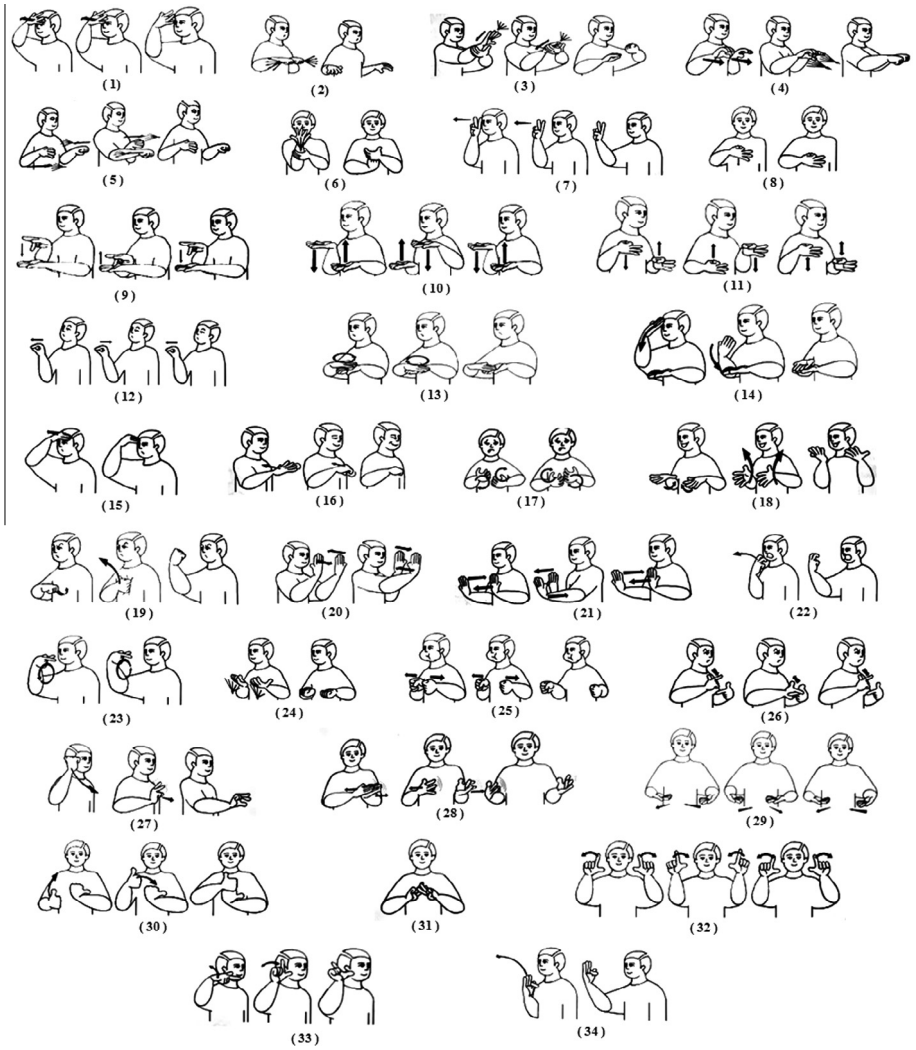


Fig. 1. The 34 signs used in our experiment. (1)"person", (2)"to spread", (3)"to copy", (4)"to catch", (5)"to gather", (6)"to disappear", (7)"to look", (8)"fair", (9)"truth", (10)"weight", (11)"justice", (12)"who", (13)"nothing", (14)"to believe", (15)"to forget", (16)"to love", (17)"to afflict", (18)"to commemorate", (19)"rancor", (20)"assembly meeting", (21)"to compare", (22)"to scream", (23)"to speak", (24)"to absorb", (25)"to fatten", (26)"to quarrel", (27)"perspicacious", (28)"to shine", (29)"maid", (30)"to replace", (31)"prison", (32)"television", (33)"yesterday", (34)"future".

developers. For these systems, general or special cameras capture images as data input. Papers describing system solutions using intensity (RGB), gray, and black and white images as their input can be found in Cui and Weng (2000), Zahedi, Keysers, and Ney (2005b), Zahedi, Keysers, Deselaers, and Ney (2005a), Haberdar and Albayrak (2005), Kawulok (2008), Dreuw, Stein, and Ney (2009), AL-Rousan, Assaleh, and Tala'a (2009) and Diraco, Leone, and Siciliano (2013). Most researchers in HGR are using RGB-D sensors as camera. With this sensor, we can obtain intensity (RGB), depth and skeleton (position) images, hence making it easier to detect and track hands. Authors in Liu and Fujimura (2004) provide a hand gesture recognition method using depth data obtained from a special camera. In recent papers, one of the most used RGB-D sensor in HGR and SLR in recent years is the Microsoft Kinect, mainly because of its low cost and the availability of a Developer Toolkit. More information about the sensor Kinect can be found in Cruz, Lucio, and Velho (2012) and Mankoff and Russo (2013). For HGR we can see the use of Kinect in the following work: (Chaaraoui, Padilla-López, Climent-Pérez, & Flórez-Revuelta, 2014; Chen et al., 2013; Dihl & Musse, in press; Frati & Prattichizzo, 2011; Li, 2012; Palacios et al., 2013; Ramey, González-Pacheco, & Salichs, 2011; Ramirez-Giraldo, Molina-Giraldo, Alvarez-Meza, Daza-Santacoloma, & Castellanos-Dominguez, 2012; Suarez & Murphy, 2012). In the specific case of SLR, Kinect has been used in the papers presented by Zafrulla, Brashear, Starner, Hamilton, and Presti (2011), Uebersax, Gall, Van den Bergh, and Van Gool (2011), Zaki and Shaheen (2011), Phadtare, Kushalnagar, and Cahill (2012), Boulares and Jemni (2012) and Oszust and Wysocki (2013). Authors in Trindade, Lobo, and Barreto (2012) propose a system design based on glove data and vision data together, joining both concepts.

Efforts to create public databases with signs in sign language have been done by some researchers. For instance, (Duduchi & Capovilla, 2006) shows an interface and a dataset for Brazilian Sign Language Recognition. Databases for American Sign Language Recognition are described in Yang, Sarkar, Loeding, and Karshmer (2006), Cooper and Bowden (2007) and Cooper, Pugeault, and Bowden (2011).

In terms of the pattern recognition and classification step of these systems, different methods have been used. In general, architectures based on Neural Networks (NN) are used, as in Huang and Huang (1998), Karami, Zanj, and Sarkaleh (2011), Sole and Tsoeu (2011), Karmokar, Alam, and Siddiquee (2012), Maraqa, Al-Zboun, Dhyabat, and Zitar (2012) and Ahmed (2012). SVM have also been used by Yang and Lee (2013). On the other hand, Hidden

Markov Models (HMM) have been used in many works for classification of the hands, as in Haberdar and Albayrak (2005), Haberdar and Albayrak (2006), Wang, Chen, Zhang, Wang, and Gao (2007), Yin, Starner, Hamilton, Essa, and Rehg (2009), AL-Rousan et al. (2009), Zafrulla et al. (2011), Zaki and Shaheen (2011) and Auephanwiriyakul, Phitakwinai, Suttapak, Chanda, and Theera-Umpon (2013). Authors in Aran, Burger, Caplier, and Akarun (2009) use HMM to detect manual and non-manual signs.

In this work we employ computer vision techniques and RGB-D sensor to segment and detect the movements of the hands. We extract features and relate them to the elements of the phonological structure of the language. The classifier system is trained to learn and detect this relation and these elements, in order to identify the sign (morpheme). In this way, scalability is straightforward since adding new signs to the system requires their description in terms of their structural elements, without needing to retrain and adjust the whole system anew.

3. Methodology for feature extraction

In this section, we describe each step in our methodology for feature extraction in SLR. First, we present, in Section 3.1, the phonological structure of BSL based on the work by Quadros and Karnopp (2004). Next, in Section 3.2, we discuss a video summarization technique employed in the methodology. In Section 3.3, we describe the database structure, and the hardware and software tools used to implement this approach and, in Section 3.4, we describe the method used to detect the region of interest (ROI) using intensity, depth and position images obtained by the RGB-D sensor.

3.1. Phonological structure in Brazilian Sign Language

We explore the phonological structure of BSL in our approach. For more details about BSL structure, see (Quadros & Karnopp, 2004). Four elements make up each sign in BSL: (i) configuration of the hands, (ii) articulation points, (iii) type of movement of the hands and (iv) orientation. However, the classification provided in Quadros and Karnopp (2004) is very extensive and detailed, and classifying each sign or morpheme exactly into values of these elements is a very complex task even for experts in BSL, let alone for a computer system. Therefore, we have to adopt a simplified set of these attributes and elements, which are summarized in Table 1. We describe each element below:

1. Configuration of the hands: This element of the phonological structure provides information about the shape of the hands, but there is no consensus about the amount of possible configurations available in Brazil for the BSL. We use as reference a study conducted in Quadros, Oliveira, and Miranda (2007), where the authors define 134 possible shapes for the hands and group them in 13 groups based on similarity. When treating this element, we have the following attributes and values: (i) group, from 1 to 13; (ii) axis alignment, either x, y, or z; (ii) variation of the configuration during the execution of the sign, yes or No.
2. Articulation points: it provides information about the location of the sign in the neutral space. The neutral space is the area ahead the body where the signs are done. We divide the neutral space into three areas, with one attribute corresponding to the (i) head, (ii) shoulder and (iii) body. The possible values for each attribute are either right, center or left.
3. Type of movement of the hands: provides information about the type of movement for the sign, in case the sign is dynamic. We adopt the following attributes for this element: (i) type, either up, down, right, left, inside or outside; (ii) frequency, either simple or repeated.
4. Orientation of the hand: provide information about the orientation of the palm of the hand in a sign. The following attributes and values are considered: (i) orientation, either up, down, inside, outside, or to the side; and (ii) change of orientation, yes or No.

Table 1
Elements of the phonological structure of BSL used for Sign Language Recognition. Each element consists of a set of attributes with possible values shown in the rightmost column.

Element	Attributes	Values
Configuration of the hands	group	$\{1,\ldots,13\}$
	axis alignment	$\{x,y,z\}$
	variation	$\{yes,no\}$
Articulation points	head	$\{right,center,left\}$
	shoulder	$\{right,center,left\}$
	body	$\{right,center,left\}$
Type of movement	type	$\{up,down,right,left,inside,outside\}$
	frequency	$\{simple,repeated\}$
Orientation	orientation	$\{up,down,inside,outside,totheside\}$
	variation	$\{yes,no\}$

3.2. Video summarization

Video summarization is an important topic in video classification and retrieval, where large videos should be compacted into more representative frames. In our case, we employ video summarization to reduce the number of frames in each sign, in order to avoid the processing of redundant frames that might be present in the recorded sign. Basic techniques for video summarization employ clustering techniques to group similar frames with respect to a given similarity metric.

We formulate the video summarization problem as a classic optimization problem known as the Maximum Diversity Problem (MDP), described in Freitas, Guimaraes, Pedrosa Silva, and Souza (2014). The MDP consists in finding a subset $M(|M|=m)$ from a set $N(|N|=n)$ of elements in such a way that the diversity of the m elements is maximized. Many relations of diversity can be used to define divergence values d_{ij} according to the practical application of the MDP. The problem is concisely described in Kuo, Glover, and Dhir (1993) and repeated in (1), where $x_i=1$ means that element i is in the subset M:

$$\max \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}x_i x_j,\ \ subject\ to \sum_{i=1}^{n} x_i = m \tag{1}$$

where $x_i \in \{0,1\}\ \forall\ i=1,\ldots,n$

The values d_{ij} in the diversity matrix $[\mathbf{D}]$ are calculated by a problem-specific metric, thus we have to define a metric of diversity between frames in videos. In this case, we can consider two aspects: temporal distance and color differences. In a temporal space, we are looking for the most distant frames in time. In a color space, we want to maximize diversity due to changes in color. The proposed diversity matrix is given by

$$[\mathbf{D}] = [\mathbf{C}] + [\mathbf{T}] \tag{2}$$

where $[\mathbf{C}]$ is a matrix with the color difference between frames i and j and $[\mathbf{T}]$ is a matrix with the time label of each frame.

The MDP is an NP-hard problem that can be solved with either exact methods or heuristics. In order to get a fast and efficient solution to the problem, we employ the Memetic Self-Adaptive Evolution Strategies (MSES) described in Freitas et al. (2014). It is an

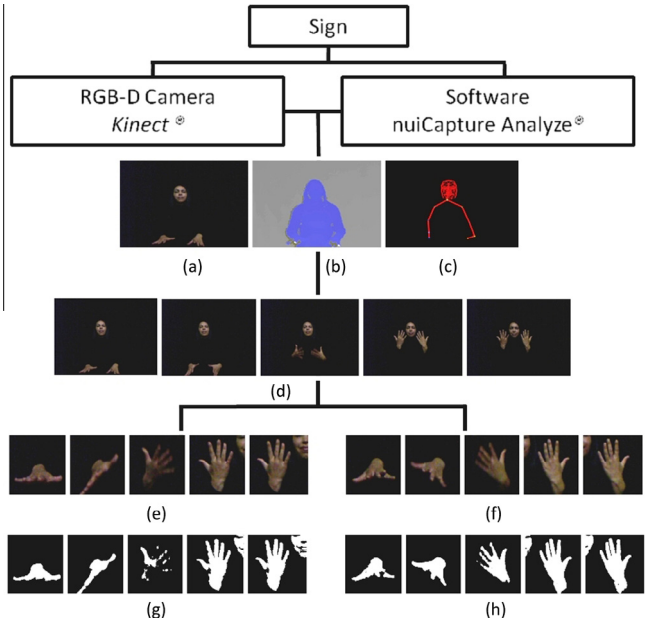


Fig. 2. Detection of region of interest – right and left hands for the sign "to commemorate". Video recorded using software nuiCaptureAnalyze working with Kinect sensor: (a) Intensity, (b) Depth and (c) Position. (d) Videos selected by video summarization using Maximum Diversity Problem to solve it. (e), (f) Region of interest in RGB detected to right and left hands, respectively. (g), (h) Region of interest detected in black and white using skin color detection.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

evolutionary algorithm with some local search heuristics designed for the MDP.

In this paper, we select a fixed number of representative frames for each sign. In Fig. 2(d), we show the $m = 5$ selected frames for the sign "Comemorar"[3] as an example.

3.3. Database and tools

The availability of databases in HGR and SLR for use by the scientific community in computer vision research is still a challenge. Some efforts are presented by Chunli, Wen, and Jiyong (2002), Dias, de Souza, and Pistori (2006) and Cooper and Bowden (2007), regarding data for SLR. Some researchers identify this need for standardized data, promoting collective actions to build large data in HGR using RGB-D sensors, see for instance (Escalera, González, Guyon, & Moeslund, 2013). The first difficulty, in the case of SLR, is reaching a consensus on which signs are representative of a given sign language and if a set of selected signs from a given sign language is useful and representative for testing and evaluating systems designed for other sign languages. Other issues are related to the resistance of the deaf community itself and regional variations as discussed in Schemer (2003), Johnston (2003) and Van Cleve (2003).

Given these difficulties, for the present work, we decided to create our own database, selecting 34 signs in BSL according to morphemes described in Capovilla et al. (2012a, 2012b), each sign being representative of many other signs due to similar characteristics. Given that there is no database available with this subset of signs, we decided to build our own database with five samples of each sign and for only one signer. We therefore focus on the variation in signs not on variations of the signer.

Signs are recorded using the Kinect sensor and the *nuiCapture Analyze* software. Kinect sensor is a low cost RGB-D camera developed by Microsoft for the XBOX video game. The integrated hardware and software in Kinect allows the detection and recording of twenty specific points of the human body. These properties can be found in Cruz et al. (2012), Microsoft (2013) and Mankoff and Russo (2013).

For the generation of the database, two requirements were given: (i) The video must contain the sign space, which is from above the waist. Recording legs is not necessary, since they do not participate in the execution of signs. (ii) The distance should be enough to capture the movements of the arms. Based on these requirements, the distance was set as 1.9 m (6.2 feet) from the Kinect sensor. This value is within the recording range of the sensor, which is between 0.8 m (2.6 feet) and 3.5 m (11 feet).

Using these tools, we obtain three videos recorded at the same time for each sample sign. We can see one frame of one sign in Fig. 2: (a) color frame from the intensity video, (b) depth frame from the depth video, (c) skeleton frame from the skeleton video. The video frame rate is 30 frames per second and all the three videos are recorded in AVI format (audio and video format from Microsoft).

3.4. Detection of the ROI

Each sign has two Regions of Interest (ROI), one for each hand. In this section we describe how we extract the ROI from the video captured by the Kinect sensor and the *nuiCapture Analyze* software.

We first remove the background by using the depth information from the depth video, see Fig. 2(b). With the skeleton video, Fig. 2(c), it is possible to find the coordinates of the hands. In this skeleton video, right hand and wrist are always recorded with blue color, while left hand and wrist are recorded with orange color. With the information about the position of the hands detected in the skeleton video, we can make the segmentation of the hands in the RGB video, already without the background.

In this way, we can extract the ROI for both hands as shown in Fig. 2(e) and (f). Finally, we detect face and hands using an algorithm for skin color detection and convert these images to black and white format, see Fig. 2(g) and (h).

The detection and extraction of the ROI represent the pre-processing part of the methodology for feature extraction. Next, we describe the techniques employed to actually obtain the features and how they relate to the elements of the phonological structure.

4. Relating features to elements in BSL

We extract seven features related to the phonological structure of BSL as described in Section 3.1, which are summarized in Table 2. These features are described in the following next subsections. It is important to highlight that all the features are extracted after the summarization and for each representative frame identified in the summarization step.

4.1. Two-dimensional distance

For obtaining this feature, we calculate the two-dimensional Euclidean distance, in pixels, between both hands and shoulder center. We select the shoulder as reference because the change in its position is negligible when the signer is performing the sign. The positions of the hands and the shoulder are given by the average positions of the pixels in the block of pixels that compose these parts extracted from the skeleton image. We can see this skeleton image in Fig. 3(a).

Individual colors are assigned to hands and shoulder positions among twenty possible body parts detected by using the RGB-D

[3] *tr. v.* to commemorate.

Table 2
Extracted features from the region of interest are shown in the left column. Structural elements in BSL associated with these features are shown in the right column.

Sec.	Analyzed features	Elements in BSL
4.1	Distance between pixels	Articulation points
4.2	Distance in millimeters	Articulation points
4.3	Velocity	Type of movement, Orientation
4.4	Area of hands	Configuration of the hands
4.5	Corners average	Articulation points, Type of movement, Orientation
4.6	Detected lines	Configuration of the hands
4.7	SURF descriptors	Type of movement, Orientation

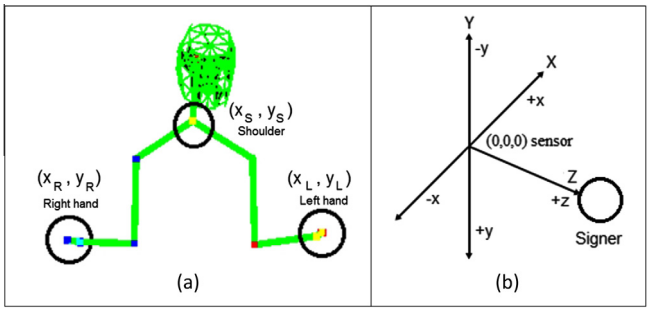


Fig. 3. (a) Feature 1 extracted: two-dimensional Euclidean distance between hands position and center of the shoulder. (b) Reference for feature 2, three-dimensional distance in millimeters.

sensor. In our case, the blue region represents the right hand, the orange region represents the left hand and the yellow color is for the shoulder center.

This feature provides information related to the articulation point of the sign in BSL.

4.2. Three-dimensional distance

Kinect records, using *nuiCapture Analyze*, the three-dimensional distance between sensor and signer in Matlab format (MathWorks, 2012). In Fig. 3(b) we can see that the origin of reference for these values is the sensor.

Three-dimensional distance provides information about the articulation point of the sign in BSL.

4.3. Velocity

We use the optical flow technique to calculate velocity for each sign. For more details about optical flow, see (Horn & Schunck, 1981). This feature provides vectors calculated using brightness difference between frames in a video. Thus, based on these vectors, we obtain information about hands tracking. In Fig. 4(a) we can see vectors calculated using this technique. This feature is related to the following structural elements in BSL: type of movement and orientation of the hands.

4.4. Area of hands

For this feature, we also use the optical flow technique, which allows to segment objects in an image based on image brightness discontinuity. In our region of interest, hands and, in some cases, the face, are the biggest objects. In Fig. 4(b) we can see these areas, all of them detecting hands. We observe, in frame 2, three detected objects: two hands and a face. This occurs because in the sign used as example, "to whip", the hands cross in front of the face. This characteristic can be preserved since this crossing is peculiar for this sign. Hands area is directly related to the Configuration of the hands in BSL.

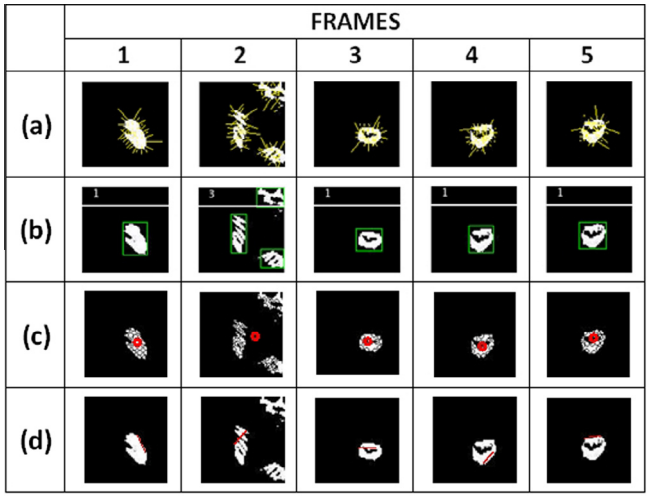


Fig. 4. Features extracted from right hand in the sign "to whip" in BSL: (a) Velocity, (b) Hands area, (c) Corners average and (d) Detected lines.

4.5. Corners average position

The area of the image where there is a large difference between pixels is named corner. Techniques to perform corner detection can be found in Harris and Stephens (1988), Shi and Tomasi (1994) and Rosten and Drummond (2006). In our approach, corners are detected using the Harris corner detector and an average of positions of these corners is the extracted feature. For more details about Harris corner detection technique, see (Harris & Stephens, 1988). This feature is related to the articulation points, the type of movement and orientation in BSL. An example of this feature is available in Fig. 4(c).

4.6. Detected lines

The sixth feature used in this work is given by the lines detected in each image. The technique used here is the Hough Transform, see (Illingworth & Kittler, 1988). This feature is an ordered pair (ρ, θ), because we use polar coordinates. In Fig. 4(d) we show the biggest lines extracted from the black and white image for each frame. These lines can contain information about the configuration of the hands in BSL.

4.7. Amount of common points between frames

The descriptor algorithm SURF (Speed-Up Robust Features) available in Matlab, (MathWorks, 2012), is used to extracted the amount of common points between frames. We can use SURF to find common points between representative frames of the sign using the template matching technique. At each two frames, SURF shows a different amount of points between them, showing the difference of movement between frames. This feature is related to the type of movement and orientation in BSL.

4.8. Assembling the feature vector

By relating each one of the seven features to each of the four structural elements in BSL, we can assemble the feature vector for each element. The number of features in each feature vector depends on the element under consideration. According to Table 2, we have the following feature vectors:

Table 3
Structure of feature vector for M features and n frames. In this study, we analyzed $M = 1, 2, 3$ features, depending on the element, and $n = 5$ frames.

Hand (right or left)			
1	...	n	Frames
1 a M	1 a M	1 a M	Features

Table 4
Structure of feature vector for M features and n frames with variable time added.

Hand (right or left)				
1	...	n		Frames
1 to M	1 to M	1 to M	t_1 to t_n	Features

Table 5
Structure of feature vector for M features and n frames with variable time multiplied.

Hand (right or left)			
1	...	n	Frames
1.t_1 to $M.t_M$	1.t_1 to $M.t_M$	1.t_1 to $M.t_M$	Features

- To the recognition of the element Articulation points, the following features are used (i) distance between pixels, (ii) distance in millimeters and (iii) corners average.
- To the recognition of the element Configuration of the hands, (i) area of hands and (ii) detected lines are used.
- To the element Type of movement we utilize (i) velocity, (ii) SURF descriptors and (iii) corners average to assemble the corresponding feature vector.
- Finally to the element Orientation, the feature vector is composed by the features (i) velocity and (ii) corners average.

The structure of the feature vector is shown in Table 3. Features (1 to M) are extracted from the first representative frame in the video ($n = 1$) and, according to the element under consideration, their values are appended sequentially in the vector. This is done for the n frames in sequence. We end up with a feature vector for each hand and for each one of the four structural elements, which are the input for the classification system presented in the next section.

The time variable is an important information to be considered in the recognition of dynamic signs. In our work, we add this information in two distinct ways in the feature vector. In the experiments, we investigate and compare the performance of the system when using

1. only the feature vector without the time information as in Table 3;
2. the feature vector with the time values of each frame added to it, see Table 4;
3. the feature vector with the time values of each frame multiplied by the values of the features, see Table 5.

5. Classification system

Fig. 5 shows a schematic of the implemented classification system. First, in (a) *Features*, we have as input the features extracted as described in Sections 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7. In (b) *Vector (n frames)*, a feature vector is assembled for each element in the phonological structure of the language, according to Table 2 and as described in Section 4.8. These vectors are the inputs of our classification system, in area (c). Our system classification is in fact implemented by Support Vector Machines (SVM) with two different kernels. This SVM identifies and classifies the linguistic elements based on the extracted features.

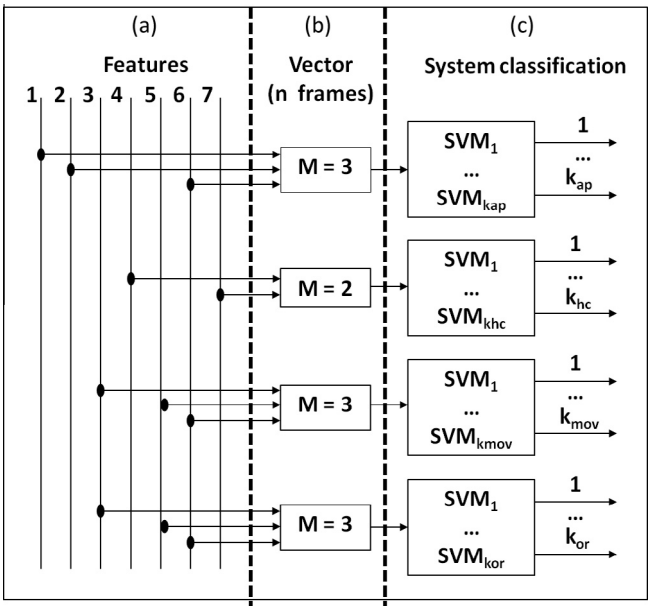


Fig. 5. System schematic. (a) Features: 7 extracted features. (b) Vector: see Table 3 for structure, $n = 5$ selected frames and M = numbers of features for a given element, see Table 2 for reference. (c) System classification: k_{ap} outputs for the articulation point ($k_{ap} = 5$ for right hand and $k_{ap} = 4$ for left hand); k_{hc} outputs for the configuration of the hands ($k_{hc} = 12$ for right hand and $k_{hc} = 10$ for left hand); k_{mov} outputs for the type of movement ($k_{mov} = 8$ for right and left hands); k_{or} outputs for the orientation ($k_{or} = 7$ for both right and left hands).

We compare results using SVM with two different kernels. First, we use a linear kernel, meaning dot product, as in Eq. (3). Second, we implement a radial basis function kernel as in Eq. (4).

$$f(x_i, x_j) = \sum_{i=1}^{S}\sum_{j=1}^{S} x_i \cdot x_j \tag{3}$$

$$f(x_i, x_j) = \sum_{i=1}^{S}\sum_{j=1}^{S} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \tag{4}$$

where x_i and x_j are feature vectors of signs in the database.

In Fig. 6 and 7 we show the attributes for system classification for each one of the 34 signs selected in this work.

6. Experiments and results

To evaluate our approach, we created a database with 34 specific signs. These signs were selected by a structural division done in Capovilla et al. (2012a), where authors describe 10.236 signs in BSL. Among these, 1.577 signs are mapped and 982 signs are composed by one or more than one of the 34 molecular morphemes chosen. Our database is composed by one example of each one of these 34 morphemes established. According to Capovilla et al. (2012a), it is possible to build other complex signs using these morphemes. In Fig. 1 we show the signs selected.

Remember that the database has five samples of each one of the 34 morphemes, making up a total of 170 examples in the database. We have performed 100 executions for each experiment, varying randomly the three samples of each sign used for training and the two samples used for testing the system, and we show averages and standard deviations for them, using statistical analysis

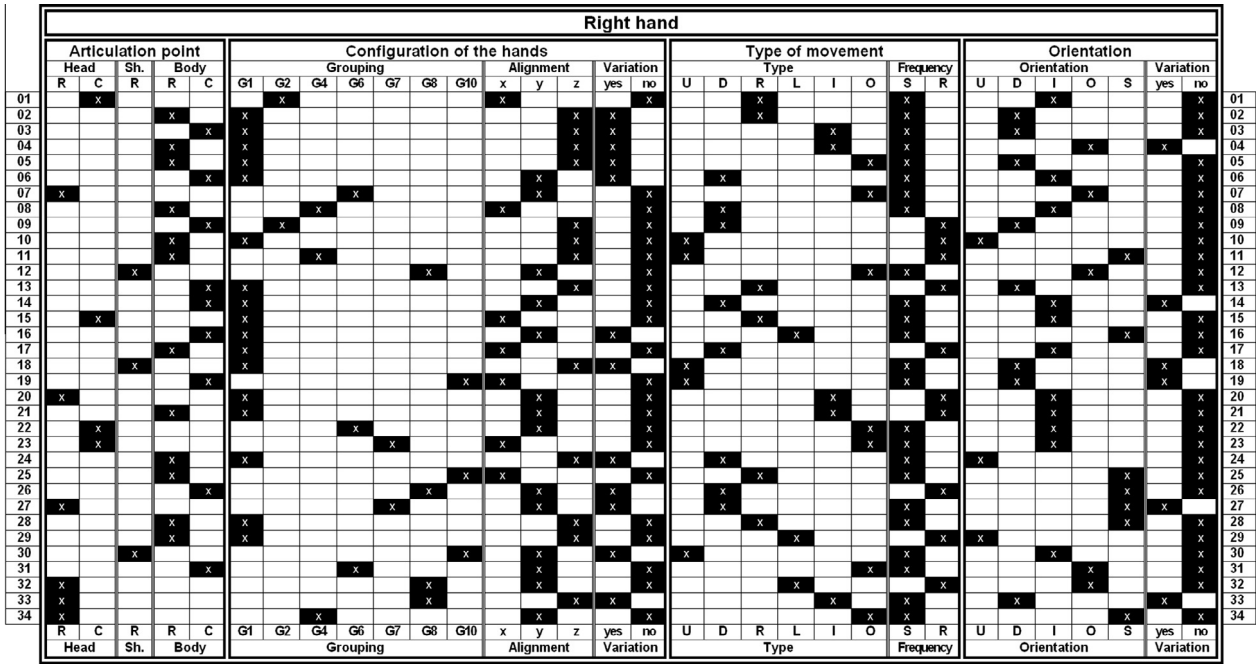


Fig. 6. Attributes for right hand *versus* 34 signs. *Articulation points*: 1. At right head. 2. At center head. 3. Sh.: shoulder. At right shoulder. 4. At right body. 5. At center body. *Configuration of the hands:* 1. G1. 2. G2. 3. G4. 4. G6. 5. G7. 6. G8. 7. G10. 8. Align in axis x. 9. Align in axis y. 10. Align in axis z. 11. Change in Configuration of the hands or alignment: yes. 12. Change in Configuration of the hands or alignment: No. *Type of movement:* 1. Up. 2. Down. 3. Right. 4. Left. 5. Inside. 6. Outside. 7. Frequency: simple. 8. Frequency: repeated. *Orientation:*1. Up. 2. Down. 3. Inside. 4. Outside. 5. To the side. 6. Variation: yes. 7. Variation: No.

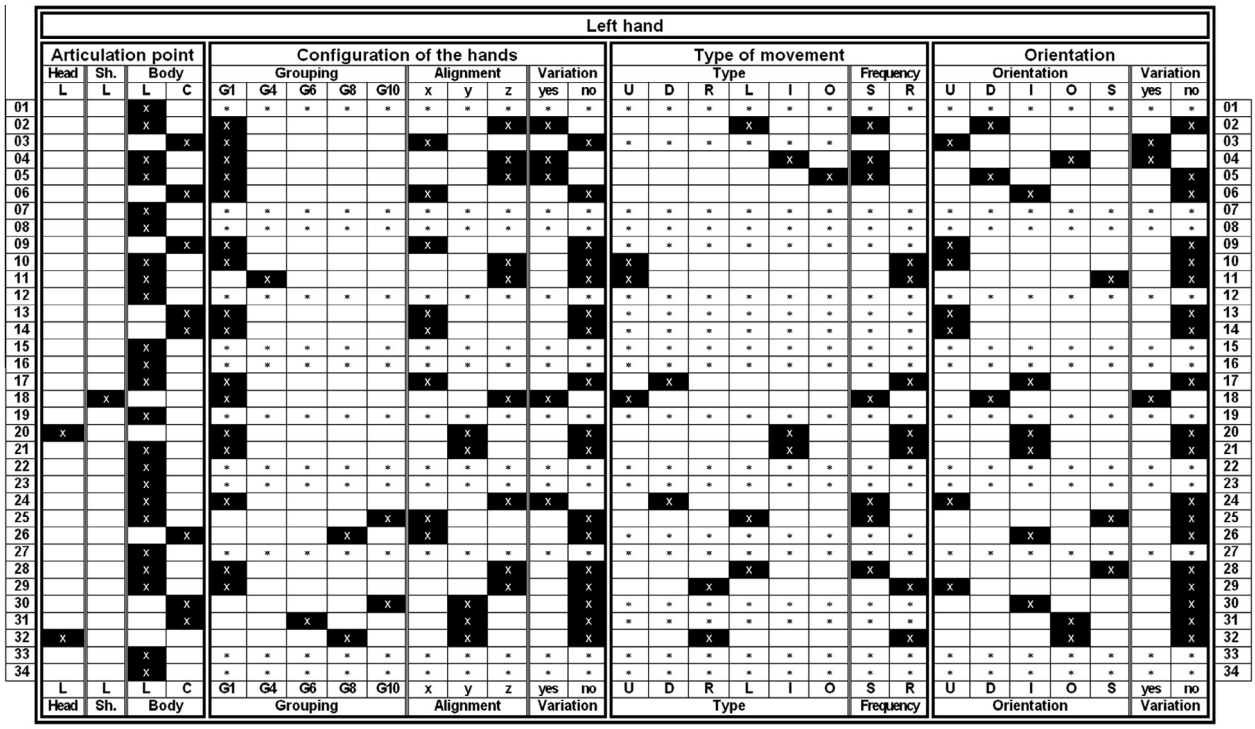


Fig. 7. Attributes for left hand *versus* 34 signs. *Articulation points*: 1. At left head. 2. Sh.: shoulder. At left shoulder. 3. At left body. 4. At center body. *Configuration of the hands:* 1. G1. 2. G4. 3. G6. 4. G8. 5. G10. 6. Align in axis x. 7. Align in axis y. 8. Align in axis z. 9. Change in Configuration of the hands or alignment: yes. 10. Change in Configuration of the hands or alignment: No. *Type of movement:* 1. Up. 2. Down. 3. Right. 4. Left. 5. Inside. 6. Outside. 7. Frequency: simple. 8. Frequency: repeated. *Orientation:*1. Up. 2. Down. 3. Inside. 4. Outside. 5. To the side. 6. Variation: yes. 7. Variation: No.

ANOVA. We present results in box plot graphics. In the 100 executions, the sources of randomness are the training and test data and the kernel parameters of the SVM in the classification system.

First, in Section 6.1, we compare two different kernels in terms of the classification results for each element in BSL for both hands. Next, in Section 6.2, we present the results of the comparison

between the feature vector with and without the time information, as explained in Tables 3–5.

6.1. Comparison between linear and RBF kernels

In Figs. 8 and 9, we present the classification results over the test data for the right and left hand respectively, using the feature vector only, as in Table 3.

We can see that, for the right hand, the average accuracy is higher for the RBF kernel in the elements Articulation points, Type of movement and Orientation. For the Configuration of the hands, the average accuracy is very similar. Also, for this element, the variation in accuracy is higher, showing the difficulty of correctly classifying the attributes of this element.

For the left hand, the RBF kernel again presents better results, specifically in the elements Configuration of the hands and Orientation. The linear kernel was better for the elements Articulation points and Type of movement. However, many signs are performed only with the right hand, with no participation of the left hand.

Based on these experiments, we selected the RBF kernel for the classification step.

6.2. Comparison of feature vectors

Using only the RBF kernel, we compared the feature vectors described in Tables 3–5,, to see the influence of the time information in the performance of the system.

For the right hand, the feature vector with the time information appended had a slightly superior accuracy, specially for the element Configuration of the hands. For the other elements, there were no statistical difference between adding time or not, and how we add

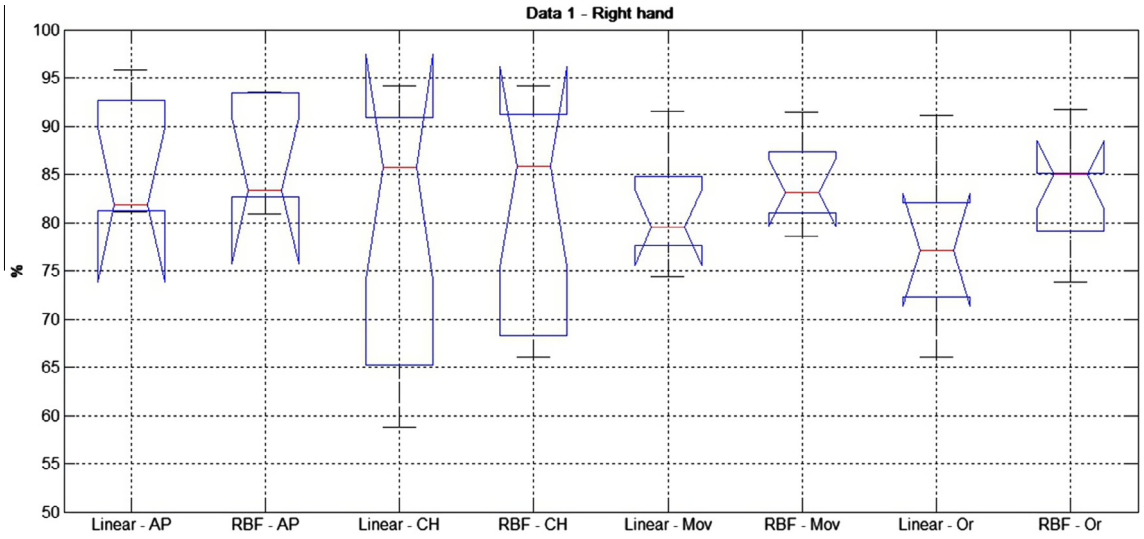


Fig. 8. Comparison between the linear and RBF kernels in the classification of the structural elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the right hand. Results obtained with the feature vector described in Table 3.

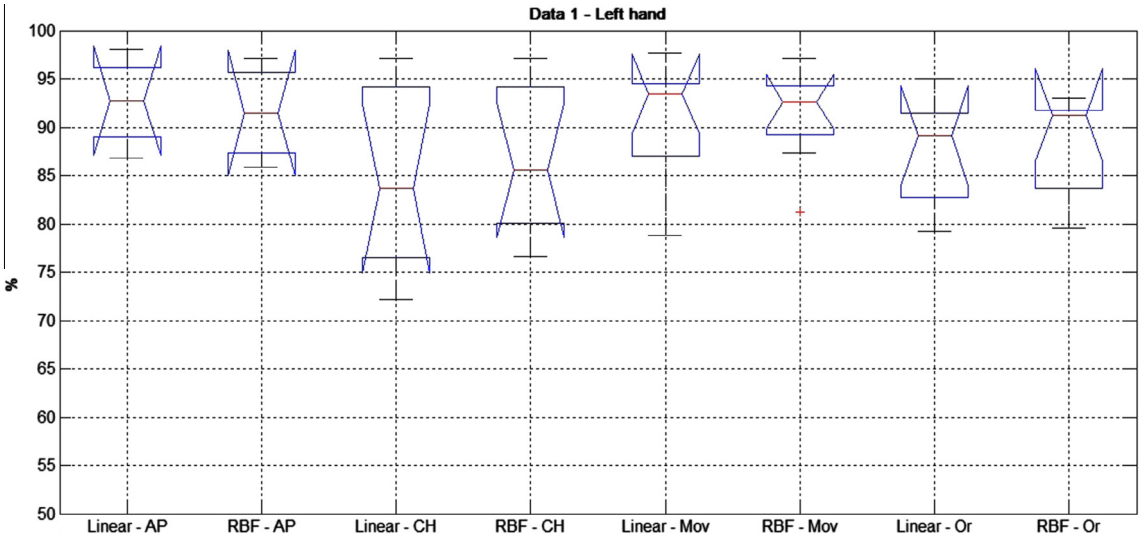


Fig. 9. Comparison between the linear and RBF kernels in the classification of the structural elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the left hand. Results obtained with the feature vector described in Table 3.

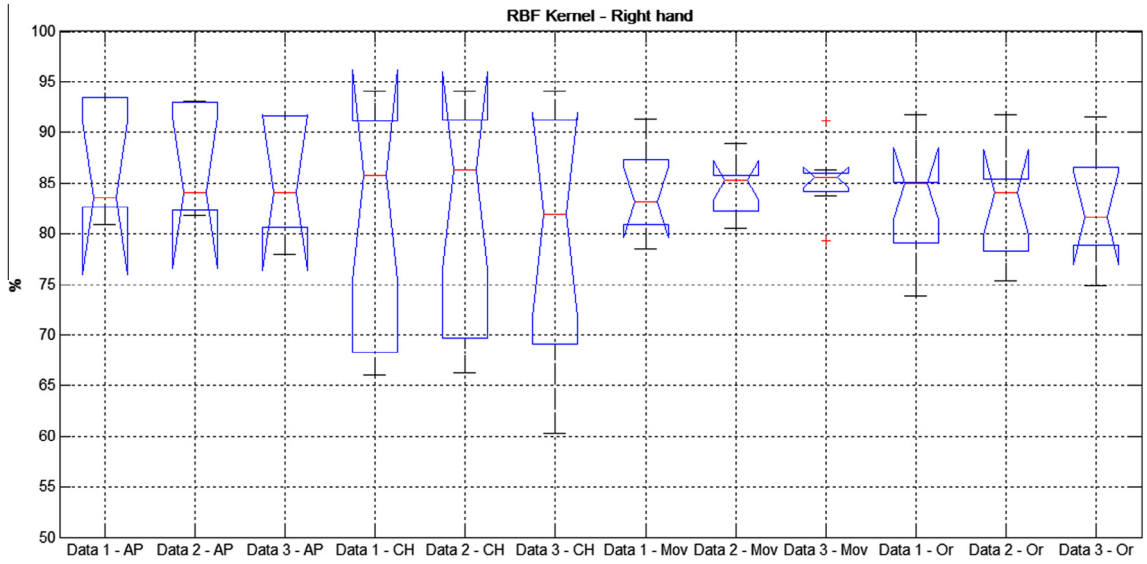


Fig. 10. Accuracy obtained by using the feature vectors described in Tables 3–5 for the RBF kernel in the classification of the elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the right hand.

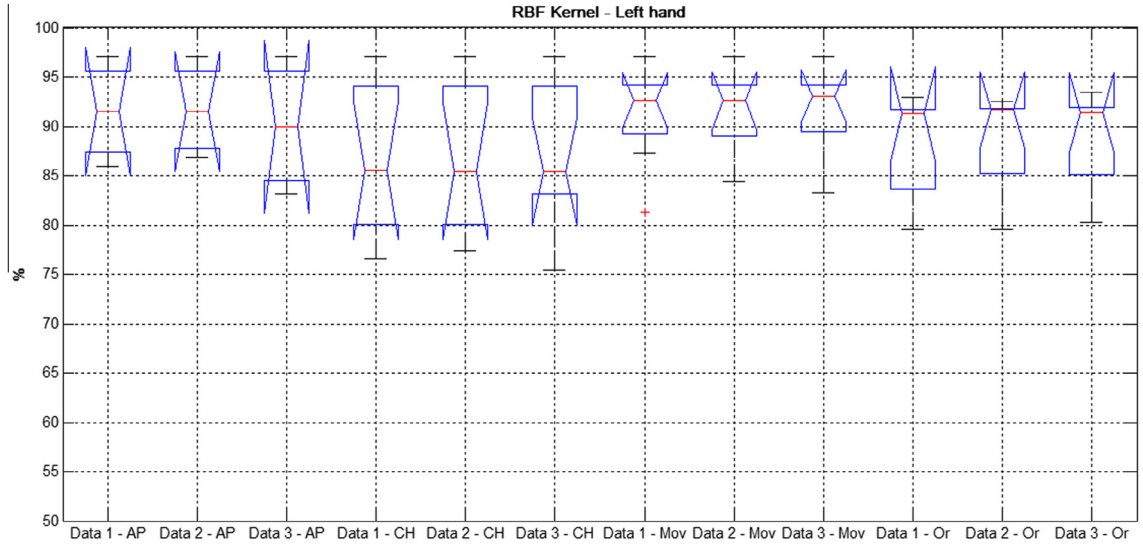


Fig. 11. Accuracy obtained by using the feature vectors described in Tables 3–5 for the RBF kernel in the classification of the elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the left hand.

this time information in the feature vector (see Figs. 10 and 11). Nevertheless, based on these experiments, we decided to adopt the feature vector with the time information appended as in Table 4.

6.3. Sign recognition

The experiments in the previous sections show that the accuracy in identifying and classifying the attributes in the structural elements was high. However, this information should be integrated somehow in order to correctly recognize one sign among the others. This can be done in different ways. In this section we perform experiments by using a simple method for combining this information: we compare the output of the classifiers for each attribute of each element in both hands to the values in the lookup tables presented in Figs. 6 and 7.

We calculate the Hamming distance between the output vector and all the lines in the lookup tables, finding the closest k points (with $k = 1, 3, 5$).

Using the RBF kernel and the feature vector in Table 4, the average accuracy is shown in Table 6. In this Table, we compute the accuracy by considering that a correct recognition is achieved when the correct sign is within the k closest neighbors in the lookup table, and we vary k from 1 to 5.

Sign Language Recognition is a very complex and challenging problem, requiring the integration of feature extraction techniques and technology and classification methods. Nonetheless, this experiment shows that extracting features and relating them to the structural elements of the language does help the recognition of the signs. The previous sections showed that the accuracy for classifying attributes of the structural elements is high, which means that the feature extraction methodology proposed is indeed useful to recognize the elements composing the sign. However, the integration of this information for actual sign recognition is still a challenge. Although some signs are correctly recognized with a high rate, other signs still need additional information such that correct recognition be achieved.

Table 6
Average accuracy for the recognition of each one of the 34 signs in the database, based on the closest point in the lookup table.

	Signs	$k = 1$ (%)	$k = 3$ (%)	$k = 5$ (%)
1.	Person	95	100	100
2.	To spread	51	76	93
3.	To copy	67	79	93
4.	To catch	0	6	7
5.	To gather	63	80	87
6.	To disappear	94	97	100
7.	To look	98	100	100
8.	Fair	74	100	100
9.	Truth	43	84	96
10.	Weight	88	98	99
11.	Justice	24	65	72
12.	Who	66	100	100
13.	Nothing	81	96	98
14.	To believe	80	92	97
15.	To forget	75	100	100
16.	To love	20	100	100
17.	To afflict	43	79	91
18.	To commemorate	7	16	21
19.	Rancor	20	99	100
20.	Assembly meeting	7	51	70
21.	To compare	30	79	92
22.	To scream	98	100	100
23.	To speak	70	100	100
24.	To absorb	86	98	97
24.	To absorb	17	43	55
26.	To quarrel	18	45	69
27.	Perspicacious	22	67	100
28.	To shine	88	100	100
29.	Maid	35	70	64
30.	To replace	38	66	78
31.	Prison	37	49	66
32.	Television	0	4	6
33.	Yesterday	0	2	5
34.	Future	1	1	1

If a new sign should be added to the sign recognition system, we just need to add the recorded sign in the database and its description in terms of the structural elements as in Figs. 6 and 7. The number of attributes and elements is always fixed, thus the number of classes in the classification system does not grow with adding new signs. Therefore, in our approach scalability can in principle be manageable given that BSL has more than 10,000 signs.

7. Conclusions

In this paper we have presented a methodology for feature extraction in Brazilian Sign Language that explores the phonological structure of the language and relied on RGB-D sensor for obtaining data. From the RGB-D images we have obtained seven vision-based features. We have related these features to structural elements based on shape, movement and position of the hands. The experiments show that the attributes of these elements can be successfully recognized in terms of the features obtained from the RGB-D images, with accuracy results individually above 80% on average.

RGB-D sensors show a great potential as a tool in hand gesture recognition in general, and in Sign Language Recognition specifically, because it produces three important types of image at the same time: intensity image, depth image and skeleton image. Joining information from these three images provide great power and flexibility in which we can improve algorithms for image processing.

The sign recognition rate was high for some signs and low for others, but in general we can conclude that the proposed feature extraction methodology and the decomposition of the signs into their phonological structural can help expert systems designed for SLR. The proposed methodology can also be applicable not only to BSL but also to other sign languages in the world, because all sign languages can be described in terms of their specific phonological structure. This methodology shows great potential in SLR, but also some challenges. The structure of sign language is very complex as described in Quadros and Karnopp (2004), which was simplified and summarized in the paper. The proposed extracted features could be selected by using feature selection techniques based on quality measures, such as ranking with F-Score or the Pearson correlation coefficient. Feature selection is left for future work. Another important point to be explored is the study of the classification system performance based on the extracted features. However, given the complexity of sign languages in general, perhaps a more fruitful avenue would be exploring a more complex description of the phonological structure in order to improve sign recognition. This paper showed that even a simplified use of this structure is beneficial in the design of systems for SLR. Finally, other ideas about how to integrate the information acquired in the classification of elements into sign recognition should be further explored.

Acknowledgments

The authors would like to thank the support given by the following Brazilian agencies: State of Minas Gerais Research Foundation – FAPEMIG; Coordination for the Improvement of Higher Level Personnel – CAPES; National Council of Scientific and Technological Development – CNPq (Grants 312276/2013-3); and by the Federal Institute of Minas Gerais (IFMG).

We also thank the anonymous reviewers for their valuable criticism, which have greatly contributed to the final version of the manuscript.

References

Ahmed, T. (2012). A neural network based real time hand gesture recognition system. *International journal of computer applications* (Vol. 59, pp. 17–22). New York, USA: Foundation of Computer Science.

AL-Rousan, M., Assaleh, K., & Tala'a, A. (2009). Video-based signer-independent arabic sign language recognition using hidden markov models. *Applied Soft Computing, 9*, 990–999.

Aran, O., Burger, T., Caplier, A., & Akarun, L. (2009). A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition, 42*, 812–822.

Auephanwiriyakul, S., Phitakwinai, S., Suttapak, W., Chanda, P., & Theera-Umpon, N. (2013). Thai sign language translation using scale invariant feature transform and hidden markov models. *Pattern Recognition Letters, 34*, 1291–1298.

Bossard, B., Braffort, A., & Jardino, M. (2003). Some issues in sign language processing. In A. Camurri & G. Volpe (Eds.), *Gesture Workshop* (pp. 90–100). Springer.

Boulares, M., & Jemni, M. (2012). 3D motion trajectory analysis approach to improve sign language 3d-based content recognition. *Procedia Computer Science, 13*, 133–143. Proceedings of the International Neural Network Society Winter Conference (INNS-WC2012).

Brashear, H., Starner, T., Lukowicz, P., & Junker, H. (2003). Using multiple sensors for mobile sign language recognition. In *Proceedings of the seventh IEEE international symposium on wearable computers* (pp. 45–52). Washington, DC, USA: IEEE Computer Society.

Capovilla, F. C., Raphael, W. D., & Maurício, A. C. L. (2012). Novo Deit-Libras: Dicionário Enciclopédico Ilustrado Trilíngue da Língua Brasileira de Sinais (Libras) baseado em Linguística e Neurociências Cognitivas, Volume I: Sinais de A a H. volume I. Edusp.

Capovilla, F. C., Raphael, W. D., & Maurício, A. C. L. (2012). Novo Deit-Libras: Dicionário Enciclopédico Ilustrado Trilíngue da Língua Brasileira de Sinais (Libras) baseado em Linguística e Neurociências Cognitivas, Volume I: Sinais de I a Z. volume II. Edusp.

Caridakis, G., Karpouzis, K., Drosopoulos, A., & Kollias, S. (2012). Non parametric, self organizing, scalable modeling of spatiotemporal inputs: The sign language paradigm. *Neural Networks, 36*, 157–166.

Chaaraoui, A. A., Padilla-López, J. R., Climent-Pérez, P., & Flórez-Revuelta, F. (2014). Evolutionary joint selection to improve human action recognition with rgb-d devices. *Expert Systems with Applications, 41*, 786–794. Methods and Applications of Artificial and Computational Intelligence.

Chakraborty, P., Sarawgi, P., Mehrotra, A., Agarwal, G., & Pradhan, R. (2008). Hand gesture recognition: A comparative study. *International MultiConference of Engineers and Computer Scientists, 1*, 388–393.

Chen, L., Wei, H., & Ferryman, J. (2013). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters, 34*, 1995–2006.

Chunli, W., Wen, G., & Jiyong, M. (2002). A real-time large vocabulary recognition system for chinese sign language. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and sign language in human-computer interaction. Lecture notes in computer science* (Vol. 2298, pp. 86–95). Berlin, Heidelberg: Springer.

Cooper, H., & Bowden, R. (2007). Large lexicon detection of sign language. In M. Lew, N. Sebe, T. Huang, & E. Bakker (Eds.), *Human-computer interaction. Lecture notes in computer science* (Vol. 4796, pp. 88–97). Berlin, Heidelberg: Springer.

Cooper, H., Pugeault, N., & Bowden, R. (2011). Reading the signs: A video based sign dictionary. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 914–919).

Cooper, H., Ong, E. J., Pugeault, N., & Bowden, R. (2012). Sign language recognition using sub-units. *Journal of Machine Learning Research, 13*, 2205–2231.

Cruz, L., Lucio, D., & Velho, L. (2012). Kinect and rgbd images: Challenges and applications. In *Proceedings of the 2012 25th SIBGRAPI conference on graphics, patterns and images tutorials* (pp. 36–49). Washington, DC, USA: IEEE Computer Society.

Cui, Y., & Weng, J. (2000). Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding, 78*, 157–176.

Dias, J., de Souza, K. P., & Pistori, H. (Eds.). (2006). Conjunto de Treinamento para Algoritmos de Reconhecimento de LIBRAS, II Workshop de Visão Computacional. UFSCar.

Dihl, L., & Musse, S. R. (2014). Recovering 3d human pose based on biomechanical constraints, postures comfort and image shading. *Expert Systems with Applications, 41*, 6305–6314.

Dipietro, L., Sabatini, A. M., & Dario, P. (2008). A survey of glove-based systems and their applications. *Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, 38*, 461–482.

Diraco, G., Leone, A., & Siciliano, P. (2013). Human posture recognition with a time-of-flight 3d sensor for in-home applications. *Expert Systems with Applications, 40*, 744–751.

Dreuw, P., Stein, D., & Ney, H. (2009). Enhancing a sign language translation system with vision-based features. In M. Sales Dias, S. Gibet, M. Wanderley, & R. Bastos (Eds.), *Gesture-based human-computer interaction and simulation. Lecture notes in computer science* (Vol. 5085, pp. 108–113). Berlin, Heidelberg: Springer.

Duduchi, M., & Capovilla, F. C. (2006). Buscasigno: A construção de uma interface computacional para o acesso ao léxico da língua de sinais brasileira. In *Proceedings of VII Brazilian symposium on human factors in computing systems* (pp. 21–30). New York, NY, USA: ACM.

Escalera, S., González, J., Guyon, I., & Moeslund, T. (2013). Multi-modal challenge – 2013.

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

Freitas, A. R. R., Guimaraes, F. G., Pedrosa Silva, R. C., & Souza, M. J. F. (2014). Memetic self-adaptive evolution strategies applied to the maximum diversity problem. *Optimization Letters, 8*, 705–714.

Futane, P. R., Dharaskar, R. V., & Thakare, V. M. (2012). A comparative study for approaches for hand sign language. In *IJCA proceedings on national conference on innovative paradigms in engineering and technology (NCIPET 2012)* (pp. 36–39). New York, USA: Foundation of Computer Science.

Haberdar, H., & Albayrak, S. (2005). Real time isolated turkish sign language recognition from video using hidden markov models with global features. In *Proceedings of the 20th international conference on computer and information sciences* (pp. 677–687). Berlin, Heidelberg: Springer-Verlag.

Haberdar, H., & Albayrak, S. (2006). A two-stage visual turkish sign language recognition system based on global and local features. In F. Esposito, Z. W. Ras, D. Malerba, & G. Semeraro (Eds.), *Foundations of intelligent systems. Lecture notes in computer science* (Vol. 4203, pp. 29–37). Berlin, Heidelberg: Springer.

Harris, C., & Stephens, M. (1988). A combined corner and edge detection. In *Proceedings of the fourth alvey vision conference* (pp. 147–151).

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*, 185–203.

Huang, C. L., & Huang, W. Y. (1998). Sign language recognition using model-based tracking and a 3d hopfield neural network. *Machine Vision and Applications, 10*, 292–307.

Illingworth, J., & Kittler, J. (1988). A survey of the hough transform. *Computer Vision, Graphics, and Image Processing, 44*, 87–116.

Johnston, A. T. (2003). Language standardization and signed language dictionaries. *Sign Language Studies*, 431–468.

Karami, A., Zanj, B., & Sarkaleh, A. K. (2011). Persian sign language (psl) recognition using wavelet transform and neural networks. *Expert Systems with Applications, 38*, 2661–2667.

Karmokar, B. C., Alam, K. M. R., & Siddiquee, M. K. (2012). Bangladeshi sign language recognition employing neural network ensemble. *International journal of computer applications* (Vol. 58, pp. 43–46). New York, USA: Foundation of Computer Science.

Kawulok, M. (2008). Dynamic skin detection in color images for sign language recognition. In A. Elmoataz, O. Lezoray, F. Nouboud, & D. Mammass (Eds.), *Image and signal processing. Lecture notes in computer science* (Vol. 5099, pp. 112–119). Berlin, Heidelberg: Springer.

Kuo, C. C., Glover, F., & Dhir, K. S. (1993). Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sciences, 24*, 1171–1185.

Li, Y. (2012). Hand gesture recognition using kinect. In *2012 IEEE third international conference on software engineering and service science (ICSESS)* (pp. 196–199).

Liu, X., & Fujimura, K. (2004). Hand gesture recognition using depth data. In *Proceedings of the sixth IEEE international conference on Automatic face and gesture recognition* (pp. 529–534). Washington, DC, USA: IEEE Computer Society.

Loeding, B. L., Sarkar, S., Parashar, A., & Karshmer, A. I. (2004). Progress in automated computer recognition of sign language. In K. Miesenberger, J. Klaus, W. Zagler, & D. Burger (Eds.), *Computers helping people with special needs. Lecture notes in computer science* (Vol. 3118, pp. 1079–1087). Berlin, Heidelberg: Springer.

Mankoff, K. D., & Russo, T. A. (2013). The kinect: A low-cost, high-resolution, short-range 3d camera. *Earth Surface Processes and Landforms, 38*, 926–936.

Maraqa, M., Al-Zboun, F., Dhyabat, M., & Zitar, R. A. (2012). Recognition of arabic sign language (arsl) using recurrent neural networks. *Journal of Intelligent Learning Systems and Applications, 4*, 41–52.

MathWorks (2012). Matlab r2012a (7.14.0.739).

Microsoft (2013). Microsoft kinect for windows.

Ong, E. J., Cooper, H., Pugeault, N., & Bowden, R. (2012). Sign language recognition using sequential pattern trees. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2200–2207).

Ong, S. C. W., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 873–891.

Oszust, M., & Wysocki, M. (2013). Polish sign language words recognition with kinect. In *2013 The sixth international conference on human system interaction (HSI)* (pp. 219–226).

Palacios, J. M., Sagüés, C., Montijano, E., & Llorente, S. (2013). Human-computer interaction based on hand gestures using rgb-d sensors. *Sensors, 13*, 11842–11860.

Parton, B. S. (2006). Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence. *Journal of Deaf Studies and Deaf Education, 11*, 94–101.

Parvini, F., Mcleod, D., Shahabi, C., Navai, B., Zali, B., & Ghandeharizadeh, S. (2009). An approach to glove-based gesture recognition. In *Proceedings of the 13th international conference on human-computer interaction. Part II: Novel interaction methods and techniques* (pp. 236–245). Berlin, Heidelberg: Springer-Verlag.

Phadtare, L., Kushalnagar, R., & Cahill, N. (2012). Detecting hand-palm orientation and hand shapes for sign language gesture recognition using 3d images. In *Image processing workshop (WNYIPW)* (pp. 29–32). Western New York.

Quadros, R. M., & Karnopp, L. B. (2004). Língua de sinais brasileira: estudos linguísticos. Artmed.

Quadros, R. M., Oliveira, J., & Miranda, R. D. (2007). Núcleo de aquisição de línguas de sinais - universidade federal de santa catarina – identificador de sinais.

Ramey, A., González-Pacheco, V., & Salichs, M. A. (2011). Integration of a low-cost rgb-d sensor in a social robot for gesture recognition. In *Proceedings of the sixth international conference on human-robot interaction* (pp. 229–230). New York, NY, USA: ACM.

Ramirez-Giraldo, D., Molina-Giraldo, S., Alvarez-Meza, A., Daza-Santacoloma, G., & Castellanos-Dominguez, G. (2012). Kernel based hand gesture recognition using kinect sensor. In *2012 XVII Symposium of image, signal processing, and artificial vision (STSIVA)* (pp. 158–161).

Ren, Z., Yuan, J., Meng, J., & Zhang, Z. (2013). Robust part-based hand gesture recognition using kinect sensor. *IEEE Transactions on Multimedia, 15*, 1110–1120.

Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. In *Proceedings of the ninth European conference on computer vision – volume part I* (pp. 430–443). Berlin, Heidelberg: Springer-Verlag.

Schemer, G. M. (2003). From variant to standard: An overview of the standardization process of the lexicon of sign language of the Netherlands. *Sign Language Studies*, 469–486.

Shi, J., & Tomasi, C. (1994). Good features to track. In *IEEE computer society conference on computer vision and pattern recognition, 1994. Proceedings CVPR* (pp. 593–600). IEEE.

Sole, M., & Tsoeu, M. (2011). Sign language recognition using the extreme learning machine. In *AFRICON* (pp. 1–6).

Starner, T., & Pentland, A. (1995). Real-time american sign language recognition from video using hidden markov models. In *International Symposium on Computer Vision, 1995. Proceedings* (pp. 265–270).

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

Suarez, J., & Murphy, R. (2012). Hand gesture recognition with depth images: A review. In *RO-MAN: The 21st IEEE international symposium on robot and human interactive communication, 2012 IEEE* (pp. 411–417).

Trindade, P., Lobo, J., & Barreto, J. (2012). Hand gesture recognition using color and depth images enhanced with hand angular pose data. In *2012 IEEE conference on multisensor fusion and integration for intelligent systems (MFI)* (pp. 71–76).

Uebersax, D., Gall, J., Van den Bergh, M., & Van Gool, L. (2011). Real-time sign language letter and word recognition from depth data. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 383–390).

Van Cleve, J. V. (2003). Lexicography and the university: Making the gaulladet dictionary of american sign language. *Sign Language Studies, 4*, 487–500.

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

Watson, R., & College, T. (1993). A survey of gesture recognition techniques. Technical Report. Trinity College Dublin.

Yang, H. D., & Lee, S. W. (2013). Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters, 34*, 2051–2056.

Yang, R., Sarkar, S., Loeding, B., & Karshmer, A. I. (2006). Efficient generation of large amounts of training data for sign language recognition: A semi-automatic tool. In K. Miesenberger, J. Klaus, W. Zagler, & A. Karshmer (Eds.), *Computers helping people with special needs. Lecture notes in computer science* (Vol. 4061, pp. 635–642). Berlin, Heidelberg: Springer.

Yin, P., Starner, T., Hamilton, H., Essa, I., & Rehg, J. M. (2009). Learning the basic units in american sign language using discriminative segmental feature selection. In *Proceedings of the 2009 IEEE international conference on acoustics, speech and signal processing* (pp. 4757–4760). Washington, DC, USA: IEEE Computer Society.

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

Zahedi, M., Keysers, D., Deselaers, T., & Ney, H. (2005a). Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Proceedings of the 27th DAGM conference on pattern recognition* (pp. 401–408). Berlin, Heidelberg: Springer-Verlag.

Zahedi, M., Keysers, D., & Ney, H. (2005b). Appearance-based recognition of words in american sign language. In *Proceedings of the second iberian conference on pattern recognition and image analysis – volume part I* (pp. 511–519). Berlin, Heidelberg: Springer-Verlag.

Zaki, M. M., & Shaheen, S. I. (2011). Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters, 32*, 572–577.

Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K., & Yang, J. (2011). A framework for hand gesture recognition based on accelerometer and emg sensors. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 41*, 1064–1076.

Zhang, C., Yang, X., & Tian, Y. (2013). Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1–8).

A.2 Proceedings - Workshop de Visão Computacional - 2013

A Methodology for Feature Extraction in Brazilian Sign Language Recognition

Sílvia Grasiella Moreira Almeida
Departamento de Engenharia Elétrica
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
Email: silvia.almeida@ifmg.edu.br

Frederico Gadelha Guimarães
Departamento de Engenharia Elétrica
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
Email: fredericoguimaraes@ufmg.br

Jaime Arturo Ramírez
Departamento de Engenharia Elétrica
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
Email: jramirez@ufmg.br

***Resumo*—Feature extraction is an essential step in any computer system designed for automatic pattern recognition. Understanding the problem at hand and finding suitable features that allow high recognition rate can greatly improve the efficiency and efficacy of the whole pattern recognition system. In this paper we present a novel methodology for feature extraction in automatic recognition of Brazilian Sign Language. The proposed methodology exploits the structure of the language in terms of location of the sign, movement and orientation, and hand configuration to develop strategies and algorithms for feature extraction in this context. The approach is based on the RGB-D sensor Kinect and computer vision methods. The results show that the features extracted are indeed related to the parameters of the Brazilian Sign Language and illustrate the potential of the methodology for sign language recognition systems.**

I. Introdução

As línguas de sinais são línguas pertencentes à modalidade linguística visual-espacial e, como toda língua, foi criada com o intuito de permitir a comunicação da forma mais natural possível entre pessoas surdas e entre surdos e ouvintes. Tais línguas possuem estrutura linguística completa, apresentando regras gramaticais, semânticas e morfológicas próprias. A língua brasileira de sinais, chamada Libras, é o objeto deste trabalho. Devido à citada característica espacial da Libras nota-se grande semelhança entre esta e línguas de sinais de países de todo o mundo, o que não necessariamente ocorre no caso de línguas de características orais-escritas.

No Brasil há cerca de 9.500 sinais de Libras documentados no dicionário dos autores em [1] e [2]. Estes sinais são descritos de três formas: como figura, em língua portuguesa e em língua inglesa. À medida em que aumenta o uso da Libras, percebe-se o dinamismo desta língua, com a quantidade de sinais documentados crescendo e abrangendo novos sinais específicos para diversas áreas de conhecimento. Cada novo sinal criado deve, obrigatoriamente, obedecer à estrututura gramatical, morfológica e semântica da língua. O trabalho a seguir apresenta uma metodologia de extração de atributos relacionados à estrutura morfológica da Libras.

Em [3], os autores apresentam uma metodologia de reconhecimento de sinais que explora a estrutura espacial da língua baseando-se em atributos de visão. Os atributos extraídos dos sinais relacionam-se à estrutura espacial da seguinte forma: a posição da *kurtosis* com a localização do sinal; a análise de componentes principais (PCA - Principal Component Analysis) com a orientação e configuração de mãos dos sinais e o código da cadeia de movimento (MCC - Motion Chain Code) representando o movimento das mãos no sinal.

Já em [4], os autores tratam o problema que consiste em estabelecer-se um ponto de referência na observação dos sinais. Para este caso, criam uma matriz de posição fundamental que contém a geometria do problema. Ela deve ser única para dois pontos de referência quando a observação dá-se de forma síncrona em um ambiente virtual de visão estéreo. O método proposto é dependente do auxílio de marcadores, uma vez que para aumentar a precisão das amostras são utilizadas luvas coloridas marcando as mãos durante as gravações.

Em [5] realiza-se a classificação dos sinais usando uma árvore de partição recursiva com a aplicação de técnicas de análise de componentes principais (PCA - Principal Component Analysis) e análise de discriminantes múlitplos (MDA - Multiple Discriminat Analysis) a cada nó da árvore. A partir de cada nó encontram-se automaticamente os melhores subconjuntos de características extraídas das imagens durante a etapa de aprendizagem.

Os autores em [6] apresentam um trabalho onde o reconhecimento de gestos é utilizado para comandar um robô. A metodologia proposta extrai atributos que incluem formas e localização das mãos, além de seu movimento. Uma câmera dedicada à gravação de vídeos de profundidade é utilizada na aquisição das imagens.

No Brasil há trabalhos desenvolvidos especificamente para a Libras em diversas vertentes. Trabalhos apresentados em [7] e [8] propõem sistemas de reconhecimento de sinais utilizando modelos ocultos de Markov (HMM - Hidden Markov Model). Os autores em [9] apresentam um modelo baseado em realidade aumentada para a criação de avatares que se comunicam em Libras. Sinais do alfabeto são reconhecidos por meio de momentos invariantes em [10]. Em [11], o autor desenvolve um *software* de reconhecimento de sinais individuais gravados em vídeos. Utiliza para tal um sistema semi-intrusivo, pois há a necessidade do uso de luvas plásticas durante a gravação dos sinais.

No trabalho descrito neste artigo são realizadas análises de sete atributos que possuem como entrada vídeos gravados por uma câmera RGB-D (Red, Green, Blue, Depth) de baixo custo. A detecção e o rastreamento das mãos é feita a partir das informações de cor (padrão RGB), profundidade (D) e posição em coordenadas lineares (x,y,z) obtidas nesta gravação. Desta forma, não há necessidade do uso de marcadores nas mãos. O

objetivo da extração e análise destes sete atributos é reconhecer automaticamente as saídas morfológicas possíveis descritas na Libras. Assim, novos sinais poderão ser detectados por um sistema de classificação a partir das saídas linguísticas e a quantidade de sinais reconhecido pode aumentar a partir da combinação de cada uma das saídas morfológicas.

II. Estrutura das línguas de sinais

As regras gramaticais, morfológicas e semânticas da Libras seguem princípios gerais que permitem a criação de uma grande quantidade de sinais a partir de um número finito de regras combinadas. Cada sinal pode ser caracterizado, morfologicamente, pelos seguintes elementos: (i) configuração de mãos, (ii) movimento e orientação, (iii) ponto de articulação e (iv) sinais não manuais (expressões faciais e corporais). Estes elementos são descritos brevemente a seguir.

Configuração de mãos: no Brasil tem-se definidas 74 configurações de mãos até o presente momento, descritas em [12]. Cada uma das mãos deve estar em uma formação específica para um sinal, que pode ser estático ou dinâmico. Se dinâmico, a configuração de mão pode sofrer ou não alteração ao longo da sua execução.

Movimento e orientação: possuem informações que se complementam. Haverá movimento e orientação somente se o sinal for dinâmico. Em [13] estão descritas as possibilidades de tipos de movimentos e orientações possíveis na língua.

Ponto de articulação: fornece a posição onde a maior parte do sinal é realizada. O ponto de articulação existe tanto para sinais estáticos quanto dinâmicos. No caso dos sinais dinâmicos, há a possibilidade de o ponto de articulação sofrer alterações durante a execução do sinal. As possíveis localizações para o ponto de articulação estão definidas em [13].

Sinais não manuais: são compostos por expressões faciais e corporais. Eles são responsáveis por fornecer os sentimentos aos sinais, tais como alegria, tristeza, dúvida, entre outros. Este reconhecimento de emoções tem sido objeto de estudo em diversos trabalhos na área de computação afetiva. Os autores em [14] e [15] já vislumbram em seus artigos a capacidade de reconhecimento de emoções por computadores. Este elemento, no entanto, não é objeto desta pesquisa.

Cada sinal em língua de sinais está completamente definido quando seus parâmetros morfológicos estão definidos. Já com relação à semântica do sinal, ou seja, seu significado, é possível haver alterações de acordo com o contexto onde este encaixa-se.

III. Metodologia

A. Detecção da região de interesse

A primeira etapa do trabalho consiste em extrair as regiões que são interesse para o trabalho. Neste caso, esta região é composta pelas mãos direita e esquerda. Um resumo desta etapa pode ser visto na Figura 1 e é descrito nesta seção.

Para gravação do sinal utiliza-se o *software nuiCapture Analyse*, dedicado para operação com o *Kinect Sensor*. O *Kinect* é uma câmera RGB-D (Red, Green, Blue e Depth) de baixo custo desenvolvida para o *videogame* XBOX da

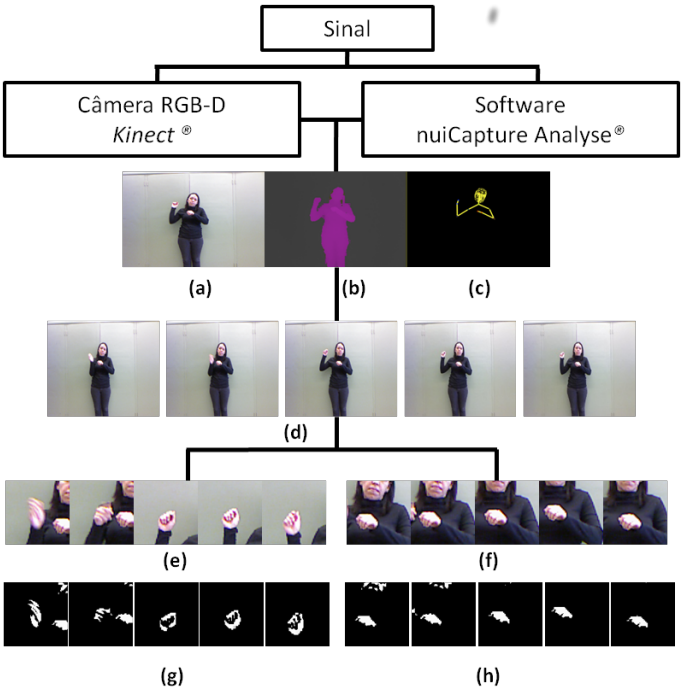


Figura 1. Detecção da região de interesse - mãos direita e esquerda para o sinal "Açoite". (a), (b), (c) Vídeo gravado utilizando o *software* nuiCaptureAnalyse operando o sensor RGB-D *Kinect*. (d) Quadros selecionados a partir da sumarização do vídeo por meio do Problema de Diversidade Máxima. (e), (f) Regiões de interesse detectadas, em RGB, para a mão direita e mão esquerda, respectivamente. (g), (h) Regiões de interesse detectadas em preto e branco a partir da detecção da cor da pele.

Microsoft que permite reconhecimento de gestos nos jogos para o citado *videogame*. Ele possui uma câmera RGB, além de um conjunto emissor/sensor infravermelho que permite a gravação da profundidade. Há também microfones embutidos disponíveis. Este conjunto *software* + *hardware* permite a detecção e gravação de vinte pontos no corpo humano. Estas características podem ser vistas em [16].

Três vídeos são gravados a partir das ferramentas de *software* e *hardware* descritas e podem ser vistos na Figura 1: (a) um vídeo colorido em RGB, (b) um vídeo de profundidade e (c) um vídeo do esqueleto. A taxa de captura do vídeo é de 30 quadros/segundo e o formato de todos os três é AVI (formato de áudio e vídeo da *Microsoft*).

A quantidade de quadros disponíveis em um vídeo faz com que existam informações redundantes nestes. Para diminuir esta redundância utilizamos uma técnica de sumarização de vídeos baseada no problema da diversidade máxima para selecionar os 5 quadros mais diversos entre si. A solução implementada pode ser vista em [17]. Na Figura 1 (d), os 5 quadros selecionados para o sinal "Açoite" são exibidos como exemplo.

A informação de profundidade gravada, Figura 1 (b), permite a delimitação do espaço de busca pelas mãos separando o fundo da imagem do corpo do sinalizador por meio de um algoritmo baseado em cores. Já a partir do vídeo do esqueleto, Figura 1 (c), é possível delimitar a localização das mãos, tendo como entrada a figura já sem fundo. O vídeo é gravado sempre

utilizando a cor azul para o conjunto mão e pulso direitos e a cor alaranjada para o conjunto mão e pulso esquerdos. Com a informação de posição da região de interesse detectada na imagem do esqueleto faz-se a segmentação das mãos na imagem RGB. As regiões de interesse extraídas em formato RGB são exibidas em Figura 1 (e) para a mão direita e Figura 1 (f) para a mão esquerda. As mãos e o rosto nesta imagem são então detectados a partir da cor da pele e a imagem destas convertidas para o formato preto e branco, conforme Figura 1 (g) e Figura 1 (h).

Concluída a detecção e posterior extração da região de interesse, a próxima etapa consiste na extração e análise de características que representem os elementos morfológicos configuração de mãos, ponto de articulação, movimento e orientação dos sinais.

B. *Extração de características*

Ao observar-se a morfologia das línguas de sinais é possível relacionar seus elementos de configuração de mãos, movimento, orientação e ponto de articulação a formas e posições ao longo do tempo. Portanto, são extraídas da região de interesse atributos que fornecem informações relacionadas a forma e posição. Um vetor de características contendo informações dos atributos para ambas as mãos é construído. Este vetor é único e contém os valores dos atributos em análise para a região de interesse em cada um dos cinco quadros selecionados. Os atributos de cada quadro são adicionados sequencialmente, primeiramente para a mão direita e em seguida para a esquerda, conforme Tabela I. São M características extraídas e n quadros utilizados em cada vídeo.

Tabela I. *Estrutura do vetor de atributos para M atributos e n quadros. Neste trabalho são analisadas $M = 7$ características e $n = 5$ quadros.*

MÃO (DIREITA OU ESQUERDA)			
1	...	n	Quadros
1 a M	1 a M	1 a M	Atributos

As relações entre os atributos extraídos e os parâmetros de forma da Libras estão descritas na Tabela II. Os três parâmetros de forma são abreviados como PA (ponto de articulação), CM (configuração de mãos) e MO (movimento e orientação).

Tabela II. *Na coluna da esquerda encontram-se os atributos extraídos da região de interesse. O vetor final é composto por estes atributos para cada uma das mãos no sinal. Na coluna da direita estão relacionados quais elementos em Libras são descritos por estes atributos: PA: ponto de articulação, CM: configuração de mãos, MO: movimento e orientação.*

	ATRIBUTOS EM ANÁLISE	ELEMENTOS EM LIBRAS
1	Distância euclidiana (x,y) em pixels	PA
2	Distância (x,y,z) em milímetros	PA
3	Velocidade	MO
4	Área das mãos	CM
5	Pontos comuns entre os quadros	MO
6	Média das quinas	PA e MO
7	Linhas detectadas	CM

1) Distância euclidiana (x,y) em pixels: A distância euclidiana para os eixos (x,y) são calculadas entre as posições das mãos e o centro do ombro> O centro do ombro é a referência porque sofre pouquíssima variação durante a execução do sinal. O valor obtido como atributo é, na realidade, a distância euclidiana média dos *pixels* da mão direita, esquerda e centro do ombro encontradas na imagem gravada do esqueleto, conforme Figura 2 (a). O cálculo é uma média porque o algoritmo procura pelas mãos por meio da cor azul para a mão direita e alaranjada para a mão esquerda e estas regiões não são compostas somente por um *pixel*, mas por vários. Este atributo de distância permite que sejam obtidas informações relativas ao parâmetro linguístico ponto de articulação do sinal.

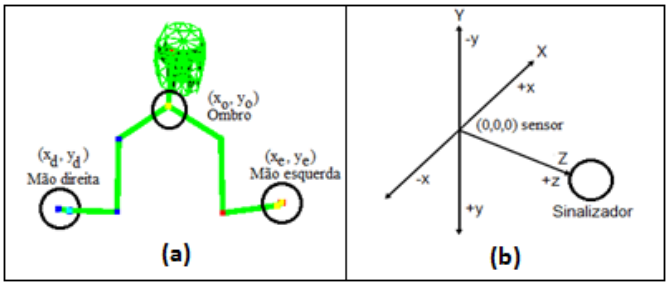


Figura 2. (a) Característica 1 extraída: distância bidimensional euclidiana em *pixels* entre as posições das mãos e do centro do ombro. (b) Referência para os cálculos da característica 2, distância das mãos nos eixos (x,y,z) em milímetros.

2) Distância (x,y,z) em milímetros: A distância em milímetros entre o sensor e o sinalizador é uma informação gravada pelo *software nuiCapture Analyze* e posteriormente exportada para o formato *Matlab*, [18]. Seus valores são as distâncias em coordenadas (*x,y,z*) entre as mãos e o sensor. A Figura 2 (b) mostra a referência (0,0,0) no sensor para o cálculo destas distâncias. Antes da normalização dos dados entre 0 e 1, há valores negativos, uma vez que tanto em x quanto em y é possível o sinal ser realizado nos quadrantes negativos de acordo com a referência utilizada. No eixo z, que representa a profundidade, não haverá valor negativo porque todo sinal é realizado à frente do sensor.

3) Velocidade: O algoritmo do fluxo óptico está descrito em [19] e é utilizado neste trabalho para detectar a velocidade da imagem por meio das diferenciações do brilho existente nestas. Na Figura 3 (a) são exibidos os vetores de velocidade calculados para cada um dos n quadros selecionados. O atributo velocidade calculado para cada um dos n quadros selecionados está relacionado aos parâmetros de movimento e orientação dos sinais estudados.

4) Área das mãos: Além de detectar velocidade por meio da descontinuidade do brilho das imagens, a técnica de fluxo óptico permite também a segmentação de objetos. Esta propriedade é utilizada para o cálculo da característica de número 4, que é a área dos objetos detectados nas imagens. Esta área detectada representa o maior objeto nas regiões de interesse extraídas, que são as mãos. Na Figura 3 (b) estão exibidas as áreas de maior região, que são as mãos. No quadro 2, há 3 regiões detectadas, pois no sinal realizado há cruzamento das mãos em frente ao rosto. Esta característica do sinal deve ser preservada, pois é uma informação importante, já que o cruzamento é característico do sinal. O atributo de área está relacionado diretamente à configuração de mãos do sinal realizado.

5) Quantidade de pontos comuns entre os quadros: O algoritmo do descritor SURF (Speed-Up Robust Features), descrito em [18], foi utilizado para o cálculo da quinta característica. Neste trabalho o SURF é aplicado na procura por pontos em comum entre dois quadros do vídeo dos sinais, técnica

chamada *template matching*. A cada dois quadros este descritor apresenta uma quantidade de pontos em comum. Quanto maior a quantidade de pontos encontrados, menor é o movimento existente entre os quadros, pois estes são semelhantes. A informação obtida, neste caso, é composta pela quantidade de pontos em comum encontrados em cada quadro e está relacionada ao parâmetro movimento e orientação da língua de sinais.

6) Média das quinas: A área da imagem onde há grande diferença entre os *pixels* é chamada quina. Métodos já bem conhecidos para realizar a detecção de quinas podem ser encontrados em [20], [21] e [22]. Neste trabalho, detectamos as quinas por meio do detector de quinas de *Harris*, [20]. O atributo calculado neste caso é uma média das posições das quinas para cada quadro. Esta média pode ser visualizada na Figura 3 (c) e está relacionada aos elementos ponto de articulação e movimento/orientação dos sinais.

7) Linhas detectadas: O sétimo atributo extraído são as linhas presentes em cada imagem. Faz-se esta extração por meio da transformada de *Hough*, [23]. Estas linhas são definidas em coordenadas polares e, portanto, o atributo é um par ordenado (ρ, θ), onde ρ é a distância do ponto à origem do vetor e θ é o ângulo formado entre o eixo polar e a reta P. Na Figura 3 (d) são exibidas as linhas de maior tamanho encontradas para cada quadro. A maior linha extraída da imagem em preto e branco contém informações sobre a configuração da mão, por estar relacionada especificamente à forma desta em cada um dos quadros escolhidos.

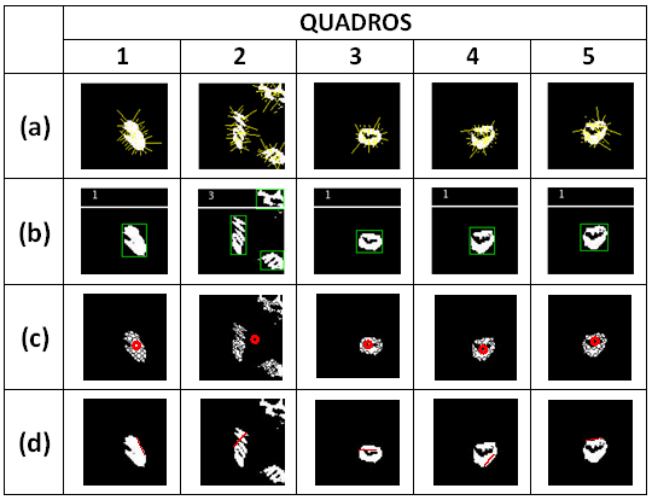


Figura 3. Exemplo da extração dos atributos: (a) Velocidade, (b) Área das mãos, (c) Média das quinas e (d) Linhas, para o sinal exemplo "Açoite", mão direita.

C. Abordagem dos parâmetros linguísticos

Todo sinal em Libras é composto por parâmetros definidos na seção II deste trabalho. Em [13] pode-se encontrar uma descrição detalhada sobre as definições de cada um destes parâmetros estruturais e suas saídas possíveis. No trabalho aqui descrito houve a necessidade de resumir estas saídas de forma que seja viável reconhecer os parâmetros na atual etapa desenvolvida. Cada um dos três parâmetros são descritos conforme a abordagem utilizada para reconhecimento neste trabalho: (A) Configuração de mãos, (B) Movimento e orientação e (C) Ponto de articulação.

1) Configuração de mãos: Em Libras, temos definidas 74 formas [12] para a configuração das mãos, muitas das quais bastante semelhantes em forma. Estas configurações semelhantes foram reunidas em 09 grupos, conforme descrito na Tabela III. Assim, em nossa abordagem a configuração de mão para cada sinal pertence a um grupo entre 09 possíveis. O vetor de saída criado possui 14 possibilidades para cada mão, composto pelos 09 agrupamentos mais 05 possibilidades de informações complementares para este parâmetro. Na Tabela IV encontra-se a descrição destas informações. As posições de 01 a 09 no vetor definem o agrupamento ao qual a configuração de mãos pertence. As posições 10, 11 e 12 descrevem se há variações na posição da configuração de mãos durante a execução do sinal. Já a posição 13 do vetor informa se há alteração na configuração de mãos ao longo do movimento. Finalmente, a posição 14 informa se a mão em questão faz parte ou não do sinal. Tanto o agrupamento quanto a definição dos outros 05 parâmetros complementares para a configuração de mãos foram estabelecidos para este trabalho pelos autores.

Tabela III. *Agrupamentos realizados por semelhança de forma para o parâmetro Configuração de mãos. Compõem o vetor de saída entre as posições 01 a 09 para este parâmetro.*

AGRUPAMENTOS	CONFIGURAÇÕES
1	01, 02, 03, 04, 05, 06
2	07, 08a, 08b, 09, 10, 12, 13, 14, 15,
	16, 24, 25, 31, 32, 38, 39, 40, 41, 42
3	11, 17, 18a, 18b
4	19, 20, 21, 22a, 22b, 23,
	29a, 29b, 30, 43, 44, 45
5	26, 27, 28
6	33, 34, 35a, 35b, 36, 37a, 37b,
	49, 54, 55, 56, 58, 59a
7	46a, 46b, 47, 48, 50, 51a, 51b
8	52, 53a, 53b, 57
9	60, 61, 62, 63, 64

Tabela IV. *Informações complementares para definição do parâmetro Configuração de mãos em um sinal. Compõem as posições 10 a 14 para o vetor de saída deste.*

ITEM	DESCRIÇÃO	NÚMERO NO VETOR
Posição da configuração de mãos	Normal	10
	90 graus no eixo x, lateralmente	11
	90 graus no eixo z, para frente	12
Mudança	Mesma configuração de mãos durante a execução do sinal (Sim ou Não)	13
Participação	Define se a mão participa ou não da execução do sinal (Sim ou Não)	14

2) Movimento e orientação: O vetor de saída para o parâmetro Movimento e Orientação tem suas posições de 01 a 05 compostas pelo tipo de movimento. Já as posições 06 a 10 definem a orientação deste movimento. A posição 11 define se a mão participa ou não do sinal. O vetor que compõe esta saída possui tamanho 22, pois é composto para as mãos direita e esquerda.

3) Ponto de articulação: As saídas definidas para o parâmetro Ponto de articulação compõem um vetor de tamanho 62 para ambas as mãos. Na Tabela VI estão descritas as saídas deste parâmetro. As posições 01 a 16 definem as saídas principais, que são as localizações possíveis para a mão durante a realização de um sinal. As posições entre 17 e 30 fornecem informações complementares ao espaço. A posição 31 define se a mão participa ou não do sinal.

Tabela V. *Definições do parâmetro Movimento-Orientação.*

ITEM	DESCRIÇÃO
1	Retilíneo
2	Helicoidal
3	Sinuoso
4	Circular
5	Semi-circular
6	Direita para esquerda
7	Esquerda para direita
8	Baixo para cima
9	De cima para baixo
10	Em diagonal
11	Define se a mão participa ou não do sinal (Sim ou Não).

Tabela VI. *Definições das saídas para o elemento Ponto de articulação.*

ITEM	DESCRIÇÃO	ITEM	DESCRIÇÃO
1	topo da cabeça/testa	17	lado direito
2	orelha	18	lado esquerdo
3	olhos/nariz/boca/bochechas	19	medial
4	zona abaixo do queixo	20	lateral
5	pescoço/ombro	21	em frente
6	busto	22	atrás
7	estômago	23	imediatamente próximo
8	cintura	24	distância média
9	braço	25	distante
10	antebraço	26	em contato
11	cotovelo	27	contato inicial
12	pulso	28	contato medial
13	palma da mão	29	contato final
14	costa da mão	30	cruzamento
15	perna	31	define se a mão participa
16	espaço neutro		ou não do sinal (Sim ou Não).

D. Sistema de classificação

Para o sistema de classificação optou-se por utilizar uma rede neural de arquitetura *feedforward* de duas camadas, uma oculta e a outra de saída. A função sigmóide é a função de ativação dos neurônios de ambas as camadas. O treinamento da rede é realizado pelo método de *Levenberg-Marquardt*.

E. Avaliação do método proposto

Para avaliação do método proposto foram realizados experimentos seguindo duas linhas distintas. Na primeira delas, o vetor de características fornecido como entrada para o sistema de classificação é composto apenas pelos atributos relacionados separadamente à cada um dos três parâmetros estruturais da Libras que se deseja identificar, conforme Tabela II. Para o segundo experimento, utiliza-se o vetor completo de características como entrada do sistema de classificação. As saídas do sistema são os vetores descritos no item C desta seção, "Abordagem dos parâmetros linguísticos'.'

F. Discussões

As características relativas ao cálculo de distâncias, 1 e 2, estão relacionadas ao parâmetro linguístico de forma ponto de articulação. Por meio dos cáculos de distância é possível inferir sobre a posição das mãos relativamente ao ombro e ao sensor. O atributo velocidade, característica 3, extraído por meio do fluxo óptico, contém informações sobre o movimento e a orientação dos sinais estudados. Já a área das mãos, atributo 4, fornece informações relativas ao parâmetro configuração de mãos. A característica 5, quantidade de pontos em comum entre quadros sequenciais contém informações a respeito do tipo de movimento do sinal. A média das quinas, extraída como característica de número 6, fornece informações sobre a orientação e ponto de articulação dos sinais. Por fim, a característica 7, que é a maior linha de cada quadro extraída por meio da transformada de *Hough* contém informações de configuração das mãos.

IV. Experimentos

Nesta seção são apresentados os resultados dos experimentos preliminares realizados para avaliação de desempenho das características extraídas dos vídeos dos sinais em Libras, conforme metodologia descrita na seção anterior.

Para implementação do sistema de classificação foi utilizado o *Toolbox* de Reconhecimento de Padrões do *Matlab*. Em todos os casos, foram realizadas trinta execuções do algoritmo e são apresentados os percentuais de acertos médios para as etapas de treinamento, validação e teste da classificação.

Para esta fase preliminar de resultados parciais foram gravados oito sinais, repetidos por cinco vezes. Eles estão descritos nas Tabelas VII, VIII e IX respectivamente com as definições para as saídas dos parâmetros estruturais Configuração de mãos, Ponto de articulação e Movimento/Orientação da Libras. Foram utilizadas as definições dos parâmetros nas Tabelas III, IV, VI e V para construção destas tabelas aqui apresentadas com os sinais.

Tabela VII. *Definição do parâmetro Configuração de mãos para os sinais gravados, mãos direita e esquerda. A) Define o grupo ao qual a configuração de mãos pertence. B) Define a posição da configuração de mãos. C) Define se há mudança de configuração de mãos durante a execução do sinal (Sim ou Não). D) Define se a mão faz parte do sinal (Sim ou Não).*

	Direita				Esquerda			
	A	B	C	D	A	B	C	D
1. Abafar	2	10	S	S	0	0	N	N
2. Acabar2	9	12	N	S	9	12	N	S
3. Acenar	9	10	N	S	0	0	N	N
4. Açoite	1	10	N	S	6	10	N	S
5. Acordeão	7	11	N	S	7	11	N	S
6. Alemão	2	10	N	S	0	0	N	N
7. Almejar	7	10	S	S	0	0	N	N
8. Apaixonado	7	10	S	S	0	0	N	N

Tabela VIII. *Definição do parâmetro Ponto de articulação para os sinais gravados, mãos direita e esquerda. A) Define a principal localização da mão. B) Definição secundária do ponto de articulação da mão. C) Define se a mão faz parte do sinal (Sim ou Não).*

	Direita			Esquerda		
	A	B	C	A	B	C
1. Abafar	3	20	S	0	0	N
2. Acabar2	16	30	S	16	30	S
3. Acenar	2	17	S	0	0	N
4. Açoite	2	17	S	16	0	S
5. Acordeão	16	30	S	16	30	S
6. Alemão	1	23	S	0	0	N
7. Almejar	6	19	S	0	0	N
8. Apaixonado	6	18	S	0	0	N

V. Resultados e discussões

Os resultados médios de trinta execuções do sistema de classificação são apresentados na Tabela X. Os ajuste do treinamento da rede baseia-se no erro quadrático médio (MSE - Mean Square Error). Já para a validação mede-se a capacidade de generalização da rede e esta é interrompida quando não há mais melhora. Por fim, os resultados de teste apresentados são independentes do treinamento. 70% das amostras são utilizadas no treinamento e 15% utilizadas tanto para validação quanto para teste. A rede neural utilizada para classificação foi

Tabela IX. *Definição do parâmetro Movimento-Orientação para os sinais gravados, mãos direita e esquerda. A) Define o principal movimento da mão. B) Definição secundária da orientação da mão, dependente do movimento. C) Define se a mão faz parte do sinal (Sim ou Não).*

	Direita			Esquerda		
	A	B	C	A	B	C
1. Abafar	1	7	S	0	0	N
2. Acabar2	1	7	S	1	6	S
3. Acenar	5	6,7	S	0	0	N
4. Açoite	1	8,9,10	S	0	0	S
5. Acordeão	3	6,7	S	3	6,7	S
6. Alemão	0	0	S	0	0	N
7. Almejar	5	0	S	0	0	N
8. Apaixonado	4	0	S	0	0	N

ajustada em todos os casos com 10 neurônios e tem seus pesos atualizados de acordo com o método de *Levenberg-Marquardt*.

Nos resultados apresentados para o Treinamento (i) e (ii), nota-se que em ambas as situações o acerto atinge um percentual acima de 90%. Para os parâmetros Ponto de Articulação e Movimento/Orientação, entretanto, o experimento com o vetor que contém os atributos relacionados aos parâmetros (i) obteve um percentual de acerto mais elevado.

Nos resultados apresentados para a Validação (i) e (ii), os melhores resultados foram obtidos quando utiliza-se o vetor completo de atributos (ii) na entrada do sistema de classificação, para os três parâmetros.

No último caso, resultados apresentados para Teste (i) e (ii) os melhores resultados ocorrem também para o experimento onde o vetor de entrada é constituído por todos os atributos extraídos (ii), embora não haja grande diferença percentual entre eles.

Tabela X. *Percentual de acerto do sistema de classificação para o treinamento, validação e teste. Os resultados para o vetor com atributos relativos à configuração de mãos (CM), ponto de articulação (PA) e movimento/orientação (MO), respectivamente, são representados nas colunas (i). Nas colunas (ii) estão os resultados para o vetor de atributos completo com todos os sete atributos extraídos.*

	TREINAMENTO		**VALIDACAO**		**TESTE**	
	(i)	(ii)	(i)	(ii)	(i)	(ii)
CM	91,14	94,25	86,44	90,79	83,60	88,90
PA	98,10	97,84	96,04	96,34	95,40	95,13
MO	93,39	91,76	86,55	87,08	83,73	85,00

VI. Conclusão

Pelos resultados de acerto apresentados na seção anterior é possível perceber que os atributos extraídos de fato relacionam-se com os parâmetros da Língua Brasileira de Sinais. A distinção nos experimentos não apresentou grande diferença quando o vetor de entrada é composto apenas pelos atributos específicos do parâmetro a ser reconhecido. Este fato ocorre porque, apesar da relação descrita na Tabela II, há informações sobre os três parâmetros contidas em todos os atributos extraídos. Os resultados apresentados neste trabalho são parciais. A questão a ser explorada e o principal foco do trabalho é extrair e selecionar atributos relacionados com a estrutura visual-espacial da Libras. Neste sentido, as características apresentadas aqui possuem um grande potencial para permitir o reconhecimento de um sinal em Libras em trabalhos futuros.

Referências

[1] F. C. Capovilla and W. D. Raphael. *Dicionário enciclopédico ilustrado trilíngüe da língua de sinais brasileira: Sinais de A a N*, volume 1. Edusp, 2001.

[2] F. C. Capovilla and W. D. Raphael. *Dicionário enciclopédico ilustrado trilíngüe da língua de sinais brasileira: Sinais de M a Z*, volume 2. Edusp, 2001.

[3] M. M. Zaki and S. I. Shaheen. Sign language recognition using a combination of new vision based features. *Pattern Recogn. Lett.*, 32(4):572–577, March 2011.

[4] Q. Wang, X. Chen, L. Zhang, C. Wang, and W. Gao. Viewpoint invariant sign language recognition. *Comput. Vis. Image Underst.*, 108(1-2):87–97, October 2007.

[5] Y. Cui and J. Weng. Appearance-based hand sign recognition from intensity image sequences. *Comput. Vis. Image Underst.*, 78(2):157–176, May 2000.

[6] X. Liu and K. Fujimura. Hand gesture recognition using depth data. In *Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*, FGR' 04, pages 529–534, Washington, DC, USA, 2004. IEEE Computer Society.

[7] J. B. Dias, K. P. Souza, and H. Pistori. Conjunto de treinamento para algoritmos de reconhecimento de libras. In *II Workshop de Visão Computacional, São Carlos*, 2006.

[8] K. P. de Souza, J. B. Dias, and H. Pistori. Reconhecimento automático de gestos da lıngua brasileira de sinais utilizando visao computacional. In *III Workshop de Visão Computacional*, 2007.

[9] J. R. Brega, C. Antônio, I. A. Rodello, E. Fusco, D. P. Silva, and F. H. Furlanetto. O emprego de realidade aumentada na viabilização da comunicação em libras, 2009.

[10] A. Carneiro, P. Cortez, and R. Costa. Reconhecimento de gestos da libras com classificadores neurais a partir dos momentos invariantes de hu. *Interaction South America*, 2009.

[11] F. B. Siola. Desenvolvimento de um software para reconhecimento de sinais em libras através de vıdeo. *Monografia, de Engenharia, Modelagem e Ciências Sociais Aplicadas - UFABC*, 2010.

[12] Filipe Antônio Castro de Farias. http://interpretelipe.blogspot.ca/, 2013.

[13] L. F. Brito. *A Língua Brasileira de Sinais*. Universidade Federal do Rio de Janeiro, MEC, 2012.

[14] R. W. Picard. Affective computing: challenges. *Int. J. Hum.-Comput. Stud.*, 59(1-2):55–64, July 2003.

[15] J. Tao and T. Tan. Affective computing: A review. In *Affective Computing and Intelligent Interaction*, pages 981–995. Springer, 2005.

[16] Microsoft Corporation. http://www.microsoft.com/en-us/kinectforwindows/, 2013.

[17] A. R. R. Freitas, F. G. Guimaraes, R. C. Pedrosa Silva, and M. J. F. Souza. Memetic self-adaptive evolution strategies applied to the maximum diversity problem. *Optimization Letters*, pages 1–10, 2013.

[18] MathWorks. Matlab r2012a (7.14.0.739), 2012.

[19] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, August 1981.

[20] C. Harris and M. Stephens. A Combined Corner and Edge Detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[21] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, June 1994.

[22] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Proceedings of the 9th European conference on Computer Vision - Volume Part I*, ECCV'06, pages 430–443, Berlin, Heidelberg, 2006. Springer-Verlag.

[23] J. Illingworth and J. Kittler. A survey of the hough transform. *Comput. Vision Graph. Image Process.*, 44(1):87–116, August 1988.